HHS Public Access

Author manuscript

Neural Comput. Author manuscript; available in PMC 2019 January 23.

Published in final edited form as:

Neural Comput. 2018 April ; 30(4): 885–944. doi:10.1162/neco_a_01056.

Information-Theoretic Bounds and Approximations in Neural Population Coding

Wentao Huang and

Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A., and Cognitive and Intelligent Lab and Information Science Academy of China Electronics Technology Group Corporation, Beijing 100846, China

Kechen Zhang

Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD 21205, U.S.A.

Abstract

While Shannon's mutual information has widespread applications in many disciplines, for practical applications it is often difficult to calculate its value accurately for high-dimensional variables because of the curse of dimensionality. This article focuses on effective approximation methods for evaluating mutual information in the context of neural population coding. For large but finite neural populations, we derive several information-theoretic asymptotic bounds and approximation formulas that remain valid in high-dimensional spaces. We prove that optimizing the population density distribution based on these approximation formulas is a convex optimization problem that allows efficient numerical solutions. Numerical simulation results confirmed that our asymptotic formulas were highly accurate for approximating mutual information for large neural populations. In special cases, the approximation formulas are exactly equal to the true mutual information. We also discuss techniques of variable transformation and dimensionality reduction to facilitate computation of the approximations.

1 Introduction

Shannon's mutual information (MI) provides a quantitative characterization of the association between two random variables by measuring how much knowing one of the variables reduces uncertainty about the other (Shannon, 1948). Information theory has become a useful tool for neuroscience research (Rieke, Warland, de Ruyter van Steveninck, & Bialek, 1997; Borst & Theunissen, 1999; Pouget, Dayan, & Zemel, 2000; Laughlin & Sejnowski, 2003; Brown, Kass, & Mitra, 2004; Quiroga & Panzeri, 2009), with applications to various problems such as sensory coding problems in the visual systems (Eckhorn & Pöpel, 1975; Optican & Richmond, 1987; Atick & Redlich, 1990; McClurkin, Gawne, Optican, & Richmond, 1991; Atick, Li, & Redlich, 1992; Becker & Hinton, 1992; Van Hateren, 1992; Gawne & Richmond, 1993; Tovee, Rolls, Treves, & Bellis, 1993; Bell & Sejnowski, 1997; Lewis & Zhaoping, 2006) and the auditory systems (Chechik et al., 2006; Gourévitch and Eggermont, 2007; Chase & Young, 2005).

One major problem encountered in practical applications of information theory is that the exact value of mutual information is often hard to compute in high-dimensional spaces. For example, suppose we want to calculate the mutual information between a random stimulus variable that requires many parameters to specify and the elicited noisy responses of a large population of neurons. In order to accurately evaluate the mutual information between the stimuli and the responses, one has to average over all possible stimulus patterns and over all possible response patterns of the whole population. This averaging quickly leads to a combinatorial explosion as either the stimulus dimension or the population size increases. This problem occurs not only when one computes MI numerically for a given theoretical model but also when one estimates MI empirically from experimental data.

Even when the input and output dimensions are not that high, an MI estimate from experimental data tends to have a positive bias due to limited sample size (Miller, 1955; Treves & Panzeri, 1995). For example, a perfectly flat joint probability distribution implies zero MI, but an empirical joint distribution with fluctuations due to finite data size appears to suggest a positive MI. The error may get much worse as the input and output dimensions increase because a reliable estimate of MI may require exponentially more data points to fill the space of the joint distribution. Various asymptotic expansion methods have been proposed to reduce the bias in an MI estimate (Miller, 1955; Carlton, 1969; Treves & Panzeri, 1995; Victor, 2000; Paninski, 2003). Other estimators of MI have also been studied, such as those based on *k*-nearest neighbor (Kraskov, Stögbauer, & Grassberger, 2004) and minimal spanning trees (Khan et al., 2007). However, it is not easy for these methods to handle the general situation with high-dimensional inputs and high-dimensional outputs.

For numerical computation of MI for a given theoretical model, one useful approach is Monte Carlo sampling, a convergent method that may potentially reach arbitrary accuracy (Yarrow, Challis, & Series, 2012). However, its stochastic and inefficient computational scheme makes it unsuitable for many applications. For instance, to optimize the distribution of a neural population for a given set of stimuli, one may want to slightly alter the population parameters and see how the perturbation affects the MI, but a tiny change of MI can be easily drowned out by the inherent noise in the Monte Carlo method.

An alternative approach is to use information-theoretic bounds and approximations to simplify calculations. For example, the Cramér-Rao lower bound (Rao, 1945) tells us that the inverse of Fisher information (FI) is a lower bound to the mean square decoding error of any unbiased decoder. Fisher information is useful for many applications partly because it is often much easier to calculate than MI (see e.g., Zhang, Ginzburg, McNaughton, & Sejnowski, 1998; Zhang & Sejnowski, 1999; Abbott & Dayan, 1999; Bethge, Rotermund, & Pawelzik, 2002; Harper & McAlpine, 2004; Toyoizumi, Aihara, & Amari, 2006).

A link between MI and FI has been studied by several researchers (Clarke & Barron, 1990; Rissanen, 1996; Brunel & Nadal, 1998; Sompolinsky, Yoon, Kang, & Shamir, 2001). Clarke and Barron (1990) first derived an asymptotic formula between the relative entropy and FI for parameter estimation from independent and identically distributed (i.i.d.) observations with suitable smoothness conditions. Rissanen (1996) generalized it in the framework of stochastic complexity for model selection. Brunel and Nadal (1998) presented an asymptotic

relationship between the MI and FI in the limit of a large number of neurons. The method was extended to discrete inputs by Kang and Sompolinsky (2001). More general discussions about this also appeared in other papers (e.g., Ganguli & Simoncelli, 2014; Wei & Stocker, 2015). However, for finite population size, the asymptotic formula may lead to large errors, especially for high-dimensional inputs, as detailed in sections 2.2 and 4.1.

In this article, our main goal is to improve FI approximations to MI for finite neural populations especially for high-dimensional inputs. Another goal is to discuss how to use these approximations to optimize neural population coding. We will present several information-theoretic bounds and approximation formulas and discuss the conditions under which they are established in section 2, with detailed proofs given in the appendix. We also discuss how our approximation formulas are related to other statistical estimators and information-theoretic bounds, such as Cramér-Rao bound and van Trees' Bayesian Cramér-Rao bound (see section 3). In order to better apply the approximation formulas in high-dimensional input space, we propose some useful techniques in section 4, including variable transformation and dimensionality reduction, which may greatly reduce the computational complexity for practical applications. Finally, in section 5, we discuss how to use the approximation formulas for optimizing information transfer for neural population coding.

2 Bounds and Approximations for Mutual Information in Neural Population Coding

2.1 Mutual Information and Notations.

Suppose the input $\mathbf{x}$ is a K-dimensional vector, $\mathbf{x} = (x_1, x_2, \ldots, x_K)^T$, and the outputs of N neurons are denoted by a vector, $\mathbf{r} = (r_1, r_2, \ldots, r_N)^T$. In this article, we denote random variables by uppercase letters (e.g., random variables X and R) in contrast to their vector values $\mathbf{x}$ and $\mathbf{r}$. The MI $I(X; R)$ (denoted as I below) between X and R is defined by Cover and Thomas (2006):

$$I = \int_{\mathcal{X}} \int_{\mathcal{R}} p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x}) \ln \frac{p(\mathbf{r} \mid \mathbf{x})}{p(\mathbf{r})} d\mathbf{r} d\mathbf{x}, \quad (2.1)$$

where $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^K$, $\mathbf{r} \in \mathcal{R} \subseteq \mathbb{R}^N$, $d\mathbf{x} = \prod_{k=1}^{K} dx_k$, $d\mathbf{r} = \prod_{n=1}^{N} dr_n$, and the integration symbol $\int$ is for the continuous variables and can be replaced by the summation symbol $\sum$ for discrete variables. The probability density function (p.d.f.) of $\mathbf{r}$, $p(\mathbf{r})$, satisfies

$$p(\mathbf{r}) = \int_{\mathcal{X}} p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x}. \quad (2.2)$$

The MI I in equation 2.1 may also be expressed equivalently as

$$I = H(X) - \left\langle \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x})} \right\rangle_{\mathbf{r}, \mathbf{x}} = H(X) - H(X \mid R), \quad (2.3)$$

where $H(X)$ is the entropy of random variable X,

$$H(X) = -\langle \ln p(\mathbf{x}) \rangle_{\mathbf{x}}, H(X \mid R) = -\langle \ln p(\mathbf{x} \mid \mathbf{r}) \rangle_{\mathbf{r}, \mathbf{x}}, \quad (2.4)$$

and $\langle \cdot \rangle$ denotes expectation:

$$\left\langle \cdot \right\rangle_{\mathbf{x}} = \int_{\mathcal{X}} p(\mathbf{x})(\cdot) d\mathbf{x}, \quad (2.5)$$

$$\left\langle \cdot \right\rangle_{\mathbf{r} \mid \mathbf{x}} = \int_{\mathcal{R}} p(\mathbf{r} \mid \mathbf{x})(\cdot) d\mathbf{r}, \quad (2.6)$$

$$\left\langle \cdot \right\rangle_{\mathbf{r}, \mathbf{x}} = \int_{\mathcal{X}} \int_{\mathcal{R}} p(\mathbf{r}, \mathbf{x})(\cdot) d\mathbf{r} d\mathbf{x}. \quad (2.7)$$

Next, we introduce the following notations,

$$l(\mathbf{r} \mid \mathbf{x}) = \ln p(\mathbf{r} \mid \mathbf{x}), \quad (2.8)$$

$$L(\mathbf{r} \mid \mathbf{x}) = \ln (p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x})), \quad (2.9)$$

$$q(\mathbf{x}) = \ln p(\mathbf{x}), \quad (2.10)$$

and

$$I_F = \frac{1}{2} \left\langle \ln \left(\det \left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e} \right) \right) \right\rangle_{\mathbf{x}} + H(X), \quad (2.11)$$

$$I_G = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X), \quad (2.12)$$

where det (·) denotes the matrix determinant, and

$$\mathbf{J}(\mathbf{x}) = \left\langle l'(\mathbf{r} \mid \mathbf{x}) l'(\mathbf{r} \mid \mathbf{x})^T \right\rangle_{\mathbf{r} \mid \mathbf{x}}, \quad (2.13)$$

$$\mathbf{G}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x}), \quad (2.14)$$

$$\mathbf{P}(\mathbf{x}) = -q''(\mathbf{x}). \quad (2.15)$$

Here **J**(**x**) is the FI matrix, which is symmetric and positive-semidefinite, and $'$ and $''$ denote the first and second derivative for **x**, respectively; that is, $l'(\mathbf{r}|\mathbf{x}) = \partial l(\mathbf{r}|\mathbf{x})/\partial \mathbf{x}$ and $q''(\mathbf{r}|\mathbf{x}) = \partial^2 \ln p(\mathbf{x})/\partial \mathbf{x} \partial \mathbf{x}^T$. If $p(\mathbf{r}|\mathbf{x})$ is twice differentiable for **x**, then

$$\mathbf{J}(\mathbf{x}) = \left\langle l'(\mathbf{r} \mid \mathbf{x}) l'(\mathbf{r} \mid \mathbf{x})^T \right\rangle_{\mathbf{r} \mid \mathbf{x}} = -\left\langle l''(\mathbf{r} \mid \mathbf{x}) \right\rangle_{\mathbf{r} \mid \mathbf{x}}. \quad (2.16)$$

We denote the Kullback-Leibler (KL) divergence as

$$D(\mathbf{x} \| \hat{\mathbf{x}}) = \int_{\mathscr{R}} p(\mathbf{r} \mid \mathbf{x}) \ln \frac{p(\mathbf{r} \mid \mathbf{x})}{p(\mathbf{r} \mid \hat{\mathbf{x}})} d\mathbf{r}, \quad (2.17)$$

and define

$$\mathscr{X}_{\omega}(\mathbf{x}) = \left\{ \check{\mathbf{x}} \in \mathbb{R}^K : (\check{\mathbf{x}} - \mathbf{x})^T \mathbf{G}(\mathbf{x}) (\check{\mathbf{x}} - \mathbf{x}) < N\omega^2 \right\}, \quad (2.18)$$

as the ω neighborhoods of **x** and its complementary set as

$$\bar{\mathscr{X}}_{\omega}(\mathbf{x}) = \mathscr{X} - \mathscr{X}_{\omega}(\mathbf{x}), \quad (2.19)$$

where ω is a positive number.

2.2 Information-Theoretic Asymptotic Bounds and Approximations.

In a large N limit, Brunei and Nadal (1998) proposed an asymptotic relationship $I \sim I_F$ between MI and FI and gave a proof in the case of one-dimensional input. Another proof is

given by Sompolinsky et al. (2001), although there appears to be an error in their proof when a replica trick is used (see equation B1 in their paper; their equation B5 does not follow directly from the replica trick). For large but finite N, $I \simeq I_F$ is usually a good approximation as long as the inputs are low dimensional. For the high-dimensional inputs, the approximation may no longer be valid. For example, suppose $p(\mathbf{r}|\mathbf{x})$ is a normal distribution with mean $\mathbf{A}^T\mathbf{x}$ and covariance matrix $\mathbf{I}_N$ and $p(\mathbf{x})$ is a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\sum$,

$$p(\mathbf{r} \mid \mathbf{x}) = \mathcal{N}(\mathbf{A}^T\mathbf{x}, \mathbf{I}_N), \quad p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (2.20)$$

where $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_N]$ is a deterministic $K \times N$ matrix and $\mathbf{I}_N$ is the $N \times N$ identity matrix. The MI I is given by (see Verdu, 1986; Guo, Shamai, & Verdu, 2005, for details)

$$I = \frac{1}{2}\ln\left(\det\left(\boldsymbol{\Sigma}^{1/2}\mathbf{A}\mathbf{A}^T\boldsymbol{\Sigma}^{1/2} + \mathbf{I}_K\right)\right). \quad (2.21)$$

If rank $(\mathbf{J}(x)) < K$, then $I_F = -\infty$. Notice that here, $\mathbf{J}(\mathbf{x}) = \mathbf{A}\mathbf{A}^T$. When $\mathbf{a} = \mathbf{a}_1 = \ldots = \mathbf{a}_N$ and $\sum = \mathbf{I}_k$, then by equation 2.21 and a matrix determinant lemma, we have

$$I = \frac{1}{2}\ln\left(\det\left(N\mathbf{a}\mathbf{a}^T + \mathbf{I}_K\right)\right) = \frac{1}{2}\ln\left(N\mathbf{a}^T\mathbf{a} + 1\right) \geq 0, \quad (2.22)$$

and by equation 2.11,

$$I_F = \frac{1}{2}\ln\left(\det\left(N\mathbf{a}\mathbf{a}^T\right)\right) = -\infty, \quad (2.23)$$

which is obviously incorrect as an approximation to I. For high-dimensional inputs, the determinant det $(\mathbf{J}(\mathbf{x}))$ may become close to zero in practical applications. When the FI matrix $\mathbf{J}(\mathbf{x})$ becomes degenerate, the regularity condition ensuring the Cramér-Rao paradigm of statistics is violated (Amari & Nakahara, 2005), in which case using I_F as a proxy for I incurs large errors.

In the following, we will show that I_G is a better approximation of I for high-dimensional inputs. For instance, for the above example, we can verify that

$$\begin{aligned} I_G &= \frac{1}{2}\ln\left(\det\left(\frac{1}{2\pi e}(\mathbf{A}\mathbf{A}^T + \boldsymbol{\Sigma}^{-1})\right)\right) + \frac{1}{2}\ln\left(\det\left(2\pi e\boldsymbol{\Sigma}\right)\right) \\ &= \frac{1}{2}\ln\left(\det\left(\boldsymbol{\Sigma}^{1/2}\mathbf{A}\mathbf{A}^T\boldsymbol{\Sigma}^{1/2} + \mathbf{I}_K\right)\right) = I, \end{aligned} \quad (2.24)$$

which is exactly equal to the MI I given in equation 2.21.

2.2.1 Regularity Conditions.—First, we consider the following regularity conditions for $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$:

C1: $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$ are twice continuously differentiable for almost every $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ is a convex set; $\mathbf{G}(\mathbf{x})$ is positive definite, and $\| \mathbf{G}^{-1}(\mathbf{x})\| = O(N^{-1})$, where $\|\cdot\|$ denotes the Frobenius norm of a matrix. The following conditions hold:

$$\|q'(\mathbf{x})\| < \infty, \quad (2.25a)$$

$$\|q''(\mathbf{x})\| < \infty, \quad (2.25b)$$

$$\left\langle \left(N^{-1} l'(\mathbf{r} \mid \mathbf{x})^T l'(\mathbf{r} \mid \mathbf{x})\right)^2 \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(1), \quad (2.25c)$$

$$\left\langle \left\| N^{-1}\left(l''(\mathbf{r} \mid \mathbf{x}) - \langle l''(\mathbf{r} \mid \mathbf{x})\rangle_{\mathbf{r} \mid \mathbf{x}}\right)\right\|^2 \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(N^{-1}), \quad (2.25d)$$

and there exists an $\omega = \omega(\mathbf{x}) > 0$ for $\forall \check{\mathbf{x}} \in \mathcal{X}_{\omega}(\mathbf{x})$ such that

$$N^{-1}\|l''(\mathbf{r} \mid \check{\mathbf{x}}) - l''(\mathbf{r} \mid \mathbf{x})\| = O(1), \quad (2.25e)$$

where O indicates the big-O notation.

C2: The following condition is satisfied,

$$\left\langle \left\| N^{-1}\left(l''(\mathbf{r} \mid \mathbf{x}) - \langle l''(\mathbf{r} \mid \mathbf{x})\rangle_{\mathbf{r} \mid \mathbf{x}}\right)\right\|^{2(m+1)} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(N^{-1}), \quad (2.26a)$$

for $m \in \mathbb{N}$, and there exists $\eta > 1$ such that

$$\mathbb{P}_{\mathbf{r} \mid \mathbf{x}}\left\{ \det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})} p(\hat{\mathbf{x}} \mid \mathbf{r}) d\hat{\mathbf{x}} > \epsilon\, p(\mathbf{x} \mid \mathbf{r}) \right\} = O(N^{-\eta}) \quad (2.26b)$$

for all $\epsilon \in (0, 1/2)$, $\hat{\omega} \in (0, \omega)$ and $\mathbf{x} \in \mathcal{X}$ with $p(\mathbf{x}) > 0$, where $\mathbb{P}_{\mathbf{r} \mid \mathbf{x}}\{\cdot\}$ denotes the probability of $\mathbf{r}$ given $\mathbf{x}$.

The regularity conditions C1 and C2 are needed to prove theorems in later sections. They are expressed in mathematical forms that are convenient for our proofs, although their meanings

may seem opaque at first glance. In the following, we will examine these conditions more closely. We will use specific examples to make interpretations of these conditions more transparent.

Remark 1. In this article, we assume that the probability distributions $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$ are piecewise twice continuously differentiable. This is because we need to use Fisher information to approximate mutual information, and Fisher information requires derivatives that make sense only for continuous variables. Therefore, the methods developed in this article apply only to continuous input variables or stimulus variables. For discrete input variables, we need alternative methods for approximating MI, which we will address in a separate publication.

Conditions 2.25a and 2.25b state that the first and the second derivatives of $q(\mathbf{x}) = \ln p(\mathbf{x})$ have finite values for any given $\mathbf{x} \in \mathcal{X}$. These two conditions are easily satisfied by commonly encountered probability distributions because they only require finite derivatives within $\mathcal{X}$, the set of allowable inputs, and derivatives do not need to be finitely bounded.

Remark 2. Conditions 2.25c to 2.26a constrain how the first and the second derivatives of $l(\mathbf{r}|\mathbf{x}) = \ln p(\mathbf{r}|\mathbf{x})$ scale with N, the number of neurons. These conditions are easily met when $p(\mathbf{r}|\mathbf{x})$ is conditionally independent or when the noises of different neurons are independent, that is, $p(\mathbf{r} \mid \mathbf{x}) = \prod_{n=1}^{N} p(r_n \mid \mathbf{x})$.

We emphasize that it is possible to satisfy these conditions even when $p(\mathbf{r}|\mathbf{x})$ is not independent or when the noises are correlated, as we show later. Here we first examine these conditions closely, assuming independence. For simplicity, our demonstration that follows is based on a one-dimensional input variable ($K = 1$). The conclusions are readily generalizable to higher-dimensional inputs ($K > 1$) because K is fixed and does not affect the scaling with N.

Assuming independence, we have $l(\mathbf{r} \mid x) = \sum_{n=1}^{N} l(r_n \mid x)$ with $l(r_n|x) = \ln p(r_n|x)$, and the left-hand side of equation 2.25c becomes

$$\begin{aligned} & N^{-2}\left\langle l'(\mathbf{r} \mid x)^4 \right\rangle_{\mathbf{r} \mid x} \\ &= N^{-2} \sum_{n_1,\ldots,n_4=1}^{N} \left\langle l'(r_{n_1} \mid x) l'(r_{n_2} \mid x) l'(r_{n_3} \mid x) l'(r_{n_4} \mid x) \right\rangle_{r_{n_1}, r_{n_2}, r_{n_3}, r_{n_4} \mid x} \\ &= N^{-2} \left(\sum_{n \neq m} \left\langle l'(r_n \mid x)^2 \right\rangle_{r_n \mid x} \left\langle l'(r_m \mid x)^2 \right\rangle_{r_m \mid x} + \sum_{n=1}^{N} \left\langle l'(r_n \mid x)^4 \right\rangle_{r_n \mid x} \right), \end{aligned} \tag{2.27}$$

where the final result contains only two terms with even numbers of duplicated indices, while all other terms in the expansion vanish because any unmatched or lone index k (from n_1, n_2, n_3, n_4) should yield a vanishing average:

$$\left\langle l'(r_k \mid x)\right\rangle_{r_k \mid x} = \int_{\mathscr{R}} p(r_k \mid x) l'(r_k \mid x) dr_k = \frac{\partial}{\partial x}\left(\int_{\mathscr{R}} p(r_k \mid x) dr_k\right) = 0 . \quad (2.28)$$

Thus, condition 2.25c is satisfied as long as $\langle l'(r_n|x)^2\rangle_{r_n|x}$ and $\langle l'(r_n|x)^4\rangle_{r_n|x}$ are bounded by some finite numbers, say, a and b, respectively, because now equation 2.27 should scale as $N^{-2}(aN(N-1)+bN) = O(1)$. For instance, a gaussian distribution always meets this requirement because the averages of the second and fourth powers are proportional to the second and fourth moments, which are both finite. Note that the argument above works even if $\langle l'(r_n|x)^4\rangle_{r_n|x}$ is not finitely bounded but scales as $O(N)$.

Similarly, under the assumption of independence, the left-hand side of equation 2.25d becomes

$$\begin{aligned}
&N^{-2}\left\langle\left(l''(\mathbf{r} \mid x) - \langle l''(\mathbf{r} \mid x)\rangle_{\mathbf{r} \mid x}\right)^2\right\rangle_{\mathbf{r} \mid x} \\
&= N^{-2}\sum_{n,m=1}^{N}\left\langle\left(l''(r_n \mid x) - \left\langle l''(r_n \mid x)\right\rangle_{r_n \mid x}\right)\left(l''(r_m \mid x) - \left\langle l''(r_m \mid x)\right\rangle_{r_m \mid x)}\right)\right\rangle_{r_n, r_m \mid x} \\
&= N^{-2}\sum_{n=1}^{N}\left\langle\left(l''(r_n \mid x) - \left\langle l''(r_n \mid x)\right\rangle_{r_n \mid x}\right)^2\right\rangle_{r_n \mid x} \\
&= N^{-2}\sum_{n=1}^{N}\left(\left\langle l''(r_n \mid x)^2\right\rangle_{r_n \mid x} - \left\langle l''(r_n \mid x)\right\rangle_{r_n \mid x}^2\right),
\end{aligned} \quad (2.29)$$

where, in the second step, the only remaining terms are the squares, while all other terms in the expansion with $n \neq m$ have vanished because $\langle l''(r_n|x) - \langle l''(r_n|x)\rangle_{r_n|x}\rangle_{r_n|x} = 0$. Thus, condition 2.25d is satisfied as long as $\langle l''(r_n|x)\rangle_{r_n|x}$ and $\langle l''(r_n|x)^2\rangle_{r_n|x}$ are bounded so that equation 2.29 scales as $N^{-2}N = N^{-1}$.

Condition 2.25e is easily satisfied under the assumption of independence. It is easy to show that this condition holds when $l''(r_n|x)$ is bounded.

Condition 2.26a can be examined with similar arguments used for equations 2.27 and 2.29. Assuming independence, we rewrite the left-hand side of equation 2.26a as

$$
\begin{aligned}
&N^{-z}\left\langle\left(l''(\mathbf{r}\mid x)-\langle l''(\mathbf{r}\mid x)\rangle_{\mathbf{r}\mid x}\right)^{z}\right\rangle_{\mathbf{r}\mid x}\\
&=N^{-z}\sum_{n_1,\ldots,n_z=1}^{N}\left\langle\left(l''(r_{n_1}\mid x)-\left\langle l''(r_{n_1}\mid x)\right\rangle_{r_{n_1}\mid x}\right)\right.\\
&\quad\left.\cdots\left(l''(r_1\mid x)-\left\langle l''(r_{n_z}\mid x)\right\rangle_{r_{n_z}\mid x}\right)\right\rangle_{r_{n_z}\mid x} \qquad (2.30)\\
&=N^{-z}\sum_{n_1,\ldots,n_{m+1}=1}^{N}\left\langle\prod_{i=1}^{m+1}\left(l''(r_{n_i}\mid x)-\left\langle l''(r_{n_i}\mid x)\right\rangle_{r_{n_i}\mid x}\right)^{2}\right\rangle_{r_{n_i}\mid x}+\cdots
\end{aligned}
$$

where $z = 2(m + 1) \geq 4$ is an even number. Any term in the expansion with an unmatched index n_k should vanish, as in the cases of equations 2.27 and 2.29. When $\langle l''(r_n|x)\rangle_{r_n|x}$ and $\langle l''(r_n|x)^2\rangle_{r_n|x}$ are bounded, the leading term with respect to scaling with N is the product of squares, as shown at the end of equation 2.30, because all the other nonvanishing terms increase more slowly with N. Thus equation 2.30 should scale as $N^{-z}N^{m+1} = N^{-m-1}$, which trivially satisfies condition 2.26a.

In summary, conditions 2.25c to 2.26a are easy to meet when $p(\mathbf{r}|\mathbf{x})$ is independent. It is sufficient to satisfy these conditions when the averages of the first and second derivatives of $l(\mathbf{r}|\mathbf{x}) = \ln p(\mathbf{r}|\mathbf{x})$, as well as the averages of their powers, are bounded by finite numbers for all the neurons.

Remark 3. For neurons with correlated noises, if there exists an invertible transformation that maps $\mathbf{r}$ to $\tilde{\mathbf{r}}$ such that $p(\tilde{\mathbf{r}} \mid \mathbf{x})$ becomes conditionally independent, then conditions C1 and C2 are easily met in the space of the new variables by the discussion in remark 2. This situation is best illustrated by the familiar example of a population of neurons with correlated noises that obey a multivariate gaussian distribution:

$$p(\mathbf{r}\mid x)=\frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{r}-\mathbf{g})^{T}\boldsymbol{\Sigma}^{-1}(\mathbf{r}-\mathbf{g})\right), \qquad (2.31)$$

where $\sum$ is an $N \times N$ invertible covariance matrix, and $\mathbf{g} = (g_1(x; \boldsymbol{\theta}_1),\ldots, g_N(x; \boldsymbol{\theta}_N))$ describes the mean resp onses with $\boldsymbol{\theta}_n$ being the parameter vector. Using the following transformation,

$$\tilde{\mathbf{r}}=\boldsymbol{\Sigma}^{-1/2}\mathbf{r}=(\tilde{r}_1,\tilde{r}_2,\ldots,\tilde{r}_N)^{T}, \qquad (2.32)$$

$$\tilde{\mathbf{g}} = \mathbf{\Sigma}^{-1/2}\mathbf{g} = (\tilde{g}_1, \tilde{g}_2, \ldots, \tilde{g}_N)^T, \quad (2.33)$$

we obtain the independent distribution:

$$p(\tilde{\mathbf{r}} \mid x) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(\tilde{r}_n - \tilde{g}_n)^2\right). \quad (2.34)$$

In the special case when the correlation coefficient between any pair of neurons is a constant c, $-1 < c < 1$, the noise covariance can be written as

$$\mathbf{\Sigma} = a\left((1-c)\mathbf{I}_N + c\mathbf{u}\mathbf{u}^T\right), \quad (2.35)$$

where $a > 0$ is a constant, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{u} = (1, 1, \ldots, 1)^T \in \mathbb{R}^{N \times 1}$. The desired transformation in equations 2.32 and 2.33 is given explicitly by

$$\mathbf{\Sigma}^{-1/2} = b_0\left(\mathbf{I}_N - b_1\mathbf{u}\mathbf{u}^T\right), \quad (2.36)$$

where

$$b_0 = \frac{1}{\sqrt{a(1-c)}}, \quad b_1 = \frac{1}{N}\left(1 \pm \sqrt{\frac{1-c}{(N-1)c+1}}\right). \quad (2.37)$$

The new response variables defined in equations 2.32 and 2.33 now read:

$$\tilde{r}_n = b_0\left(r_n - b_1 \sum_{m=1}^{N} r_m\right), \quad (2.38)$$

$$\tilde{g}_n = b_0\left(g_n - b_1 \sum_{m=1}^{N} g_m\right). \quad (2.39)$$

Now we have the derivatives:

$$l'(\tilde{r}_n \mid x) = (\tilde{r}_n - \tilde{g}_n)\frac{\partial \tilde{g}_n}{\partial x}, \quad (2.40)$$

$$l''(\tilde{r}_n \mid x) = \left\langle l''(\tilde{r}_n \mid x) \right\rangle_{r_n \mid x} = (\tilde{r}_n - \tilde{g}_n)\frac{\partial^2 \tilde{g}_n}{\partial x^2}, \quad (2.41)$$

where $\partial \tilde{g}_n / \partial x$ and $\partial^2 \tilde{g}_n / \partial x^2$ are finite as long as $\partial g_n / \partial x$ and $\partial^2 g_n / \partial x^2$ are finite. Conditions C1 and C2 are satisfied when the derivatives and their powers are finitely bounded as shown before.

The example above shows explicitly that it is possible to meet conditions C1 and C2 even when the noises of different neurons are correlated. More generally, if a nonlinear transformation exists that maps correlated random variables into independent variables, then by similar argument, conditions C1 and C2 are satisfied when the derivatives of the log likelihood functions and their powers in the new variables are finitely bounded. Even when the desired transformation does not exist or is unknown, it does not necessarily imply that conditions C1 and C2 must be violated.

While the exact mathematical conditions for the existence of the desired transformation are unclear, let us consider a specific example. If a joint probability density function can be morphed smoothly and reversibly into a flat or constant density in a cube (hypercube), which is a special case of an independent distribution, then this morphing is the desired transformation. Here we may replace the flat distribution by any known independent distribution and the argument above should still work. So the desired transformation may exist under rather general conditions.

For correlated random variables, one may use algorithms such as independent component analysis to find an invertible linear mapping that makes the new random variables as independent as possible (Bell & Sejnowski, 1997) or use neural networks to find related nonlinear mappings (Huang & Zhang, 2017). These methods do not directly apply to the problem of testing conditions C1 and C2 because they work for a given network size N and further development is needed to address the scaling behavior in the large network limit $N \to \infty$.

Finally, we note that the value of the MI of the transformed independent variables is the same as the MI of the original correlated variables because of the invariance of MI under invertible transformation of marginal variables. A related discussion is in theorem 3, which involves a transformation of the input variables rather than a transformation of the output variables as needed here.

Remark 4. Condition 2.26b is satisfied if a positive number δ and a positive integer m exist such that

$$\det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})} \int_{\mathcal{B}_{m,\delta}(\mathbf{x})} p(\mathbf{r} \mid \hat{\mathbf{x}}) p(\hat{\mathbf{x}}) d\mathbf{r} d\hat{\mathbf{x}} = O(N^{-\eta}) \quad (2.42)$$

for all $\hat{\mathbf{x}} \in \bar{\mathscr{X}}_{\hat{\omega}}(\mathbf{x})$, where

$$\mathscr{B}_{m,\delta}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathscr{R} \colon -\delta N^{\frac{\eta-1}{2m}} \mathbf{G}(\mathbf{x}) < l''(\mathbf{r} \mid \mathbf{x}) - \left\langle l''(\mathbf{r} \mid \mathbf{x}) \right\rangle_{\mathbf{r} \mid \mathbf{x}} < \delta N^{\frac{\eta-1}{2m}} \mathbf{G}(\mathbf{x}) \right\}, \quad (2.43)$$

and $\mathbf{A} < \mathbf{B}$ means that the matrix $\mathbf{A} - \mathbf{B}$ is negative definite. A proof is as follows.

First note that in equation 2.43, if $\eta \to 1$ or $m \to \infty$, then $N^{\frac{\eta-1}{2m}} \to 1$. Following Markov's inequality, condition C2 and equation A.19 in the appendix, for the complementary set of $\mathscr{B}_{m,\delta}(\mathbf{x})$, $\bar{\mathscr{B}}_{m,\delta}(\mathbf{x})$, we have

$$\begin{aligned} \mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \bar{\mathscr{B}}_{m,\delta}(\mathbf{x}) \right\} &\le \mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \left\| \mathbf{B}_0 \right\|^2 \ge \delta^2 N^{\frac{\eta-1}{m}} \right\} \\ &\le \delta^{-2m} N^{-(\eta-1)} \left\langle \left\| \mathbf{B}_0 \right\|^{2m} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &= O\left(N^{-\eta}\right), \end{aligned} \quad (2.44)$$

where

$$\mathbf{B}_0 = \mathbf{G}^{-1/2}(\mathbf{x}) \left(l''(\mathbf{r} \mid \mathbf{x}) - \left\langle l''(\mathbf{r} \mid \mathbf{x}) \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right) \mathbf{G}^{-1/2}(\mathbf{x}). \quad (2.45)$$

Define the set

$$\mathscr{A}_{\hat{\omega}}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathscr{R} : \int_{\bar{\mathscr{X}}_{\hat{\omega}}(\mathbf{x})} \frac{p(\hat{\mathbf{x}} \mid \mathbf{r})}{p(\mathbf{x} \mid \mathbf{r})} d\hat{\mathbf{x}} > \det\left(\mathbf{G}(\mathbf{x})\right)^{-1/2} \epsilon \right\}. \quad (2.46)$$

Then it follows from Markov's inequality and equation 2.42 that

$$\begin{aligned} &\mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \mathscr{A}_{\hat{\omega}}(\mathbf{x}) \cap \mathscr{B}_{m,\delta}(\mathbf{x}) \right\} \\ &\le \epsilon^{-1} \det\left(\mathbf{G}(\mathbf{x})\right)^{1/2} \int_{\mathscr{B}_{m,\delta}(\mathbf{x})} \int_{\bar{\mathscr{X}}_{\hat{\omega}}(\mathbf{x})} \frac{p(\mathbf{r} \mid \hat{\mathbf{x}}) p(\hat{\mathbf{x}})}{p(\mathbf{x})} d\hat{\mathbf{x}} d\mathbf{r} \\ &= O\left(N^{-\eta}\right). \end{aligned} \quad (2.47)$$

Hence, we get

$$\mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \mathscr{A}_{\hat{\omega}}(\mathbf{x}) \right\} \le \mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \mathscr{A}_{\hat{\omega}}(\mathbf{x}) \cap \mathscr{B}_{m,\delta}(\mathbf{x}) \right\} + \mathbb{P}_{\mathbf{r} \mid \mathbf{x}} \left\{ \bar{\mathscr{B}}_{m,\delta}(\mathbf{x}) \right\} = O\left(N^{-\eta}\right),$$

which yields condition 2.26b.

Condition 2.42 is satisfied if there exists a positive number ς such that

$$\ln \frac{p(\mathbf{r} \mid \mathbf{x})}{p(\mathbf{r} \mid \hat{\mathbf{x}})} \geq N_{\varsigma} \quad (2.48)$$

for all $\hat{\mathbf{x}} \in \bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})$ and $\mathbf{r} \in \mathcal{B}_{m,\delta}(\mathbf{x})$. This is because

$$\begin{aligned} &\det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})} \int_{\mathcal{B}_{m,\delta}(\mathbf{x})} p(\mathbf{r} \mid \hat{\mathbf{x}}) p(\hat{\mathbf{x}}) d\mathbf{r} d\hat{\mathbf{x}} \\ &= \det(\mathbf{G}(\mathbf{x}))^{1/2} \int_{\bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})} p(\hat{\mathbf{x}}) \int_{\mathcal{B}_{m,\delta}(\mathbf{x})} p(\mathbf{r} \mid \mathbf{x}) \exp\left(-\ln \frac{p(\mathbf{r} \mid \mathbf{x})}{p(\mathbf{r} \mid \hat{\mathbf{x}})}\right) d\mathbf{r} d\hat{\mathbf{x}} \quad (2.49) \\ &\leq \det(\mathbf{G}(\mathbf{x}))^{1/2} \exp(-N_{\varsigma}) = O(N^{K/2} e^{-N_{\varsigma}}). \end{aligned}$$

Here notice that $\det(\mathbf{G}(\mathbf{x}))^{1/2} = O(N^{K/2})$ (see equation A.23).

Inequality 2.48 holds if $p(\mathbf{r}|\mathbf{x})$ is conditionally independent, namely, $p(\mathbf{r} \mid \mathbf{x}) = \prod_{n=1}^{N} p(r_n \mid \mathbf{x})$, with

$$\ln \frac{p(r_n \mid \mathbf{x})}{p(r_n \mid \hat{\mathbf{x}})} \geq \varsigma, \forall n = 1, 2, \ldots, N, \quad (2.50)$$

for all $\hat{\mathbf{x}} \in \bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})$ and $\mathbf{r} \in \mathcal{B}_{m,\delta}(\mathbf{x})$. Consider the inequality $\left\langle \ln p(r_n \mid \mathbf{x}) / p(r_n \mid \hat{\mathbf{x}}) \right\rangle_{r_n \mid \mathbf{x}} \geq 0$ where the equality holds when $\mathbf{x} = \hat{\mathbf{x}}$. If there is only one extreme point at $\hat{\mathbf{x}} = \mathbf{x}$ for $\hat{\mathbf{x}} \in \mathcal{X}_{\omega}(\mathbf{x})$, then generally it is easy to find a set $\mathcal{B}_{m,\delta}(\mathbf{x})$ that satisfies equation 2.50, so that equation 2.26b holds.

2.2.2 Asymptotic Bounds and Approximations for Mutual Information.—

Let

$$\xi = N^{-1} \left\langle \left\| \left(l''(\mathbf{r} \mid \mathbf{x}) - \langle l''(\mathbf{r} \mid \mathbf{x}) \rangle_{\mathbf{r} \mid \mathbf{x}} \right) \mathbf{G}^{-1}(\mathbf{x}) l'(\mathbf{r} \mid \mathbf{x}) \right\|^2 \right\rangle_{\mathbf{r} \mid \mathbf{x}}, \quad (2.51)$$

and it follows from conditions C1 and C2 that

$$\begin{aligned}\xi &\le \left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|^2 \left\langle \left\| N^{-1}\left(l''(\mathbf{r}\mid\mathbf{x}) - \langle l''(\mathbf{r}\mid\mathbf{x})\rangle_{\mathbf{r}\mid\mathbf{x}}\right)\right\|^4 \right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2} \\ &\quad \times \left\langle \left(N^{-1}l'(\mathbf{r}\mid\mathbf{x})^T l'(\mathbf{r}\mid\mathbf{x})\right)^2 \right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2} \\ &= O(N^{-1/2}).\end{aligned} \tag{2.52}$$

Moreover, if $p(\mathbf{r}|\mathbf{x})$ is conditionally independent, then by an argument similar to the discussion in remark 2, we can verify that the condition $\xi = O(N^{-1})$ is easily met.

In the following we state several conclusions about the MI; their proofs are given in the appendix.

Lemma 1. *If condition C1 holds, then the MI I has an asymptotic upper bound for integer N,*

$$I \le I_G + O(N^{-1}). \tag{2.53}$$

Moreover, if equations 2.25c and 2.25d are replaced by

$$\left\langle \left| N^{-1}l'(\boldsymbol{r}\mid\boldsymbol{x})^T l'(\boldsymbol{r}\mid\boldsymbol{x}) \right|^{1+\tau} \right\rangle_{\boldsymbol{r}\mid\boldsymbol{x}} = O(1), \tag{2.54a}$$

$$\left\langle \left\| N^{-1}\left(l''(\boldsymbol{r}\mid\boldsymbol{x}) - \langle l''(\boldsymbol{r}\mid\boldsymbol{x})\rangle_{\boldsymbol{r}\mid\boldsymbol{x}}\right)\right\|^2 \right\rangle_{\boldsymbol{r}\mid\boldsymbol{x}} = o(1), \tag{2.54b}$$

for some $\tau \in (0,1)$, *where o indicates the Little-O notation, then the MI has the following asymptotic upper bound for integer N:*

$$I \le I_G + o(1). \tag{2.55}$$

Lemma 2. *If conditions C1 and C2 hold,* $\xi = O(N^{-1})$, *then the MI has an asymptotic lower bound for integer N,*

$$I \ge I_G + O(N^{-1}). \tag{2.56}$$

Moreover, if condition C1 holds but equations 2.25c *and* 2.25d *are replaced by 2.54a and 2.54b, and inequality 2.26b in C2 also holds for* $\eta > 0$, *then the MI has the following asymptotic lower bound for integer N:*

$$I \geq I_G + o(1). \quad (2.57)$$

Theorem 1. *If conditions C1 and C2 hold,* $\xi = O(N^{-1})$, *then the MI has the following asymptotic equality for integer N:*

$$I = I_G + O(N^{-1}). \quad (2.58)$$

For more relaxed conditions, suppose condition C1 holds but equations 2.25c *and* 2.25d *are replaced by 2.54a and 2.54b, and inequality 2.26b in C2 also holds for* $\eta > 0$, *then the MI has an asymptotic equality for integer N:*

$$I = I_G + o(1). \quad (2.59)$$

Theorem 2. *Suppose* $\boldsymbol{J}(\boldsymbol{x})$ *and* $\boldsymbol{G}(\boldsymbol{x})$ *are symmetric and positive-definite. Let*

$$\varsigma = \langle Tr(\boldsymbol{\Psi}(x)) \rangle_x, \quad (2.60)$$

$$\boldsymbol{\Psi}(\boldsymbol{x}) = \boldsymbol{J}^{-1/2}(\boldsymbol{x})\boldsymbol{P}(\boldsymbol{x})\boldsymbol{J}^{-1/2}(\boldsymbol{x}). \quad (2.61)$$

Then

$$I_G \leq I_F + \frac{\varsigma}{2}, \quad (2.62)$$

where $Tr(\cdot)$ *indicating matrix trace; moreover, if* $\boldsymbol{P}(\boldsymbol{x})$ *is positive-semidefinite, then*

$$0 \leq I_G - I_F \leq \frac{\varsigma}{2}. \quad (2.63)$$

But if

$$\varsigma_1 = \langle \|\boldsymbol{\Psi}(\boldsymbol{x})\| \rangle_x = O(N^{-\beta}) \quad (2.64)$$

for some $\beta > 0$, *then*

$$I_G = I_F + O(N^{-\beta}). \quad (2.65)$$

Remark 5. In general, we need only to assume that $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$ are piecewise twice continuously differentiable for $\mathbf{x} \in \mathcal{X}$. In this case, lemmas 1 and 2 and theorem 1 can still be established. For more general cases, such as discrete or continuous inputs, we have also derived a general approximation formula for MI from which we can easily derive formula for I_G (this will be discussed in separate paper).

2.3 Approximations of Mutual Information in Neural Populations with Finite Size.

In the preceding section, we provided several bounds, including both lower and upper bounds, and asymptotic relationships for the true MI in the large N (network size) limit. Now, we discuss effective approximations to the true MI in the case of finite N. Here we consider only the case of continuous inputs (we will discuss the case of discrete inputs in another paper).

Theorem 1 tells us that under suitable conditions, we can use I_G to approximate I for a large but finite N (e.g., $N \gg K$), that is,

$$I \simeq I_G. \quad (2.66)$$

Moreover, by theorem 2, we know that if $\varsigma \approx 0$ with positive-semidefinite $\mathbf{P}(\mathbf{x})$ or $\varsigma_1 \approx 0$ holds (see equations 2.60 and 2.64), then by equations 2.63, 2.65, and 2.66, we have

$$I \simeq I_G \simeq I_F. \quad (2.67)$$

Define

$$\widetilde{\mathbf{G}}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x}) + \mathbf{Q}(\mathbf{x}), \quad (2.68)$$

$$\widetilde{I}_G = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\widetilde{\mathbf{G}}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X), \quad (2.69)$$

where $\widetilde{\mathbf{G}}(\mathbf{x})$ is positive-definite and $\mathbf{Q}(\mathbf{x})$ is a symmetric matrix depending on $\mathbf{x}$ and $\|\mathbf{Q}(\mathbf{x})\| = O(1)$. Suppose $\left\|\widetilde{\mathbf{G}}^{-1}(\mathbf{x})\right\| = O(N^{-1})$. If we replace I_G by $\widetilde{I}_G$ in theorem 1, then we can prove equations 2.58 and 2.59 in a manner similar to the proof of that theorem. Considering a special case where $\|\mathbf{P}(\mathbf{x})\| \to 0$, $\det(\mathbf{J}(\mathbf{x})) = O(1)$ (e.g., rank $(\mathbf{J}(\mathbf{x})) < K$) and $\|\mathbf{G}^{-1}(\mathbf{x})\| \neq O(N^{-1})$, then we can no longer use the asymptotic formulas in theorem 1. However, if we

substitute $\widetilde{\mathbf{G}}(\mathbf{x})$ for $\mathbf{G}(\mathbf{x})$ by choosing an appropriate $\mathbf{Q}(\mathbf{x})$ such that $\widetilde{\mathbf{G}}(\mathbf{x})$ is positive-definite and $\left\|\widetilde{\mathbf{G}}^{-1}(\mathbf{x})\right\| = O(N^{-1})$, then we can use equation 2.58 and 2.59 as the asymptotic formula.

If we assume $\mathbf{G}(\mathbf{x})$ and $\widetilde{\mathbf{G}}(\mathbf{x})$ are positive-definite and

$$\zeta = \left\langle \left\| \mathbf{Q}(\mathbf{x})\widetilde{\mathbf{G}}^{-1}(\mathbf{x}) \right\| \right\rangle_{\mathbf{x}} = O(N^{-\beta}), \beta > 0, \quad (2.70)$$

then similar to the proof of theorem 2, we have

$$\begin{aligned} &\langle \ln(\det(\mathbf{G}(\mathbf{x}))) \rangle_{\mathbf{x}} \\ &= \left\langle \ln\left(\det\left(\widetilde{\mathbf{G}}(\mathbf{x})\right)\right) \right\rangle_{\mathbf{x}} + \left\langle \ln\left(\det\left(\mathbf{I}_K - \mathbf{Q}(\mathbf{x})\widetilde{\mathbf{G}}^{-1}(\mathbf{x})\right)\right) \right\rangle_{\mathbf{x}} \quad (2.71) \\ &= \left\langle \ln\left(\det\left(\widetilde{\mathbf{G}}(\mathbf{x})\right)\right) \right\rangle_{\mathbf{x}} + O(N^{-\beta}) \end{aligned}$$

and

$$\widetilde{I}_G = I_G + O(N^{-\beta}).$$

For large N, we usually have $\widetilde{I}_G \simeq I_G$.

It is more convenient to redefine the following quantities:

$$\mathbf{Q}(\mathbf{x}) = \mathbf{P}_+ - \mathbf{P}(\mathbf{x}), \quad (2.72)$$

$$\mathbf{P}_+ = \left\langle \frac{\partial \ln p(\mathbf{x})}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{x})}{\partial \mathbf{x}^T} \right\rangle_{\mathbf{x}}, \quad (2.73)$$

$$\mathbf{G}_+(\mathbf{x}) = \widetilde{\mathbf{G}}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + \mathbf{P}_+, \quad (2.74)$$

and

$$I_{G_+} = \widetilde{I}_G = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}_+(\mathbf{x})}{2\pi e}\right)\right) \right\rangle_{\mathbf{x}} + H(X). \quad (2.75)$$

Notice that if $p(\mathbf{x})$ is twice differentiable for $\mathbf{x}$ and

$$\int_{\mathcal{X}} \frac{\partial^2 p(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}^T} d\mathbf{x} = \mathbf{0}, \quad (2.76)$$

then

$$\mathbf{P}_+ = \left\langle \mathbf{P}(\mathbf{x}) \right\rangle_{\mathbf{x}} = \left\langle \frac{1}{p(\mathbf{x})} \frac{\partial^2 p(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}^T} \right\rangle_{\mathbf{x}} - \left\langle \frac{\partial^2 \ln p(\mathbf{x})}{\partial \mathbf{x}\, \partial \mathbf{x}^T} \right\rangle_{\mathbf{x}}. \quad (2.77)$$

For example, if $p(\mathbf{x})$ is a normal distribution, $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then

$$\mathbf{P}(\mathbf{x}) = \mathbf{P}_+ = \boldsymbol{\Sigma}^{-1}. \quad (2.78)$$

Similar to the proof of theorem 2, we can prove that

$$0 \le I_{G_+} - I_F \le \frac{\varsigma_+}{2}, \quad (2.79)$$

where

$$\varsigma_+ = \left\langle \mathrm{Tr}\left(\mathbf{P}_+ \mathbf{J}^{-1}(\mathbf{x})\right) \right\rangle_{\mathbf{x}}. \quad (2.80)$$

We find that I_G is often a good approximation of MI I even for relatively small N. However, we cannot guarantee that $\mathbf{P}(\mathbf{x})$ is always positive-semidefinite in equation 2.14, and as a consequence, it may happen that det $(\mathbf{G}(\mathbf{x}))$ is very small for small N, $\mathbf{G}(\mathbf{x})$ is not positive-definite, and ln (det $(\mathbf{G}(\mathbf{x}))$) is not a real number. In this case, I_G is not a good approximation to I but I_{G_+} is still a good approximation. Generally, if $\mathbf{P}(\mathbf{x})$ is always positive-semidefinite, then I_G or I_{G_+} is a better approximation than I_F, especially when $p(\mathbf{x})$ is close to a normal distribution.

In the following, we give an example of 1D inputs. High-dimensional inputs are discussed in section 4.1.

2.3.1 A Numerical Comparison for 1D Stimuli.—Considering the Poisson neuron model (see equation 5.7 in section 5.1 for details), the tuning curve of the nth neuron, $f(x; \theta_n)$, takes the form of circular normal or von Mises distribution

$$f(x; \theta_n) = A \exp\left(-\left(\frac{T}{2\pi\sigma_f} \right)^2 \left(1 - \cos\left(\frac{2\pi}{T}(x - \theta_n) \right) \right) \right), \quad (2.81)$$

where $x \in [-T/2, T/2)$, $\theta_n \in [-T_\theta/2, T_\theta/2]$, $n \in \{1,2, \ldots, N\}$, with $T = \pi$, $T_\theta = 1$, $\sigma_f = 0.5$, and $A = 20$, and the centers $\theta_1, \theta_2, \ldots, \theta_N$ of the N neurons are uniformly distributed on interval $[-T_\theta/2, T_\theta/2]$, that is, $\theta_n = (n-1)\, d_\theta - T_\theta/2$, with $d_\theta = T_\theta/(N-1)$ and $N \geq 2$. Suppose the distribution of 1D continuous input $x\,(K = 1)$ $p(x)$ has the form

$$p(x) = \mathrm{Z}^{-1} \exp\left(-\left(\frac{T}{2\pi\sigma_p}\right)^2 \left(1 - \cos\left(\frac{2\pi}{T}x\right)\right)\right), \quad (2.82)$$

where σ_p is a constant set to $\pi/4$ and Z is the normalization constant. Figure 1A shows graphs of the input distribution $p(x)$ and the tuning curves $f(x; \theta)$ with different centers $\theta = -\pi/4, 0, \pi/4$.

To evaluate the precision of the approximation formulas, we use Monte Carlo (MC) simulation to approximate MI I. For MC simulation, we first sample an input x_j by the distribution $p(x)$, then generate the neural response $\mathbf{r}_j$ by the conditional distribution $p(\mathbf{r}_j|x_j)$, where $j = 1, 2, \ldots, j_{\max}$. The value of MI by MC simulation is calculated by

$$I^*_{MC} = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln\left(\frac{p(\mathbf{r}_j \mid x_j)}{p(\mathbf{r}_j)}\right), \quad (2.83)$$

where $p(\mathbf{r}_j)$ is given by

$$p(\mathbf{r}_j) = \sum_{m=1}^{M} p(\mathbf{r}_j \mid x_m) p(x_m) \quad (2.84)$$

and $x_m = (m-1)\, T/M - T/2$ for $m \in \{1,2, \ldots, M\}$.

To evaluate the accuracy of MC simulation, we compute the standard deviation,

$$I_{std} = \sqrt{\frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} \left(I^i_{MC} - I_{MC}\right)^2}, \quad (2.85)$$

where

$$I^i_{MC} = \frac{1}{j_{\max}} \sum_{j=1}^{j_{\max}} \ln\left(\frac{p(\mathbf{r}_{\Gamma_{j,i}} \mid x_{\Gamma_{j,i}})}{p(\mathbf{r}_{\Gamma_{j,i}})}\right), \quad (2.86)$$

$$I_{MC} = \frac{1}{i_{\max}} \sum_{i=1}^{i_{\max}} I_{MC}^{i}, \quad (2.87)$$

and $\Gamma_{j,i} \in \{1,2, \ldots, j_{\max}\}$ 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 $\{1, 2, \ldots, j_{\max}\}$ by a uniform distribution. Here we set $j_{\max} = 5 \times 10^5$, $i_{\max} = 100$ and $M = 10^3$.

For different $N \in \{2, 3, 4, 6, 10, 14, 20, 30, 50, 100, 200,400, 700, 1000\}$, we compare I_{MC} with I_G, I_{G+}, and I_F, which are illustrated in Figures 1B to 1D. Here we define the relative error of approximation, for example, for I_G, as

$$DI_G = \frac{I_G - I_{MC}}{I_{MC}}, \quad (2.88)$$

and the relative standard deviation

$$DI_{std} = \frac{I_{std}}{I_{MC}}. \quad (2.89)$$

Figure 1B shows how the values of I_{MC}, I_G, I_{G+}, and I_F change with neuron number N, and Figures 1C and 1D show their relative errors and the absolute values of the relative errors with respect to I_{MC}. From Figures 1B to 1D, we can see that the values of I_G, I_{G+}, and I_F are all very close to one another and the absolute values of their relative errors are all very small. The absolute values are less than 1% when $N \geq 10$ and less than 0.1% when $N \geq 100$. However, for the high-dimensional inputs, there will be a big difference between I_G, I_{G+}, and I_F in many cases (see section 4.1 for more details).

3 Statistical Estimators and Neural Population Decoding

Given the neural response $\mathbf{r}$ elicited by the input $\mathbf{x}$, we may infer or estimate the input $\mathbf{x}$ from the response. This procedure is sometimes referred to as decoding from the response. We need to choose an efficient estimator or a function $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{r})$ that maps the response $\mathbf{r}$ to an estimate $\hat{\mathbf{x}}$ of the true stimulus $\mathbf{x}$. The maximum likelihood (ML) estimator defined by

$$\hat{\mathbf{x}}(\mathbf{r}) = \underset{\mathbf{x}}{\operatorname{argmax}} p(\mathbf{r} \mid \mathbf{x}) = \underset{\mathbf{x}}{\operatorname{argmax}} l(\mathbf{r} \mid \mathbf{x}) \quad (3.1)$$

is known to be efficient in large N limit. According to the Cramér-Rao lower bound (Rao, 1945), we have the following relationship between the covariance matrix of any unbiased estimator $\mathbf{\Sigma}_{\hat{\mathbf{x}}}$ and the FI matrix $\mathbf{J}(\mathbf{x})$,

$$\mathbf{\Sigma}_{\hat{\mathbf{x}}} = \left\langle (\hat{\mathbf{x}}(\mathbf{r}) - \mathbf{x})(\hat{\mathbf{x}}(\mathbf{r}) - \mathbf{x})^T \right\rangle_{\mathbf{r} \mid \mathbf{x}} \geq \mathbf{J}^{-1}(\mathbf{x}), \quad (3.2)$$

where $\hat{\mathbf{x}}(\mathbf{r})$ is an unbiased estimation of $\mathbf{x}$ from the response $\mathbf{r}$, and $\mathbf{A} \geq \mathbf{B}$ means that matrix $\mathbf{A} - \mathbf{B}$ is positive-semidefinite. Thus,

$$\begin{aligned} I_F &= \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X) \\ &\geq \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{\Sigma}_{\hat{\mathbf{x}}}^{-1}}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X) = I_{\text{var}}. \end{aligned} \quad (3.3)$$

The MI between X and $\hat{X}$ is given by

$$\hat{I} = H(\hat{X}) - \left\langle H(\hat{X} \mid X)\right\rangle_{\hat{\mathbf{x}}, \mathbf{x}}, \quad (3.4)$$

where $H(\hat{X})$ is the entropy of random variable $\hat{X}$ and $H(\hat{X} \mid X)$ is its conditional entropy of random variable $\hat{X}$ given X. Since the maximum entropy probability distribution is gaussian, $H(\hat{X} \mid X)$ satisfies

$$H(\hat{X} \mid X) \leq \frac{1}{2}\ln\left(\det(2\pi e \mathbf{\Sigma}_{\hat{\mathbf{x}}})\right). \quad (3.5)$$

Therefore, from equations 3.4 and 3.5, we get

$$\hat{I} \geq \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{\Sigma}_{\hat{\mathbf{x}}}^{-1}}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(\hat{X}) = \hat{I}_{\text{var}}. \quad (3.6)$$

The data processing inequality (Cover & Thomas, 2006) states that postprocessing cannot increase information, so that we have

$$I \geq \hat{I} \geq \hat{I}_{\text{var}}. \quad (3.7)$$

Here we can not directly obtain $I \geq I_F$ as in Brunel and Nadal (1998) when $H(\hat{X}) = H(X)$ and $I_{\text{var}} = \hat{I}_{\text{var}}$. The simulation results in Figure 1 also show that I_F is not a lower bound of I.

For biased estimators, the van Trees' Bayesian Cramér-Rao bound (Van Trees & Bell, 2007) provides a lower bound:

$$\left\langle \boldsymbol{\Sigma}_{\hat{\mathbf{x}}} \right\rangle_{\mathbf{x}} = \left\langle \left\langle (\hat{\mathbf{x}}(\mathbf{r}) - \mathbf{x})(\hat{\mathbf{x}}(\mathbf{r}) - \mathbf{x})^T \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \geq (\langle \mathbf{J}(\mathbf{x}) \rangle_{\mathbf{x}} + \mathbf{P}_+)^{-1} = \left\langle \mathbf{G}_+(\mathbf{x}) \right\rangle_{\mathbf{x}}^{-1}. \quad (3.8)$$

It follows from equations 2.75, 3.6, and 3.8 that

$$I_{G_+} \leq \frac{1}{2} \ln \left(\det \left(\frac{\left\langle \mathbf{G}_+(\mathbf{x}) \right\rangle_{\mathbf{x}}}{2\pi e} \right) \right) + H(X) = I_{VT}, \quad (3.9)$$

$$I_{VT} \geq \frac{1}{2} \ln \left(\det \left(\frac{\left\langle \boldsymbol{\Sigma}_{\hat{\mathbf{x}}} \right\rangle_{\mathbf{x}}^{-1}}{2\pi e} \right) \right) + H(X) = \widetilde{I}_{\text{var}}, \quad (3.10)$$

$$I_{\text{var}} \geq \widetilde{I}_{\text{var}}. \quad (3.11)$$

We may also regard decoding as Bayesian inference. By Bayes' rule,

$$p(\mathbf{x} \mid \mathbf{r}) = \frac{p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x})}{p(\mathbf{r})}. \quad (3.12)$$

According to the Bayesian decision theory, if we know the response **r**, from the prior $p(\mathbf{x})$ and the likelihood $p(\mathbf{r}|\mathbf{x})$, we can infer an estimation of the true stimulus **x**, $\hat{\mathbf{x}}(\mathbf{r})$—for example,

$$\hat{\mathbf{x}}(\mathbf{r}) = \underset{\mathbf{x}}{\operatorname{argmax}} p(\mathbf{x} \mid \mathbf{r}) = \underset{\mathbf{x}}{\operatorname{argmax}} L(\mathbf{r} \mid \mathbf{x}), \quad (3.13)$$

which is also called maximum a posteriori (MAP) estimation.

Consider a loss function $\varphi(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{x})$ for estimation,

$$\varphi(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{x}) = -\ln p(\mathbf{x} \mid \mathbf{r}), \quad (3.14)$$

which is minimized when $p(\mathbf{x}|\mathbf{r})$ reaches its maximum. Now the conditional risk is

$$R(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{r}) = \left\langle \varphi(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{x}) \right\rangle_{\mathbf{x} \mid \mathbf{r}}, \quad (3.15)$$

and the overall risk is

$$R_o = \langle R(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{r})\rangle_{\mathbf{r}} = \left\langle \langle \varphi(\hat{\mathbf{x}}(\mathbf{r}) \mid \mathbf{x})\rangle_{\mathbf{x} \mid \mathbf{r}} \right\rangle_{\mathbf{r}} = -\langle \ln p(\mathbf{x} \mid \mathbf{r})\rangle_{\mathbf{x},\mathbf{r}}. \quad (3.16)$$

Then it follows from equations 2.3 and 3.16 that

$$I = \langle \ln p(\mathbf{x} \mid \mathbf{r})\rangle_{\mathbf{r},\mathbf{x}} + H(X) = -R_o + H(X). \quad (3.17)$$

Comparing equations 2.12, 2.66, and 3.17, we find

$$R_o \simeq -\frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}}. \quad (3.18)$$

Hence, maximizing MI I (or I_G) means minimizing the overall risk R_o for a determinate $H(X)$. Therefore, we can get the optimal Bayesian inference via optimizing MI I (or I_G).

By the Cramér-Rao lower bound, we know that the inverse of FI matrix $\mathbf{J}^{-1}(\mathbf{x})$ reflects the accuracy of decoding (see equation 3.2). $\mathbf{P}(\mathbf{x})$ provides some knowledge about the prior distribution $p(\mathbf{x})$; for example, $\mathbf{P}^{-1}(\mathbf{x})$ is the covariance matrix of input $\mathbf{x}$ when $p(\mathbf{x})$ is a normal distribution. $\|\mathbf{P}(\mathbf{x})\|$ is small for a flat prior (poor prior) and large for a sharp prior (good prior). Hence, if the prior $p(\mathbf{x})$ is flat or poor and the knowledge about model is rich, then the MI I is governed by the knowledge of model, which results in a small ς_1 (see equation 2.64) and $I \simeq I_G \simeq I_F$. Otherwise, the prior knowledge has a great influence on MI I, which results in a large ς_1 and $I \simeq I_G \not\simeq I_F$.

4 Variable Transformation and Dimensionality Reduction in Neural Population Coding

For low-dimensional input $\mathbf{x}$ and large N, both I_G are I_F are good approximations of MI I, but for high-dimensional input $\mathbf{x}$, a large value of ς_1 may lead to a large error of I_F, in which case I_G (or I_{G+}) is a better approximation. It is difficult to directly apply the approximation formula $I \simeq I_G$ when we do not have an explicit expression of $p(\mathbf{x})$ or $\mathbf{P}(\mathbf{x})$. For many applications, we do not need to know the exact value of I_G and care only about the value of $\langle \ln(\det(\mathbf{G}(\mathbf{x})))\rangle_{\mathrm{x}}$ (see section 5). From equations 2.12, 2.22, and 2.78, we know that if $p(\mathbf{x})$ is close to a normal distribution, we can easily approximate $\mathbf{P}(\mathbf{x})$ and $H(X)$ to obtain $\langle \ln(\det(\mathbf{G}(\mathbf{x})))\rangle_{\mathrm{x}}$ and I_G. When $p(\mathbf{x})$ is not a normal distribution, we can employ a technique of variable transformation to make it closer to a normal distribution, as discussed below.

4.1 Variable Transformation.

Suppose $\mathbf{T}: \mathcal{X} \to \widetilde{\mathcal{X}}$ is an invertible and differentiable mapping:

$$\widetilde{\mathbf{x}} = \mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x}), \ldots, T_K(\mathbf{x}))^T, \quad (4.1)$$

$\mathbf{x} = \mathbf{T}^{-1}(\tilde{\mathbf{x}})$, and $\tilde{\mathbf{x}} \in \tilde{\mathcal{X}} \subseteq \mathbb{R}^K$. Let $p(\tilde{\mathbf{x}})$ denote the p.d.f. of random variable $\tilde{X}$ and

$$p(\mathbf{r} \mid \tilde{\mathbf{x}}) = \quad p(\mathbf{r} \mid \mathbf{x}) \mid_{\mathbf{x} = \mathbf{T}^{-1}(\tilde{\mathbf{x}})}. \quad (4.2)$$

Then we have the following conclusions, the proofs of which are given in the appendix.

Theorem 3. *The MI is equivariant under the invertible transformations. More specifically, for the above invertible transformation* $\mathbf{T}$*, the MI I(X; R) in* equation 2.1 *is equal to*

$$I(\tilde{X}; R) = \left\langle ln \frac{p(\boldsymbol{r} \mid \tilde{\boldsymbol{x}})}{p(\boldsymbol{r})} \right\rangle_{\boldsymbol{r}, \tilde{\boldsymbol{x}}}. \quad (4.3)$$

Furthermore, suppose $p(\tilde{x})$ *and* $p(r \mid \tilde{x})$ *fulfill the conditions C1, C2 and* $\xi = O(N^{-1})$*. Then we have*

$$I(\tilde{X}; R) = \tilde{I}_G + O(N^{-1}), \quad (4.4)$$

$$\begin{aligned} \tilde{I}_G &= \frac{1}{2} \left\langle ln \left(det \left(\frac{\boldsymbol{G}(\tilde{\boldsymbol{x}})}{2\pi e} \right) \right) \right\rangle_{\tilde{\boldsymbol{x}}} + H(\tilde{X}) \\ &= \frac{1}{2} \left\langle ln \left(det \left(\frac{\boldsymbol{G}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_{x} + H(X) \\ &= I_G, \end{aligned} \quad (4.5)$$

where $H(\tilde{X})$ *is the entropy of random variable* $\tilde{X}$ *and satisfies*

$$H(\tilde{X}) = -\langle ln\, p(\tilde{\mathbf{x}}) \rangle_{\tilde{x}} = H(X) + \langle ln \mid det\, (D\boldsymbol{T}(\boldsymbol{x})) \mid \rangle_{\boldsymbol{x}}, \quad (4.6)$$

and $D\boldsymbol{T}(\mathbf{x})$ *denotes the Jacobian matrix of* $\mathbf{T}(\mathbf{x})$*,*

$$(D\boldsymbol{T}(\boldsymbol{x}))_{i,j} = \frac{\partial T_i(\boldsymbol{x})}{\partial x_j}, \quad \forall i,\ j = 1, 2, \ldots, K. \quad (4.7)$$

Corollary 1. *Suppose* $p(\mathbf{r}|\mathbf{x})$ *is a normal distribution,*

$$p(\boldsymbol{r} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{A}^T \boldsymbol{y}, \boldsymbol{I}_N), \quad (4.8)$$

where $\mathbf{y} = \mathbf{f}(\mathbf{B}^T\mathbf{x}) = (y_1, y_2, \ldots, y_K)^{\mathrm{T}}$, $y_k = f_k(\mathbf{b}_k^T\mathbf{x})$ *for* $k = 1, 2, \ldots, K$, $\mathbf{A}$ *is a deterministic* $K \times N$ *matrix,* $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_{\mathrm{K}}]$ *is a deterministic invertible matrix, and* f_k *is an invertible and differentiable function. If* Y *has also a normal distribution,* $p(\mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_f, \boldsymbol{\Sigma}_f)$, *then*

$$\begin{aligned} I_G &= I_{G_+} = I(X;R) = I(Y;R) \\ &= \frac{1}{2} ln\left(det\left(\frac{1}{2\pi e}\left(\mathbf{A}\mathbf{A}^T + \boldsymbol{\Sigma}_f^{-1}\right)\right)\right) + H(Y) \\ &= \frac{1}{2}\left\langle ln\left(det\left(\frac{1}{2\pi e}(\mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x}))\right)\right)\right\rangle_{\mathbf{x}} + H(X), \end{aligned} \tag{4.9}$$

where

$$H(Y) = \frac{1}{2} ln\left(det\left(2\pi e\, \boldsymbol{\Sigma}_f\right)\right) = H(X) + \langle ln \mid det\left(\mathbf{D}(\mathbf{x})\right) \mid \rangle_{\mathbf{x}}, \tag{4.10}$$

$$\mathbf{D}(\mathbf{x}) = \left(f_1'(\mathbf{b}_1^T\mathbf{x})\mathbf{b}_1,\ f_2'(\mathbf{b}_2^T\mathbf{x})\mathbf{b}_2, \ldots, f_K'(\mathbf{b}_K^T\mathbf{x})\mathbf{b}_K\right)^T, \tag{4.11}$$

$$f_k'(\mathbf{b}_k^T\mathbf{x}) = \left.\frac{\partial f_k(y_k)}{\partial y_k}\right|_{y_k = b_k^T x}, \quad \forall k = 1, 2, \ldots, K. \tag{4.12}$$

Remark 6. From corollary 1 and equation 2.78, we know that the approximation accuracy for $I_G \simeq I(X; R)$ is improved when we employ an invertible transformation on the input random variable X to make the new random variable Y closer to a normal distribution (see section 4.3).

Consider the eigendecompositions of $\mathbf{A}\mathbf{A}^T$ and $\sum\mathbf{f}$ as given by

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}_{\mathbf{A}}\widehat{\boldsymbol{\Sigma}}\mathbf{U}_{\mathbf{A}}^T, \tag{4.13}$$

$$\boldsymbol{\Sigma}_{\mathbf{f}} = \mathbf{U}_{\mathbf{f}}\widetilde{\boldsymbol{\Sigma}}\mathbf{U}_{\mathbf{f}}^T, \tag{4.14}$$

where $\mathbf{U}_{\mathbf{A}}$ and $\mathbf{U}_{\mathbf{f}}$ are $K \times K$ orthogonal matrices; $\widehat{\boldsymbol{\Sigma}} = \mathrm{diag}\,(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots, \hat{\sigma}_K^2)$ and $\widetilde{\boldsymbol{\Sigma}} = \mathrm{diag}\,(\tilde{\sigma}_1^2, \tilde{\sigma}_2^2, \ldots, \tilde{\sigma}_K^2)$ are $K \times K$ eigenvalue matrices; and $\hat{\sigma}_1 \geq \hat{\sigma}_2 \geq \cdots \geq \hat{\sigma}_K > 0$ and $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_K > 0$. Then by equations 2.11 and 4.9, we have

$$\begin{aligned} I_G &= I_{G_+} = I(X;R) = I(Y;R) \\ &= \frac{1}{2}\ln\left(\det\left(\frac{1}{2\pi e}(\mathbf{U_A}\widehat{\mathbf{\Sigma}}\mathbf{U_A}^T + \mathbf{U_f}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{U_f}^T)\right)\right) + H(Y), \end{aligned} \tag{4.15}$$

$$I_F = \frac{1}{2}\ln\left(\det\left(\frac{\widehat{\mathbf{\Sigma}}}{2\pi e}\right)\right) + H(Y), \tag{4.16}$$

and

$$I_F - I_G = -\frac{1}{2}\ln\left(\det\left(\mathbf{I}_K + \widehat{\mathbf{\Sigma}}^{-1/2}\mathbf{U_A}^T\mathbf{U_f}\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{U_f}^T\mathbf{U_A}\widehat{\mathbf{\Sigma}}^{-1/2}\right)\right). \tag{4.17}$$

Now consider two special cases. If $\widetilde{\mathbf{\Sigma}} = \mathbf{I}_K$, then by equation 4.17, we get

$$I_F - I_G = -\frac{1}{2}\sum_{k=1}^{K}\ln\left(1 + \hat{\sigma}_k^{-2}\right). \tag{4.18}$$

If $\mathbf{U_A} = \mathbf{U_f}$, then

$$I_F - I_G = -\frac{1}{2}\sum_{k=1}^{K}\ln\left(1 + \hat{\sigma}_k^{-2}\tilde{\sigma}_k^{-2}\right). \tag{4.19}$$

Here $\mathbf{J}(\mathbf{x}) = \mathbf{U_A}\widehat{\mathbf{\Sigma}}\mathbf{U_A}^T$, $\mathbf{P}^{-1}(\mathbf{x}) = \mathbf{U_f}\widetilde{\mathbf{\Sigma}}\mathbf{U_f}^T$. The FI matrices $\mathbf{J}(\mathbf{x})$ and $\mathbf{P}^{-1}(\mathbf{x})$ become degenerate when $\hat{\sigma}_K^2 \to 0$ and $\tilde{\sigma}_K^2 \to 0$.

From equations 4.18 and 4.19, we see that if either $\mathbf{J}(\mathbf{x})$or $\mathbf{P}^{-1}(\mathbf{x})$ becomes degenerate, then $(I_F - I_G) \to -\infty$. This may happen for high-dimensional stimuli. For a specific example, consider a random matrix $\mathbf{A}$ defined as follows. Here we first generate $K \times N$ elements $A_{k,n}$, ($k = 1, 2, \ldots, K$; $n = 1, 2, \ldots, N$) from a normal distribution $\mathcal{N}(0,1)$. Then each column of matrix $\mathbf{A}$ is normalized by $A_{k,n} \leftarrow A_{k,n} / \sqrt{\sum_{k=1}^{K} A_{k,n}^2}$. We randomly sample M (set to 2×10^4) image patches with size $\omega \times \omega$ from Olshausen's nature image data set (Olshausen & Field, 1996) as the inputs. Each input image patch was centered by subtracting its mean: $\mathbf{x}_m \leftarrow \mathbf{x}_m - \frac{1}{K}\sum_{k=1}^{K} x_{k,m}$. Then let $\mathbf{x}_m \leftarrow \mathbf{x}_m - \frac{1}{M}\sum_{m'=1}^{M}\mathbf{x}_{m'}$ for $\forall m \in \{1, 2, \ldots, M\}$. Define matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M]$ and compute eigendecomposition

$$\frac{1}{M}\mathbf{X}\mathbf{X}^T = \mathbf{U}_\mathbf{x}\breve{\mathbf{\Sigma}}\mathbf{U}_\mathbf{x}^T, \quad (4.20)$$

where $\mathbf{U}_\mathbf{x}$ is a $K \times \mathbf{K}$ orthogonal matrix and $\breve{\mathbf{\Sigma}} = \text{diag}\left(\breve{\sigma}_1^2, \breve{\sigma}_2^2, \ldots, \breve{\sigma}_K^2\right)$ is a $K \times K$ eigenvalue matrix with $\breve{\sigma}_1 \geq \breve{\sigma}_2 \geq \cdots \geq \breve{\sigma}_K > 0$. Define

$$\mathbf{y} = \mathbf{U}_\mathbf{x}^T\mathbf{x}. \quad (4.21)$$

Then

$$\frac{1}{M}\sum_{m=1}^{M}\mathbf{y}_m\mathbf{y}_m^T = \breve{\mathbf{\Sigma}}. \quad (4.22)$$

The distribution of random variable Y can be approximated by a normal distribution (see section 4.3 for more details). When $p(\mathbf{y}) = \mathcal{N}(\breve{\boldsymbol{\mu}}, \breve{\mathbf{\Sigma}})$, we have

$$I_G = I_{G_+} = I(X;R) = I(Y;R), \quad (4.23)$$

$$\begin{aligned} I_G &= \frac{1}{2}\ln\left(\det\left(\frac{1}{2\pi e}(\mathbf{A}\mathbf{A}^T + \breve{\mathbf{\Sigma}}^{-1})\right)\right) + H(Y) \\ &= \frac{1}{2}\ln\left(\det\left(\frac{1}{2\pi e}(\breve{\mathbf{\Sigma}}^{1/2}\mathbf{A}\mathbf{A}^T\breve{\mathbf{\Sigma}}^{1/2} + \mathbf{I}_K)\right)\right), \end{aligned} \quad (4.24)$$

$$I_F = \frac{1}{2}\ln\left(\det\left(\frac{\mathbf{A}\mathbf{A}^T}{2\pi e}\right)\right) + H(Y). \quad (4.25)$$

The error of approximation I_F is given by

$$\begin{aligned} dI_F &= I_F - I(X;R) = I_F - I_G \\ &= -\frac{1}{2}\ln\left(\det\left(\mathbf{I}_K + (\mathbf{A}\mathbf{A}^T)^{-1}\breve{\mathbf{\Sigma}}^{-1}\right)\right), \end{aligned} \quad (4.26)$$

and the relative error for I_F is

$$DI_F = \frac{dI_F}{I_G}. \quad (4.27)$$

Figure 2A shows how the values of I_G and I_F vary with the input dimension $K = \omega \times \omega$ and the number of neurons N (with $\omega = 2, 4, 6, \ldots, 30$ and $N = 10^4, 2 \times 10^4, 5 \times 10^4, 10^5$). The relative error DI_F is shown in Figure 2B. The absolute value of the relative error tends to decrease with N but may grow quite large as K increases. In Figure 2B, the largest absolute value of relative error $|DI_F|$ is greater than 5000%, which occurs when $K = 900$ and $N = 10^4$. Even the smallest $|DI_F|$ is still greater than 80%, which occurs when $K = 100$ and $N = 10^5$. In this example, I_F is a bad approximation of MI I, whereas I_G and I_{G+} are strictly equal to the true MI I across all parameters.

4.2 Dimensionality Reduction for Asymptotic Approximations.

Suppose $\mathbf{x} = (x_1, \ldots x_K)^T$ is partitioned into two sets of components, $\mathbf{x} = (\mathbf{x}_1^T, \mathbf{x}_2^T)^T$ with

$$\mathbf{x}_1 = (x_1, x_2, \ldots, x_{K_1})^T, \quad (4.28)$$

$$\mathbf{x}_2 = (x_{K_1+1}, x_{K_1+2}, \ldots, x_K)^T, \quad (4.29)$$

where $\mathbf{x}_1 \in \mathcal{X}_1 \subseteq \mathbb{R}^{K_1}$, $\mathbf{x}_2 \in \mathcal{X}_2 \subseteq \mathbb{R}^{K_2}$, $K_1 + K_2 = K$, $K \geq 2$, $K_1 \geq 1$ and $K_2 \geq 1$.

Then by Fubini's theorem, the MI I in equation 2.1 can be written as

$$I = \int_{\mathcal{X}_2} \int_{\mathcal{X}_1} \int_{\mathcal{R}} p(\mathbf{r} \mid \mathbf{x}_1, \mathbf{x}_2) p(\mathbf{x}_1, \mathbf{x}_2) \ln \frac{p(\mathbf{r} \mid \mathbf{x}_1, \mathbf{x}_2)}{p(\mathbf{r})} d\mathbf{r}\, d\mathbf{x}_1 d\mathbf{x}_2, \quad (4.30)$$

where $p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{r}|\mathbf{x})$.

First define

$$\mathbf{G}(\mathbf{x}) = \begin{pmatrix} \mathbf{G}_{1,1}(\mathbf{x}) & \mathbf{G}_{1,2}(\mathbf{X}) \\ \mathbf{G}_{2,1}(\mathbf{x}) & \mathbf{G}_{2,2}(\mathbf{x}) \end{pmatrix}, \quad (4.31a)$$

$$\mathbf{G}_{i,j}(\mathbf{x}) = \mathbf{J}_{i,j}(\mathbf{x}) + \mathbf{P}_{i,j}(\mathbf{x}), \quad (4.31b)$$

where $i, j \in \{1, 2\}$, and

$$\mathbf{J}_{i,j}(\mathbf{x}) = \left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}_i} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}_j^T} \right\rangle_{\mathbf{r} \mid \mathbf{x}}, \quad (4.32a)$$

$$\mathbf{P}_{i,j}(\mathbf{x}) = -\frac{\partial^2 \ln p(\mathbf{x})}{\partial \mathbf{x}_i \partial \mathbf{x}_j^T}. \quad (4.32b)$$

Then we have the following results, their proofs are given in the appendix.

Theorem 4. *Suppose matrices* $\boldsymbol{G}(\boldsymbol{x})$, $\boldsymbol{G}_{1,1}(\boldsymbol{x})$, *and* $\boldsymbol{G}_{2,2}(\boldsymbol{x})$ *are positive-definite. If the matrix* $\boldsymbol{A}_{\boldsymbol{x}} \in \mathbb{R}^{K \times K}$ *satisfies*

$$\left| Tr(\langle \boldsymbol{A}_{\boldsymbol{x}} \rangle_{\boldsymbol{x}}) \right| \lll 1 \quad (4.33)$$

with

$$\boldsymbol{A}_{\boldsymbol{x}} = \boldsymbol{G}_{2,2}^{-1/2}(\boldsymbol{x}) \boldsymbol{G}_{2,1}(\boldsymbol{x}) G_{1,1}^{-1}(\boldsymbol{x}) \boldsymbol{G}_{1,2}(\boldsymbol{x}) \boldsymbol{G}_{2,2}^{-1/2}, \quad (4.34)$$

then we have

$$I_G \simeq I_{G_1} \quad (4.35)$$

with strict equality if and only if

$$\boldsymbol{G}_{2,1}(\boldsymbol{x}) \boldsymbol{G}_{1,1}^{-1}(\boldsymbol{x}) \boldsymbol{G}_{1,2}(\boldsymbol{x}) = \mathbf{0}, \quad (4.36)$$

where

$$I_{G_1} = \frac{1}{2} \left\langle ln \left(\det \left(\frac{\boldsymbol{G}_{1,1}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_x + \frac{1}{2} \left\langle ln \left(det \left(\frac{\boldsymbol{G}_{2,2}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_x + H(X). \quad (4.37)$$

Theorem 5. *Suppose matrices* $\boldsymbol{G}(\boldsymbol{x})$, $\boldsymbol{G}_{1,1}(\boldsymbol{x})$ and $\boldsymbol{P}_{2,2}(\boldsymbol{x})$ *are positive-definite. If the matrix* $\boldsymbol{B}_{\boldsymbol{x}} \in \mathbb{R}^{K_2 \times K_2}$ *is positive-semidefinite and satisfies*

$$0 \leq Tr(\langle \boldsymbol{B}_x \rangle_x) \ll 1 \quad (4.38)$$

with

$$\boldsymbol{B}_{\boldsymbol{x}} = \boldsymbol{P}_{2,2}^{-1/2}(\boldsymbol{x})\, \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{P}_{2,2}^{-1/2}(\boldsymbol{x}), \quad (4.39)$$

$$\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{J}_{2,2}(\boldsymbol{x}) - \boldsymbol{G}_{2,1}(\boldsymbol{x}) \boldsymbol{G}_{1,1}^{-1}(\boldsymbol{x})\, \boldsymbol{G}_{1,2}(\boldsymbol{x}), \quad (4.40)$$

then we have

$$I_G \simeq I_{G_2}, \quad (4.41)$$

with strict equality if and only if

$$\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{0}, \quad (4.42)$$

where

$$I_{G_2} = \frac{1}{2}\left\langle ln\left(det\left(\frac{\boldsymbol{G}_{1,1}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_{\boldsymbol{x}} + \frac{1}{2}\left\langle ln\left(det\left(\frac{\boldsymbol{P}_{2,2}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_{\boldsymbol{x}} + H(X). \quad (4.43)$$

Corollary 2. *If the random variables X_1 and X_2 are independent so that $p(\boldsymbol{x}) = p(\boldsymbol{x}_1)p(\boldsymbol{x}_2), p(\boldsymbol{x}_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{\boldsymbol{x}_2})$ is a normal distribution, and $\boldsymbol{G}(\boldsymbol{x})$, $\boldsymbol{G}_{1,1}(\boldsymbol{x})$, $\boldsymbol{P}_{1,1}(\boldsymbol{x})$ and $\boldsymbol{P}_{2,2}(\boldsymbol{x})$ are all positive-definite and satisfy* equation 4.38, *then we have*

$$I_G \simeq I_{G_1'}, \quad (4.44)$$

$$I_{G_1'} = \frac{1}{2}\left\langle ln\left(det\left(\frac{\boldsymbol{G}_{1,1}(\boldsymbol{x})}{2\pi e} \right) \right) \right\rangle_{\boldsymbol{x}} + H(X_1), \quad (4.45)$$

with strict equality if and only if

$$\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{J}_{2,2}(\boldsymbol{x}) - \boldsymbol{J}_{2,1}(\boldsymbol{x})\boldsymbol{G}_{1,1}^{-1}(\boldsymbol{x})\boldsymbol{J}_{1,2}(\boldsymbol{x}) = \mathbf{0}, \quad (4.46)$$

where

$$H(X_1) = -\left\langle ln\, p(\boldsymbol{x}_1)\right\rangle_{\boldsymbol{x}_1}, \quad (4.47a)$$

$$\boldsymbol{G}_{1,1}(\boldsymbol{x}) = \boldsymbol{J}_{1,1}(\boldsymbol{x}) + \boldsymbol{P}_{1,1}(\boldsymbol{x}), \quad (4.47b)$$

$$\boldsymbol{P}_{1,1}(\boldsymbol{x}) = -\frac{\partial^2 ln\, p(\boldsymbol{x}_1)}{\partial \boldsymbol{x}_1\, \partial \boldsymbol{x}_1^T}. \quad (4.47c)$$

Remark 7. Sometimes we are concerned only with calculating the determinant of matrix **G**(**x**) with a given p(**x**). Theorems 3 and 4 provide a dimensionality reduction method for computing **G** (**x**) or det (**G** (**x**)), by which we need only to compute $\mathbf{G}_{1,1}$ (**x**) and $\mathbf{G}_{2,2}$ (**x**) separately. To apply the approximation 4.35, we do not need to strictly require |Tr ($\langle \mathbf{A}_{\mathbf{x}} \rangle_{\mathbf{x}}$)| ≪ 1. Instead we need to require only

$$\left| \mathrm{Tr}\left(\left\langle \mathbf{A}_{\mathbf{x}} \right\rangle_{\mathbf{x}}\right) \right| \ll \left| \left\langle \ln\left(\det\left(\mathbf{G}_{1,1}(\mathbf{x})\right)\det\left(\mathbf{G}_{2,2}(\mathbf{x})\right)\right)\right\rangle_{\mathbf{x}} \right|. \quad (4.48)$$

Similarly, the inequality |Tr ($\langle \mathbf{B}_{\mathbf{x}} \rangle_{\mathbf{x}}$)| ≪ 1 can be substituted by

$$\left| \mathrm{Tr}\left(\left\langle \mathbf{B}_{\mathbf{x}} \right\rangle_{\mathbf{x}}\right) \right| \ll \left| \left\langle \ln\left(\det\left(\mathbf{G}_{1,1}(\mathbf{x})\right)\det\left(\mathbf{P}_{2,2}(\mathbf{x})\right)\right)\right\rangle_{\mathbf{x}} \right|. \quad (4.49)$$

By equation 4.44 and the second mean value theorem for integrals, we get

$$I_{G_1'} = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}_{1,1}(\mathbf{x}_1, \ddot{\mathbf{x}}_2)}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}_1} + H(X_1) \quad (4.50)$$

for some fixed $\ddot{\mathbf{x}}_2 \in \mathcal{X}_2$. When $\|\boldsymbol{\Sigma}_{\mathbf{x}_2}\|$ is small, $\ddot{\mathbf{x}}_2$ should be close to the mean: $\ddot{\mathbf{x}}_2 \approx \boldsymbol{\mu}_2$. It follows from theorem 1 and corollary 2 that the approximate relationship $I \simeq I_{G_1'}$ holds. However, equation 4.50 implies that $I_{G_1'}$ is determined only by the first component $\mathbf{x}_1$. Hence, there is little impact on information transfer by the minor component (i.e., $\mathbf{x}_2$) for the

high-dimensional input $\mathbf{x}$. In other words, the information transfer is mainly determined by the first component $\mathbf{x}_1$, and we can omit the minor component $\mathbf{x}_2$.

4.3 Further Discussion.

Suppose $\mathbf{x}$ is a zero-mean vector; if it is not, then let $\mathbf{x} \leftarrow \mathbf{x} - \langle \mathbf{x} \rangle_{\mathbf{x}}$. The covariance matrix of $\mathbf{x}$ is given by

$$\mathbf{\Sigma}_{\mathbf{x}} = \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\mathbf{x}} = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T, \quad (4.51)$$

where $\mathbf{U}$ is a $K \times K$ orthogonal matrix whose kth column is the eigenvector $\mathbf{u}_k$ of $\sum_{\mathbf{x}}$ and $\sum$ is a diagonal matrix whose diagonal elements are the corresponding eigenvalues—$\mathbf{\Sigma} = \mathrm{diag}\left(\sigma_1^2, \sigma_2^2, \ldots, \sigma_K^2\right)$ with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_K > 0$. With the whitening transformation,

$$\tilde{\mathbf{x}} = \mathbf{\Sigma}^{-1/2}\mathbf{U}^T\mathbf{x}, \quad (4.52)$$

the covariance matrix of $\tilde{\mathbf{x}}$ becomes an identity matrix:

$$\mathbf{\Sigma}_{\tilde{\mathbf{x}}} = \left\langle \tilde{\mathbf{x}}\tilde{\mathbf{x}}^T \right\rangle_{\tilde{\mathbf{x}}} = \mathbf{\Sigma}^{-1/2}\mathbf{U}^T \left\langle \mathbf{x}\mathbf{x}^T \right\rangle_{\mathbf{x}} \mathbf{U}\mathbf{\Sigma}^{-1/2} = \mathbf{I}_K. \quad (4.53)$$

By the central limit theorem, the distribution of random variable $\tilde{X}$ should be closer to a normal distribution than the distribution of the original random variable X; that is, $p(\tilde{\mathbf{x}}) \simeq \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$. Using Laplace's method asymptotic expansion (MacKay, 2003), we get

$$\mathbf{P}(\tilde{\mathbf{x}}) = -\frac{\partial^2 \ln p(\tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}} \partial \tilde{\mathbf{x}}^T} \simeq \mathbf{\Sigma}_{\tilde{\mathbf{x}}}^{-1} = \mathbf{I}_K, \quad (4.54)$$

$$\mathbf{P}_+ = \langle \mathbf{P}(\tilde{\mathbf{x}}) \rangle_{\tilde{\mathbf{x}}} \simeq \mathbf{\Sigma}_{\tilde{\mathbf{x}}}^{-1} = \mathbf{I}_K. \quad (4.55)$$

In principal component analysis (PCA), the data set is modeled by a multivariate gaussian. By a PCA-like whitening transformation equation 4.52, we can use the approximation 4.55 with Laplace's method, which requires only that the peak be close to the mean and the random variable $\tilde{X}$ does not need to be an exact gaussian distribution.

By theorem 3, we have

$$I(\tilde{X}; R) \simeq I_G = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\tilde{\mathbf{x}})}{2\pi e}\right)\right)\right\rangle_{\tilde{\mathbf{x}}} + H(\tilde{X}), \quad (4.56)$$

where

$$\mathbf{G}(\tilde{\mathbf{x}}) = \mathbf{J}(\tilde{\mathbf{x}}) + \mathbf{I}_K, \quad (4.57)$$

$$\mathbf{J}(\tilde{\mathbf{x}}) = \left\langle \frac{\partial \ln p(\mathbf{r} \mid \tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}} \frac{\partial \ln p(\mathbf{r} \mid \tilde{\mathbf{x}})}{\partial \tilde{\mathbf{x}}^T} \right\rangle_{\mathbf{r} \mid \tilde{\mathbf{x}}} \quad (4.58)$$

$$= \mathbf{\Sigma}^{1/2} \mathbf{U}^T \left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}^T} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \mathbf{U} \mathbf{\Sigma}^{1/2} \quad (4.59)$$

$$= \mathbf{\Sigma}^{1/2} \mathbf{U}^T \mathbf{J}(\mathbf{x}) \mathbf{U} \mathbf{\Sigma}^{1/2}, \quad (4.60)$$

$$H(\tilde{X}) = -\left\langle \ln p(\tilde{\mathbf{x}}) \right\rangle_{\tilde{\mathbf{x}}} = H(X) - \frac{1}{2} \ln \left(\det(\mathbf{\Sigma}) \right). \quad (4.61)$$

Given a $K \times K$ orthogonal matrix $\mathbf{B} \in \mathbb{R}^{K \times K}$, we define

$$\mathbf{y} = \mathbf{B}^T \tilde{\mathbf{x}}. \quad (4.62)$$

Then it follows from equations 4.56 to 4.62 that

$$I(Y; R) \simeq I_G = \frac{1}{2} \left\langle \ln \left(\det \left(\frac{\mathbf{G}(\mathbf{y})}{2\pi e} \right) \right) \right\rangle_{\mathbf{y}} + H(Y), \quad (4.63)$$

where

$$\mathbf{G}(\mathbf{y}) = \mathbf{J}(\mathbf{y}) + \mathbf{I}_K, \quad (4.64)$$

$$\mathbf{J}(\mathbf{y}) = \mathbf{B}^T \mathbf{J}(\tilde{\mathbf{x}}) \mathbf{B}, \quad (4.65)$$

$$H(Y) = H(\tilde{X}). \quad (4.66)$$

Suppose $\mathbf{y}$ is partitioned into two sets of components, $\mathbf{y} = (\mathbf{y}_1^T, \mathbf{y}_2^T)^T$ and

$$\mathbf{y}_1 = (y_1, y_2, \ldots, y_{K_1})^T, \quad (4.67)$$

$$\mathbf{y}_2 = (y_{K_1+1}, y_{K_1+2}, \ldots, y_K)^T, \quad (4.68)$$

where $K_1 + K_2 = K$, $K \geq 2$, $K_1 \geq 1$ and $K_2 \geq 1$. Let

$$\mathbf{G}(\mathbf{y}) = \begin{pmatrix} \mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1} & \mathbf{J}_{1,2}(\mathbf{y}) \\ \mathbf{J}_{2,1}(\mathbf{y}) & \mathbf{J}_{2,2}(\mathbf{y}) + \mathbf{I}_{K_2} \end{pmatrix}, \quad (4.69)$$

where

$$\mathbf{J}_{i,j}(\mathbf{y}) = \left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_i} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_j^T} \right\rangle_{\mathbf{r} \mid \mathbf{y}}, \quad \forall i,\ j = 1, 2. \quad (4.70)$$

When $K \gg 1$, suppose we can find an orthogonal matrix $\mathbf{B}$ and K_1 that satisfy condition 4.38 in theorem 5 or condition 4.49—that is,

$$0 \leq \left\langle \mathrm{Tr}\,(\mathbf{B}_{\mathbf{y}}) \right\rangle_{\mathbf{y}} \ll \gamma, \quad (4.71)$$

$$\mathbf{B}_{\mathbf{y}} = \mathbf{J}_{2,2}(\mathbf{y}) - \mathbf{J}_{2,1}(\mathbf{y}) \left(\mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1}\right)^{-1} \mathbf{J}_{1,2}(\mathbf{y}), \quad (4.72)$$

$$\gamma = \left\langle \ln \left(\det \left(\mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1}\right)\right) \right\rangle_{\mathbf{y}}. \quad (4.73)$$

Here matrix $\mathbf{B}_{\mathbf{y}}$ is positive-semidefinite because

$$\mathbf{J}_{2,2}(\mathbf{y}) - \mathbf{J}_{2,1}(\mathbf{y}) \left(\mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1}\right)^{-1} \mathbf{J}_{1,2}(\mathbf{y}) = \left\langle \boldsymbol{\rho}(\mathbf{r} \mid \mathbf{y}) \boldsymbol{\rho}(\mathbf{r} \mid \mathbf{y})^T \right\rangle_{\mathbf{r} \mid \mathbf{y}}, \quad (4.74)$$

where

$$\boldsymbol{\rho}(\mathbf{r} \mid \mathbf{y}) = \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_2} - \mathbf{J}_{2,1}(\mathbf{y}) \left(\mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1}\right)^{-1} \left(\frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_1} + \mathbf{a}(\mathbf{r})\right) \quad (4.75)$$

and $\mathbf{a}(\mathbf{r})$ is a K_1-dimensional random vector that satisfies

$$\left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_2} \mathbf{a}(\mathbf{r})^T \right\rangle_{\mathbf{r} \mid \mathbf{y}} = \left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}_2} \right\rangle_{\mathbf{r} \mid \mathbf{y}} \left\langle \mathbf{a}(\mathbf{r})^T \right\rangle_{\mathbf{r} \mid \mathbf{y}} = \mathbf{0}, \quad (4.76)$$

$$\left\langle \mathbf{a}(\mathbf{r}) \mathbf{a}(\mathbf{r})^T \right\rangle_{\mathbf{r} \mid \mathbf{y}} = \mathbf{I}_{K_1}. \quad (4.77)$$

Assuming that $\mathbf{J}_{1,1}(\mathbf{y})$ is positive-definite, $\left\| \mathbf{J}_{1,1}^{-1}(\mathbf{y}) \right\| = O(N^{-1})$ and $\|\mathbf{J}_{1,2}(\mathbf{y})\| = \|\mathbf{J}_{2,1}(\mathbf{y})\| = O(N)$, we have

$$\left(\mathbf{J}_{1,1}(\mathbf{y}) + \mathbf{I}_{K_1}\right)^{-1} = \mathbf{J}_{1,1}^{-1}(\mathbf{y}) - \mathbf{J}_{1,1}^{-2}(\mathbf{y}) + O\left(\mathbf{J}_{1,1}^{-3}(\mathbf{y})\right) \quad (4.78)$$

and

$$\begin{aligned} \mathrm{Tr}(\mathbf{C}_{\mathbf{x}}) = {} & \mathrm{Tr}\left(\mathbf{J}_{2,2}(\mathbf{y}) - \mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-1}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y})\right) \\ & + \mathrm{Tr}\left(\mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-2}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y})\right) + O(N^{-1}). \end{aligned} \quad (4.79)$$

Hence, if

$$\left| \mathrm{Tr}\left(\mathbf{J}_{2,2}(\mathbf{y}) - \mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-1}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y})\right) \right| \ll \gamma, \quad (4.80)$$

$$\left| \mathrm{Tr}\left(\mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-2}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y})\right) \right| \ll \gamma, \quad (4.81)$$

then equation 4.71 holds. Notice that the matrix $(\mathbf{J}_{2,2}(\mathbf{y}) - \mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-1}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y}))$ is positive-semidefinite, which is similar to equation 4.74 and $0 \leq \mathrm{Tr}(\mathbf{J}_{2,1}(\mathbf{y}) \mathbf{J}_{1,1}^{-1}(\mathbf{y}) \mathbf{J}_{1,2}(\mathbf{y})) \leq \mathrm{Tr}(\mathbf{J}_{2,2}(\mathbf{y}))$. Hence, if

$$\mathrm{Tr}(\mathbf{J}_{2,2}(\mathbf{y})) \ll \gamma, \quad (4.82)$$

then equations 4.80 and 4.81 hold so does equation 4.71.

5 Optimization of Information Transfer in Neural Population Coding –

5.1 Population Density Distribution of Parameters in Neural Populations.

If $p(\mathbf{r}|\mathbf{x})$ is conditional independent, we can write

$$p(\mathbf{r} \mid \mathbf{x}) = \prod_{n=1}^{N} p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n), \quad (5.1)$$

where $\boldsymbol{\theta}_n \in \mathbb{R}^{\tilde{K}}$ denotes a $\tilde{K}$-dimensional vector for parameters of the nth neuron, and $p(r_n|\mathbf{x}; \boldsymbol{\theta}_n)$ is the conditional p.d.f. of the output r_n given $\mathbf{x}$. With the definition in equation 2.13, we have following proposition.

Proposition 1. If $p(\mathbf{r}|\mathbf{x})$ *is conditional independent as in* equation 5.1, *we have*

$$\boldsymbol{J}(\boldsymbol{x}) = N \int_{\Theta} p(\boldsymbol{\theta}) \boldsymbol{S}(\boldsymbol{x}; \boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (5.2)$$

where

$$\boldsymbol{S}(\boldsymbol{x}; \boldsymbol{\theta}) = \int_{\mathfrak{R}} p(r \mid \boldsymbol{x}; \boldsymbol{\theta}) \frac{\partial \ln p(r \mid \boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{x}} \frac{\partial \ln p(r \mid \boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{x}^T} dr, \quad (5.3)$$

$r \in \mathfrak{R} \subseteq \mathbb{R}$, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^{\tilde{K}}$, *and* $p(\boldsymbol{\theta})$ *is the population density function of parameter vector* $\boldsymbol{\theta}$*:*

$$p(\boldsymbol{\theta}) = \frac{1}{N} \sum_{n=1}^{N} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_n), \quad (5.4)$$

with $\delta(\cdot)$ *being the Dirac delta function.*

Proof.

$$\begin{aligned}\mathbf{J}(\mathbf{x}) &= \int_{\mathscr{R}} p(\mathbf{r} \mid \mathbf{x}) \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}^T} d\mathbf{r} \\ &= \sum_{n=1}^{N} \int_{\mathfrak{R}} p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n) \frac{\partial \ln p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n)}{\partial \mathbf{x}} \frac{\partial \ln p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n)}{\partial \mathbf{x}^T} dr_n \\ &= \int_{\Theta} \sum_{n=1}^{N} \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_n) \left(\int_{\mathfrak{R}} p(r \mid \mathbf{x}; \boldsymbol{\theta}) \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T} dr \right) d\boldsymbol{\theta} \\ &= N \int_{\Theta} p(\boldsymbol{\theta}) \mathbf{S}(\mathbf{x}; \boldsymbol{\theta}) d\boldsymbol{\theta}\,. \end{aligned} \tag{5.5}$$

□

Remark 8. Proposition 1 shows that $\boldsymbol{J}(\boldsymbol{x})$ can be regarded as a function of the population density of parameters, $p(\boldsymbol{\theta})$. If the p.d.f. of the input $p(\boldsymbol{x})$ is given, we can find an appropriate $p(\boldsymbol{\theta})$ to maximize MI I.

For neuron model with Poisson spikes, we have

$$p(\mathbf{r} \mid \mathbf{x}) = \prod_{n=1}^{N} p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n), \tag{5.6}$$

$$p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n) = \frac{f(\mathbf{x}; \boldsymbol{\theta}_n)^{r_n}}{r_n!} \exp(-f(\mathbf{x}; \boldsymbol{\theta}_n)), \tag{5.7}$$

where $f(\mathbf{x}; \boldsymbol{\theta}_n)$ is the tuning curve of the nth neuron, $n = 1, 2, \ldots, N$. Now we have

$$\begin{aligned}\mathbf{S}(\mathbf{x}; \boldsymbol{\theta}) &= \int_{\mathfrak{R}} p(r \mid \mathbf{x}; \boldsymbol{\theta}) \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T} dr \\ &= \frac{1}{f(\mathbf{x}; \boldsymbol{\theta})} \frac{\partial f(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial f(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T} \\ &= \frac{\partial g(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial g(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T}, \end{aligned} \tag{5.8}$$

$$g(\mathbf{x}; \boldsymbol{\theta}) = 2\sqrt{f(\mathbf{x}; \boldsymbol{\theta})}\,. \tag{5.9}$$

Similarly, for a neuron response model with gaussian noise, we have

$$p(\mathbf{r} \mid \mathbf{x}) = \prod_{n=1}^{N} p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n), \quad (5.10)$$

$$p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(r_n - f(\mathbf{x}; \boldsymbol{\theta}_n))^2}{2\sigma^2}\right), \quad (5.11)$$

where σ is a constant standard deviation. Now we get

$$\mathbf{S}(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sigma^2} \frac{\partial f(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial f(\mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T}. \quad (5.12)$$

5.2 Optimal Population Distribution for Neural Population Coding.

Suppose $p(\mathbf{x})$ and $p(\mathbf{r}|\mathbf{x})$ fulfill conditions C1 and C2 and equation 5.1. Following the discussion in section 2.2, we define the following objective for maximizing MI I,

$$\text{maximize } I_G[p(\boldsymbol{\theta})] = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X), \quad (5.13)$$

or, equivalently,

$$\text{minimize } Q_G[p(\boldsymbol{\theta})] = -\frac{1}{2}\langle \ln(\det(\mathbf{G}(\mathbf{x})))\rangle_{\mathbf{x}}, \quad (5.14)$$

where

$$\mathbf{G}(\mathbf{x}) = \mathbf{J}(\mathbf{x}) + \mathbf{P}(\mathbf{x}), \quad (5.15)$$

$$\mathbf{J}(\mathbf{x}) = N \int_{\Theta} p(\boldsymbol{\theta}) \mathbf{S}(\mathbf{x}; \boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (5.16)$$

$$\mathbf{S}(\mathbf{x}; \boldsymbol{\theta}) = \left\langle \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}} \frac{\partial \ln p(r \mid \mathbf{x}; \boldsymbol{\theta})}{\partial \mathbf{x}^T} \right\rangle_{r \mid \mathbf{x}; \boldsymbol{\theta}}. \quad (5.17)$$

Here $\mathbf{P}(\mathbf{x})$ is given in equation 2.15, and it generally can be substituted by $\mathbf{P}_+$ (see equation 2.78).

When $\varsigma_1 \approx 0$ (see equation 2.64), the object function, equation 5.13, can be reduced to

$$\text{maximize } I_F[p(\boldsymbol{\theta})] = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X), \quad (5.18)$$

or, equivalently,

$$\text{minimize } Q_F[p(\boldsymbol{\theta})] = -\frac{1}{2}\langle \ln(\det(\mathbf{J}(\mathbf{x})))\rangle_{\mathbf{x}}. \quad (5.19)$$

The constraint condition for $p(\boldsymbol{\theta})$ is given by

$$\text{subject to } \int_{\Theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1, \quad p(\boldsymbol{\theta}) \geq 0. \quad (5.20)$$

However, without further constraints on the neural populations, especially a limit on the peak firing rate, the capacity of the system may grow indefinitely: $I(X; R) \rightarrow \infty$. The most common limitation on neural populations is the energy or power constraint. For neuron models with Poisson noise or gaussian noise, a useful constraint is a limitation on the peak power,

$$| f(\mathbf{x}; \boldsymbol{\theta}_n) | \leq E_{\max}, \quad \forall \mathbf{x} \in \mathcal{X} \quad \text{and} \quad \forall n = 1, 2, \ldots, N, \quad (5.21)$$

where $E_{\max} > 0$ is the peak power. Under this constraint, maximizing $I_G[p(\boldsymbol{\theta})]$ or $I_F[p(\boldsymbol{\theta})]$ for independent neurons will result in $\max_{\mathbf{x}} | f(\mathbf{x}; \boldsymbol{\theta}_n)| = E_{\max}$ for $\forall n = 1,2, \ldots, N$.

Another constraint is a limitation on average power. For Poisson neurons given in equation 5.7,

$$\frac{1}{N}\sum_{n=1}^{N}\left\langle \langle r_n p(r_n \mid \mathbf{x}; \boldsymbol{\theta}_n)\rangle_{r_n \mid \mathbf{x}}\right\rangle_{\mathbf{x}} \leq E_{\text{avg}}, \quad (5.22)$$

which can also be written as

$$\langle \langle f(\mathbf{x}; \boldsymbol{\theta})\rangle_{\mathbf{x}}\rangle_{\boldsymbol{\theta}} \leq E_{\text{avg}}, \quad (5.23)$$

and for gaussian noise neurons given in equation 5.11,

$$\left\langle \left\langle f(\mathbf{x}; \boldsymbol{\theta})^2\right\rangle_{\mathbf{x}}\right\rangle_{\boldsymbol{\theta}} \leq E_{\text{avg}}, \quad (5.24)$$

where $E_{\text{avg}} > 0$ is the maximum average energy cost.

In equation 5.15, we can approximate the continuous integral by a discrete summation for numerical computation,

$$\mathbf{J}(\mathbf{x}) = N \sum_{k=1}^{K_1} \alpha_k \mathbf{S}(\mathbf{x}; \boldsymbol{\theta}_k), \quad (5.25)$$

where the positive integer $K_1 \leq N$ denotes the number of subclasses in the neural population and

$$\sum_{k=1}^{K_1} \alpha_k = 1, \quad \alpha_k > 0, \forall k = 1, 2, \ldots, K_1 . \quad (5.26)$$

If we do not know the specific form of $p(\mathbf{x})$ but have M samples, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$, which are i.i.d. samples drawn from the distribution $p(\mathbf{x})$, then we can approximate the integral in equation 5.13 by the sample average:

$$\langle \ln (\det (\mathbf{G}(\mathbf{x}))) \rangle_{\mathbf{x}} \simeq \frac{1}{M} \sum_{m=1}^{M} \ln (\det (\mathbf{G}(\mathbf{x}_m))) . \quad (5.27)$$

Optimizing the objective 5.13 or 5.18 is a convex optimization problem (see the appendix for a proof).

Proposition 2. *The functions $I_G[p(\boldsymbol{\theta})]$ and $I_F[p(\boldsymbol{\theta})]$ are concave about $p(\boldsymbol{\theta})$.*

Remark 9. For a low-dimensional input $\mathbf{x}$, we may use equation 5.18 or 5.19 as the objective. Since $I_G[p(\boldsymbol{\theta})]$ and $I_F[p(\boldsymbol{\theta})]$ are concave functions of $p(\boldsymbol{\theta})$, we can directly use efficient numerical methods to get the optimal solution for small K. However, for high-dimensional input $\mathbf{x}$, we need to use other methods (e.g., Huang & Zhang, 2017).

5.3 Necessary and Sufficient Conditions for Optimal Population Distribution.

Applying the method of Lagrange multipliers for the optimization problems 5.13 and 5.20 yields

$$L[p(\boldsymbol{\theta})] = I_G[p(\boldsymbol{\theta})] - \lambda_1 \left(\int_{\Theta} p(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \right) + \int_{\Theta} \lambda_2(\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}, \quad (5.28)$$

where λ_1 is a constant and $\lambda_2(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$. According to Karush-Kuhn-Tucker (KKT) conditions (Boyd & Vandenberghe, 2004), we have

$$\lambda_2(\boldsymbol{\theta})p(\boldsymbol{\theta}) = 0, \quad \lambda_2(\boldsymbol{\theta}) \geq 0, \quad (5.29)$$

and the necessary condition for optimal population density,

$$\frac{\partial L[p(\boldsymbol{\theta})]}{\partial\, p(\boldsymbol{\theta})} = \frac{1}{2}\left\langle \mathrm{Tr}\left(N\mathbf{G}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x};\boldsymbol{\theta})\right)\right\rangle_{\mathbf{x}} - \lambda_1 + \lambda_2(\boldsymbol{\theta}) = 0\,. \quad (5.30)$$

It follows from equations 5.29 and 5.30 that

$$\frac{1}{2}\left\langle \mathrm{Tr}\left(N\mathbf{G}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x};\boldsymbol{\theta})\right)\right\rangle_{\mathbf{x}} = \lambda_1, \quad p(\boldsymbol{\theta}) \neq 0, \quad (5.31)$$

$$\frac{1}{2}\left\langle \mathrm{Tr}\left(N\mathbf{G}(\mathbf{x})^{-1}\mathbf{S}(\mathbf{x};\boldsymbol{\theta})\right)\right\rangle_{\mathbf{x}} = \lambda_1 - \lambda_2(\boldsymbol{\theta}), \quad p(\boldsymbol{\theta}) = 0\,. \quad (5.32)$$

Since $I_G[p(\boldsymbol{\theta})]$ is a concave function of $p(\boldsymbol{\theta})$, equations 5.31 and 5.32 are the necessary and sufficient conditions for the optimization problems 5.13 and 5.20.

5.4 Channel Capacity for Neural Population Coding.

If $p(\mathbf{x})$ is unknown, then by Jensen's inequality, we have

$$\begin{aligned} I \simeq I_G[p(\mathbf{x})] &= \int_{\mathcal{X}} p(\mathbf{x}) \ln\left(p(\mathbf{x})^{-1}\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)^{1/2}\right)d\mathbf{x} \\ &\leq \ln \int_{\mathcal{X}} \det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)^{1/2} d\mathbf{x}, \end{aligned} \quad (5.33)$$

and the equality holds if and only if $p(\mathbf{x})^{-1}$ det $(\mathbf{G}(\mathbf{x}))^{1/2}$ is a constant. Thus,

$$I_G[p^*(\mathbf{x})] = \max_{p(\mathbf{x})}(I_G[p(\mathbf{x})]) = \ln \int_{\mathcal{X}} \det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)^{1/2} d\mathbf{x}, \quad (5.34)$$

$$p^*(\mathbf{x}) = \frac{\det(\mathbf{G}(\mathbf{x}))^{1/2}}{\int_X \det(\mathbf{G}(\hat{\mathbf{x}}))^{1/2} d\hat{\mathbf{x}}}, \quad (5.35)$$

assuming $\int_{\mathcal{X}} \det(\mathbf{G}(\hat{\mathbf{x}}))^{1/2} d\hat{\mathbf{x}} < \infty$.

Let us consider a specific example. Suppose $\mathbf{J}(\mathbf{x}) = \mathbf{J}_0$ is a constant matrix; then it follows from equation 2.12 that

$$I_G = \frac{1}{2}\left\langle \ln\left(\det\left(\frac{\mathbf{J}_0 + \mathbf{P}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + H(X). \quad (5.36)$$

According to the maximum entropy probability distribution, we know that maximizing $H(X)$ results in a uniformly distributed $p(\mathbf{x})$. Hence we have $\mathbf{G}(\mathbf{x}) = \mathbf{J}_0$, and $p^*(\mathbf{x})$ coincides with the uniform distribution (see equation 5.35). In this case, the maximum $I_G[p^*(\mathbf{x})]$ can be regarded as the channel capacity for this neural population.

If we consider a constraint on random variables X and assume that the covariance matrix of X is $\sum_0$ and satisfies

$$\mathbf{\Sigma}_0^{-1} = \mathbf{P}(\mathbf{x}), \quad (5.37)$$

then it follows from the maximum entropy probability distribution that

$$H(X) \leq \frac{1}{2}(\det(2\pi e\mathbf{\Sigma}_0)), \quad (5.38)$$

and the equality holds if and only if the p.d.f. of the input is a normal distribution: $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_0)$. Hence,

$$\begin{aligned} I_G &= \frac{1}{2}\ln\left(\det\left(\frac{\mathbf{J}_0 + \mathbf{\Sigma}_0^{-1}}{2\pi e}\right)\right) + H(X) \\ &\leq \frac{1}{2}\ln(\det(\mathbf{\Sigma}_0\mathbf{J}_0 + \mathbf{I}_K)) = I_G[p^*(\mathbf{x})], \end{aligned} \quad (5.39)$$

where $I_G[p^*(\mathbf{x})]$ is the channel capacity of neural population. Here the equality holds if and only if $p^*(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_0)$, which is consistent with equation 5.37.

Furthermore, if $\varsigma_1 \approx 0$ (see equation 2.64), we have

$$I \simeq I_G[p(\mathbf{x})] \simeq I_F[p(\mathbf{x})] = \int_{\mathcal{X}} p(\mathbf{x}) \ln\left(p(\mathbf{x})^{-1}\det\left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e}\right)^{1/2}\right) d\mathbf{x}. \quad (5.40)$$

Similarly, we also get

$$I_F[p^*(\mathbf{x})] = \max_{p(\mathbf{x})}(I_F[p(\mathbf{x})]) = \ln \int_{\mathcal{X}} \det\left(\frac{\mathbf{J}(\mathbf{x})}{2\pi e}\right)^{1/2} d\mathbf{x}, \quad (5.41)$$

$$p^*(\mathbf{x}) = \frac{\det(\mathbf{J}(\mathbf{x}))^{1/2}}{\int_{\mathcal{X}} \det(\mathbf{J}(\hat{\mathbf{x}}))^{1/2} d\hat{\mathbf{x}}}, \quad (5.42)$$

assuming $\int_{\mathcal{X}} \det(\mathbf{J}(\hat{\mathbf{x}}))^{1/2} d\hat{\mathbf{x}} < \infty$. Here $I_F[p^*(\mathbf{x})]$ is the channel capacity of the neural population. The distribution $p^*(\mathbf{x})$ coincides with the Jeffrey's prior in Bayesian probability (Jeffreys, 1961). In this case, if we suppose the covariance matrix of X is $\sum_0$, then similar to equations 5.38 and 5.39, we can get the channel capacity

$$I_F[p^*(\mathbf{x})] = \frac{1}{2}\ln(\det(\mathbf{\Sigma}_0\mathbf{J}_0)) \quad (5.43)$$

with $p^*(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma}_0)$.

For another example, consider the Poisson neuron model given in equation 5.7 and suppose the input x is one dimension, $K = 1$. It follows from equations 5.8 and 5.42 that

$$p^*(x) = \frac{\left(\int_\Theta p(\boldsymbol{\theta})\left(\frac{\partial g(x;\boldsymbol{\theta})}{\partial x}\right)^2 d\boldsymbol{\theta}\right)^{1/2}}{\int_{\mathcal{X}}\left(\int_\Theta p(\boldsymbol{\theta})\left(\frac{\partial g(\hat{x};\boldsymbol{\theta})}{\partial \hat{x}}\right)^2 d\boldsymbol{\theta}\right)^{1/2} d\hat{x}}. \quad (5.44)$$

If $p(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$, equation 5.44 becomes

$$p^*(x) = \frac{\left|\frac{\partial g(x;\boldsymbol{\theta}_0)}{\partial x}\right|}{\int_{\mathcal{X}}\left|\frac{\partial g(\hat{x};\boldsymbol{\theta}_0)}{\partial \hat{x}}\right| d\hat{x}}. \quad (5.45)$$

Atick and Redlich (1990) presented a redundancy measure to approximate Barlow's optimality principle:

$$\mathcal{R} = 1 - \frac{I(X;R)}{C(R)}, \quad (5.46)$$

where $C(R)$ is the channel capacity. Here for neural population coding, we have $C(R) \simeq I_G[p^*(\mathbf{x})]$ and $I(X; R) \simeq I_G$ (or $C(R) - I_F[p^*(\mathbf{x})]$ and $I(X; R) \simeq I_F$). Hence, we can minimize

$\mathscr{R}$ by choosing an appropriate $\mathbf{J}(\mathbf{x})$ to maximize I_G (or I_F) and simultaneously satisfy equation 5.35 (or 5.42) (see Huang & Zhang, 2017, for further details).

6 Discussion

In this article, we have derived several information-theoretic bounds and approximations for effective approximation of MI in the context of neural population coding for large but finite population size. We have found some regularity conditions under which the asymptotic bounds and approximations hold. Generally these regularity conditions are easy to meet. Special examples that satisfy these conditions include the cases when the likelihood function $p(\mathbf{r}|\mathbf{x})$ for the neural population responses is conditionally independent or has correlated noises with a multivariate gaussian distribution. Under the general regularity conditions, we have derived several asymptotic bounds and approximations of MI for a neural population and found some relationships among different approximations.

How to choose among these different asymptotic approximations of MI in a neural population with finite size N? For a flat prior distribution $p(\mathbf{x})$, we have $I_G \simeq I_F$; that is, the two approximations I_G and I_F are about equally valid. For a sharply peaked prior distribution $p(\mathbf{x})$, I_G is generally a better approximation to MI I than I_F. Under suitable conditions (e.g., cases C1 and C2) for low-dimensional inputs, I_G and I_F are good approximations of MI I not only for large N but also for small N. For high-dimensional inputs, the FI matrix $\mathbf{J}(\mathbf{x})$ (see equation 2.11) or matrix $\mathbf{P}^{-1}(\mathbf{x})$ (see equation 2.15) often becomes degenerate, which causes a large error between I_F and MI I. Hence, in this situation, I_G is a better approximation to MI I than I_F. For more convenient computation of the approximation, we have also introduced the approximation formula I_{G+} which may substitute for I_G as a proxy of MI I. For some special cases (see corollary 1), I_G and I_{G+} are strictly equal to the true MI I. Our simulation results for the one-dimensional case show that the approximations I_G, I_{G+}, and I_F are all highly precise compared with the true MI I, even for small N (see Figure 1).

These approximation formulas satisfy additional constraints. By the Cramér-Rao lower bound, we know that I_F is related to the covariance matrix of an unbiased estimator (see equation 3.3). By van Trees' Bayesian Cramér-Rao bound, we get a link between I_{G+} and the covariance matrix of a biased estimator (see equation 3.9). From the point of view of neural population decoding and Bayesian inference, there is a connection between MI (or I_G) and MAP (see equation 3.17).

For more efficient calculation of the approximation I_G (or I_{G+}) for high-dimensional inputs, we propose to apply an invertible transformation on the input variable so as to make the new variable closer to a normal distribution (see section 4.1). Another useful technique is dimensionality reduction, which effectively approximates MI by further reducing the computational complexity for high-dimensional inputs. We found that I_F could lead to huge errors as a proxy of the true MI I for high-dimensional inputs even when I_G and I_{G+} are strictly equal to the true MI I.

These approximation formulas are potentially useful for optimization problems of information transfer in neural population coding. We have proven that optimizing the

population density distribution of parameters $p(\boldsymbol{\theta})$ is a convex optimization problem and have found a set of necessary and sufficient conditions. The approximation formulas are also useful for discussion of the channel capacity of neural population coding (see section 5.4).

Information theory is a powerful tool for neuroscience and other disciplines, including diverse fields such as physics, information and communication technology, machine learning, computer vision, and bioinformatics. Finding effective approximation methods for computing MI is a key for many practical applications of information theory. Generally the FI matrix is easier to evaluate or approximate than MI. This is because calculation of MI involves averaging over both the input variable **x** and the output variable **r** (see equation 2.1), and typically $p(\mathbf{r})$ also needs to be calculated from $p(\mathbf{r}|\mathbf{x})$ by another average over **x** (see equation 2.2). By contrast, the FI matrix $\mathbf{J}(\mathbf{x})$ involves averaging over **r** only (see equation 2.13). Furthermore, it is often easier to find analytical forms of FI for specific models such as a population of tuning curves with Poisson spike statistics. Taking into account the computational efficiency, for practical applications we suggest using I_G or I_{G+} as a proxy of the true MI I for most cases. These approximations could be very useful even when we do not need to know the exact value of MI. For example, for some optimization and learning problems, we only need to know how MI is affected by the conditional p.d.f. or likelihood function $p(\mathbf{r}|\mathbf{x})$. In such situations, we may easily solve for the optimal parameters using the approximation formulas (Huang & Zhang, 2017; Huang, Huang, & Zhang, 2017). Further discussions of the applications will be given in separate publications.

Acknowledgments

This work was supported by an NIH grant R01 DC013698.

Appendix: The Proofs

We consider a Taylor expanding of $L(\mathbf{r} \mid \widehat{\mathbf{x}})$ around **x**. If $L(\mathbf{r} \mid \widehat{\mathbf{x}})$ is twice differentiable for $\forall \widehat{\mathbf{x}} \in \mathcal{X}_{\omega}(\mathbf{x})$, then by condition **C1** we get

$$\begin{aligned} & L(\mathbf{r} \mid \widehat{\mathbf{x}}) - L(\mathbf{r} \mid \mathbf{x}) \\ &= (\widehat{\mathbf{x}} - \mathbf{x})^T L'(\mathbf{r} \mid \mathbf{x}) + \frac{1}{2}(\widehat{\mathbf{x}} - \mathbf{x})^T L''(\mathbf{r} \mid \check{\mathbf{x}})(\widehat{\mathbf{x}} - \mathbf{x}) \\ &= \mathbf{y}^T \widetilde{\mathbf{v}} - \frac{1}{2}\mathbf{y}^T \mathbf{y} + \frac{1}{2}\mathbf{y}^T \mathbf{B}\mathbf{y}, \end{aligned} \tag{A.1}$$

where

$$\mathbf{y} = \mathbf{G}^{1/2}(\mathbf{x})(\widehat{\mathbf{x}} - \mathbf{x}), \tag{A.2}$$

$$\widetilde{\mathbf{v}} = \mathbf{v} + \mathbf{v}_1, \mathbf{v} = \mathbf{G}^{-1/2}(\mathbf{x}) l'(\mathbf{r} \mid \mathbf{x}), \mathbf{v}_1 = \mathbf{G}^{-1/2}(\mathbf{x}) q'(\mathbf{x}), \tag{A.3}$$

$$\check{\mathbf{x}} = \mathbf{x} + t\,(\hat{\mathbf{x}} - \mathbf{x}) \in \mathcal{X}_{\omega}(\mathbf{x}), t \in (0, 1), \quad \text{(A.4)}$$

$$\begin{cases} \mathbf{B} &= \mathbf{G}^{-1/2}(\mathbf{x})\,\mathbf{C}\mathbf{G}^{-1/2}(\mathbf{x}) = \mathbf{B}_0 + \mathbf{B}_1 + \mathbf{B}_2, \\ \mathbf{C} &= \mathbf{C}_0 + \mathbf{C}_1 + \mathbf{C}_2, \end{cases} \quad \text{(A.5)}$$

and

$$\begin{cases} \mathbf{B}_0 &= \mathbf{G}^{-1/2}(\mathbf{x})\,\mathbf{C}_0\mathbf{G}^{-1/2}(\mathbf{x}), \\ \mathbf{B}_1 &= \mathbf{G}^{-1/2}(\mathbf{x})\,\mathbf{C}_1\mathbf{G}^{-1/2}(\mathbf{x}), \\ \mathbf{B}_2 &= \mathbf{G}^{-1/2}(\mathbf{x})\,\mathbf{C}_2\mathbf{G}^{-1/2}(\mathbf{x}), \\ \mathbf{C}_0 &= l''(\mathbf{r} \mid \mathbf{x}) - \langle l''(\mathbf{r} \mid \mathbf{x})\rangle_{\mathbf{r} \mid \mathbf{x}}, \\ \mathbf{C}_1 &= l''(\mathbf{r} \mid \check{\mathbf{x}}) - l''(\mathbf{r} \mid \mathbf{x}), \\ \mathbf{C}_2 &= q''(\check{\mathbf{x}}) - q''(\mathbf{x}). \end{cases} \quad \text{(A.6)}$$

By condition C1, we know that the matrix $\mathbf{B}_1 + \mathbf{B}_2$ is continuous and symmetric for $\check{\mathbf{x}} \in \mathcal{X}_{\omega}$ and $\|\mathbf{B}_1 + \mathbf{B}_2\| = O(1)$. By the definition of continuous functions, we can prove the following: for any $\epsilon \in (0, 1)$, there is an $\varepsilon \in (0, \omega)$ such that for all $\mathcal{Y} \in \mathcal{Y}_{\varepsilon}$,

$$-\epsilon\mathbf{I}_K \le \mathbf{B}_1 + \mathbf{B}_2 \le \epsilon\mathbf{I}_K, \quad \text{(A.7)}$$

where

$$\mathcal{Y}_{\varepsilon} = \left\{\mathbf{y} \in \mathbb{R}^K : \|\mathbf{y}\| < \varepsilon\sqrt{N}\right\}. \quad \text{(A.8)}$$

Hence,

$$\left|\, \mathbf{y}^T(\mathbf{B}_1 + \mathbf{B}_2)\,\mathbf{y} \,\right| < \epsilon\|\mathbf{y}\|^2. \quad \text{(A.9)}$$

Here $\check{\mathbf{x}} = \mathbf{x} + t\mathbf{G}^{-1/2}(\mathbf{x})\mathbf{y}$, ε is a function of $\mathbf{r}$, $\varepsilon = \varepsilon(\mathbf{r}) = O(1)$, and

$$\mathcal{Y}_{\varepsilon} \subseteq \mathcal{Y}_{\omega} = \left\{\mathbf{y} \in \mathbb{R}^K : \|\mathbf{y}\| < \omega\sqrt{N}\right\}. \quad \text{(A.10)}$$

We define the sets

$$\begin{cases} \bar{\mathcal{Y}}_{\varepsilon} = \left\{ \mathbf{y} \in \mathbb{R}^K : \|\mathbf{y}\| \geq \varepsilon\sqrt{N} \right\}, \\ \mathcal{Z}_{\hat{\varepsilon}} = \left\{ \mathbf{z} \in \mathbb{R}^K : |z_k| < \hat{\varepsilon}\sqrt{N/K}, \forall k = 1, 2, \ldots, K \right\}, \\ \bar{\mathcal{Z}}_{\hat{\varepsilon}} = \left\{ \mathbf{z} \in \mathbb{R}^K : |z_k| \geq \hat{\varepsilon}\sqrt{N/K}, \forall k = 1, 2, \ldots, K \right\}, \\ \widetilde{\mathcal{Z}}_{\varepsilon} = \left\{ \mathbf{z} \in \mathbb{R}^K : \left\| \mathbf{z} + \tilde{\mathbf{v}} 1_{\mathcal{R}_{\hat{\varepsilon}}} \right\| < \varepsilon\sqrt{N} \right\}, \end{cases} \quad \text{(A.11)}$$

where

$$\hat{\varepsilon} = \varepsilon / 2, \quad \text{(A.12)}$$

$1_{(\cdot)}$ denotes an indicator random variable,

$$1_{\mathcal{R}_{\hat{\varepsilon}}} = \begin{cases} 1, \mathbf{r} \in \mathcal{R}_{\hat{\varepsilon}}(\mathbf{x}) \\ 0, \mathbf{r} \notin \mathcal{R}_{\hat{\varepsilon}}(\mathbf{x}) \end{cases}, \quad 1_{\bar{\mathcal{R}}_{\hat{\varepsilon}}} = \begin{cases} 1, \mathbf{r} \in \bar{\mathcal{R}}_{\hat{\varepsilon}}(\mathbf{x}) \\ 0, \mathbf{r} \notin \bar{\mathcal{R}}_{\hat{\varepsilon}}(\mathbf{x}) \end{cases}, \quad \text{(A.13)}$$

and

$$\begin{cases} \mathcal{R}_{\hat{\varepsilon}}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathcal{R} : \|\tilde{\mathbf{v}}\| < \hat{\varepsilon}\sqrt{N} \right\}, \\ \bar{\mathcal{R}}_{\hat{\varepsilon}}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathcal{R} : \|\tilde{\mathbf{v}}\| \geq \hat{\varepsilon}\sqrt{N} \right\}. \end{cases} \quad \text{(A.14)}$$

For all $\mathbf{z} \in \mathcal{Z}_{\hat{\varepsilon}}$, we have $\left\| \mathbf{z} + \tilde{\mathbf{v}} 1_{\mathcal{R}_{\hat{\varepsilon}}} \right\|_2 \leq \left\| \mathbf{z} \right\|_2 + \left\| \mathbf{v} 1_{\mathcal{R}_{\hat{\varepsilon}}} \right\|_2 < \varepsilon\sqrt{N};$; then

$$\mathcal{Z}_{\hat{\varepsilon}} \subseteq \widetilde{\mathcal{Z}}_{\varepsilon}. \quad \text{(A.15)}$$

It follows from equations A.3 and A.6 that

$$\langle \mathbf{v} \rangle_{\mathbf{r} \mid \mathbf{x}} = 0, \langle \mathbf{B}_0 \rangle_{\mathbf{r} \mid \mathbf{x}} = 0 \quad \text{(A.16)}$$

and

$$\begin{aligned}\left\langle\left\langle\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} &= \left\langle\left\langle L'(\mathbf{r}\mid\mathbf{x})^T\mathbf{G}^{-1}(\mathbf{x})L'(\mathbf{r}\mid\mathbf{x})\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}}\\ &= \left\langle \mathrm{Tr}\left(\left\langle L'(\mathbf{r}\mid\mathbf{x})L'(\mathbf{r}\mid\mathbf{x})^T\right\rangle_{\mathbf{r}\mid\mathbf{x}}\mathbf{G}^{-1}(\mathbf{x})\right)\right\rangle_{\mathbf{x}}\\ &= K+\zeta\\ &= K+O(N^{-1}),\end{aligned} \tag{A.17}$$

and it follows from condition C1 that

$$\begin{aligned}\zeta &= \left\langle \mathrm{Tr}\left(\frac{1}{p(\mathbf{x})}\frac{\partial^2 p(\mathbf{x})}{\partial\mathbf{x}\,\partial\mathbf{x}^T}\mathbf{G}^{-1}(\mathbf{x})\right)\right\rangle_{\mathbf{x}}\\ &= \left\langle \mathrm{Tr}\left(\left(q'(\mathbf{x})^Tq'(\mathbf{x})+q''(\mathbf{x})\right)\mathbf{G}^{-1}(\mathbf{x})\right)\right\rangle_{\mathbf{x}}\\ &= \left\langle N^{-1}\left(\left\|q'(\mathbf{x})^Tq'(\mathbf{x})\right\|+\left\|q''(\mathbf{x})\right\|\right)\left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|\right\rangle_{\mathbf{x}}\\ &= O(N^{-1}).\end{aligned} \tag{A.18}$$

Combining conditions C1 and C2 and equations A.3, A.4, and A.6, we find

$$\begin{cases}\left\langle\left\|\mathbf{B}_0\right\|^{2m}\right\rangle_{\mathbf{r}\mid\mathbf{x}} \le \left\langle\left\langle\left\|N^{-1}\mathbf{C}_0\right\|^{2m}\left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|^{2m}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} = O(N^{-1}),\\ \left\langle\left\|\mathbf{B}_0\right\|^{2m+1}\right\rangle_{\mathbf{r}\mid\mathbf{x}} \le \left\langle\left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|^{2m+1}\left\langle\left\|N^{-1}\mathbf{C}_0\right\|^2\right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2}\left\langle\left\|N^{-1}\mathbf{C}_0\right\|^{4m}\right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2}\right\rangle_{\mathbf{x}}\\ \qquad\qquad = O(N^{-1}),\\ \left\langle\left\|\mathbf{v}\right\|^{2m_0}\right\rangle_{\mathbf{r}\mid\mathbf{x}} \le \left\langle\,\left|\,N^{-1}l'(\mathbf{r}\mid\mathbf{x})^Tl'(\mathbf{r}\mid\mathbf{x})\,\right|^{m_0}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|^{m_0} = O(1),\\ \left\|\mathbf{v}_1\right\|^{2m_0} \le \left|\,N^{-1}q'(\mathbf{x})^Tq'(\mathbf{x})\,\right|^{m_0}\left\|N\mathbf{G}^{-1}(\mathbf{x})\right\|^{m_0} = O(N^{-m_0}),\end{cases} \tag{A.19}$$

together with the power mean inequality,

$$\begin{aligned}\left\langle\left(\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}\right)^{m_0}\right\rangle_{\mathbf{r}\mid\mathbf{x}} &\le \left\langle\left(\left\|\mathbf{v}\right\|+\left\|\mathbf{v}_1\right\|\right)^{2m_0}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\\ &\le 2^{2m_0-1}\left\langle\left\|\mathbf{v}\right\|^{2m_0}+\left\|\mathbf{v}_1\right\|^{2m_0}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\\ &= O(1),\end{aligned} \tag{A.20}$$

where $m\in\mathbb{N}$, $m_0\in\{1,2\}$. Notice that $\|\mathbf{G}^{-1}(\mathbf{x})\| = O(N^{-1})$. Here we note that for all conformable matrices **A** and **B**,

$$\begin{cases} |\operatorname{Tr}(\mathbf{AB})| \le \|\mathbf{A}\|\,\|\mathbf{B}\|, \\ \|\mathbf{AB}\| \le \|\mathbf{A}\|\,\|\mathbf{B}\|. \end{cases} \tag{A.21}$$

By equation 2.25c, we have

$$\begin{aligned} \operatorname{Tr}\left(N^{-1}\mathbf{J}(\mathbf{x})\right)^2 &= \left\langle N^{-1} l'(\mathbf{r}\mid\mathbf{x})^T l'(\mathbf{r}\mid\mathbf{x})\right\rangle_{\mathbf{r}\mid\mathbf{x}}^2 \\ &\le \left\langle \left(N^{-1} l'(\mathbf{r}\mid\mathbf{x})^T l'(\mathbf{r}\mid\mathbf{x})\right)^2\right\rangle_{\mathbf{r}\mid\mathbf{x}} = O(1). \end{aligned} \tag{A.22}$$

Then it follows from equations 2.25b and A.22 that

$$\det(\mathbf{G}(\mathbf{x})) = O(N^K). \tag{A.23}$$

A.1 Proof of Lemma 1. It follows from equation A.1 that

$$\begin{aligned} \Gamma_\omega &= \left\langle \left\langle \ln \int_{\mathcal{X}_\omega(\mathbf{x})} \exp\left(L(\mathbf{r}\mid\widehat{\mathbf{x}}) - L(\mathbf{r}\mid\mathbf{x})\right) d\widehat{\mathbf{x}} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \right\rangle_{\mathbf{x}} \\ &= -\left\langle \frac{1}{2}\ln\left(\det(\mathbf{G}(\mathbf{x}))\right)\right\rangle_{\mathbf{x}} \\ &\quad + \underbrace{\left\langle \left\langle \ln\left(\int_{\mathcal{Y}_\omega} \exp\left(\mathbf{y}^T\tilde{\mathbf{v}} - \frac{1}{2}\mathbf{y}^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{B}\mathbf{y}\right) d\mathbf{y}\right)\right\rangle_{\mathbf{r}\mid\mathbf{x}} \right\rangle_{\mathbf{x}}}_{\widehat{\Gamma}_\omega}. \end{aligned} \tag{A.24}$$

For $\mathbf{y} \in \mathcal{Y}_\varepsilon$, according to the definitions in equations A.13 and A.14, we have

$$\begin{aligned} |\mathbf{y}^T\tilde{\mathbf{v}} 1_{\bar{\mathcal{R}}_{\hat{\varepsilon}}}| &\le \left\|\mathbf{y}\right\| \left\|\tilde{\mathbf{v}} 1_{\bar{\mathcal{R}}_{\hat{\varepsilon}}}\right\| \\ &\le (N\varepsilon^2)^{1/2} \left\|\tilde{\mathbf{v}} 1_{\bar{\mathcal{R}}_{\hat{\varepsilon}}}\right\| \\ &\le 2\tilde{\mathbf{v}}^T\tilde{\mathbf{v}} 1_{\bar{\mathcal{R}}_{\hat{\varepsilon}}}. \end{aligned} \tag{A.25}$$

Then by condition C1, we get

$$\begin{aligned}\left\langle \tilde{\mathbf{v}}^T\tilde{\mathbf{v}}1_{\bar{\mathscr{R}}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{r}\mid\mathbf{x}} &\le \left\langle \frac{\|\tilde{\mathbf{v}}\|^4}{(\hat{\varepsilon}\sqrt{N})^2}\right\rangle_{\mathbf{r}\mid\mathbf{x}} \\ &\le N^{-1}(\hat{\varepsilon}_0)^{-2}\left\langle \|\tilde{\mathbf{v}}\|^4\right\rangle_{\mathbf{r}\mid\mathbf{x}} = O(N^{-1}),\end{aligned} \tag{A.26}$$

where $\hat{\varepsilon}_0$ is a positive constant and $\hat{\varepsilon}_0 \in [\min \hat{\varepsilon}(\mathbf{r}), \max \hat{\varepsilon}(\mathbf{r})]$. By equations A.9, A.17, and A.24, we get

$$\begin{aligned}\hat{\Gamma}_\omega &\ge \left\langle\left\langle \ln\left(\int_{\mathscr{Y}_\varepsilon} \exp\left(\mathbf{y}^T\tilde{\mathbf{v}} - \frac{1}{2}(1+\epsilon)\mathbf{y}^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{B}_0\mathbf{y}\right)d\mathbf{y}\right)\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\ &\ge \left\langle\left\langle \ln\left(\int_{\mathscr{Z}_{\hat{\varepsilon}}} \exp\left(\frac{1}{2}\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathscr{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)^T\mathbf{B}_0\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathscr{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)\right)\phi_{\hat{\varepsilon}}(\mathbf{z})d\mathbf{z}\right)\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\ &\quad + \left\langle\left\langle \ln(\boldsymbol{\Psi}_{\hat{\varepsilon}}) + \frac{\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}}{2(1+\epsilon)^2} - \frac{5\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}1_{\overline{\mathscr{R}}_{\hat{\varepsilon}}}}{2(1+\epsilon)^2}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\ &\le \frac{1}{2}\left\langle\left\langle\left\langle\left(\int_{\mathscr{Z}_{\hat{\varepsilon}}}\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathscr{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)^T\mathbf{B}_0\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathscr{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)\phi_{\hat{\varepsilon}}(\mathbf{z})d\mathbf{z}\right)\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\ &\quad + \left\langle\left\langle \ln(\boldsymbol{\Psi}_{\hat{\varepsilon}})\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} + \frac{K+\zeta}{2(1+\epsilon)^2} + O(N^{-1}),\end{aligned} \tag{A.27}$$

where $\mathbf{z} = \mathbf{y} - \tilde{\mathbf{v}}1_{\mathscr{R}_{\hat{\varepsilon}}(\mathbf{x})}$, the last step in equation A.27 follows from Jensen's inequality, and

$$\begin{cases}\phi_{\hat{\varepsilon}}(\mathbf{z}) = \boldsymbol{\Psi}_{\hat{\varepsilon}}^{-1}\exp\left(-\frac{1+\epsilon}{2}\mathbf{z}^T\mathbf{z}\right), \\ \boldsymbol{\Psi}_{\hat{\varepsilon}} = \int_{\mathscr{Z}_{\hat{\varepsilon}}} \exp\left(-\frac{1+\epsilon}{2}\mathbf{z}^T\mathbf{z}\right)d\mathbf{z}.\end{cases} \tag{A.28}$$

Integrating by parts yields

$$\left\langle 1_{\bar{\mathscr{Z}}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}}\int_{\bar{\mathscr{Z}}_{\hat{\varepsilon}}}\left(\frac{1+\epsilon}{2\pi}\right)^{K/2}\exp\left(-\frac{1+\epsilon}{2}\mathbf{z}^T\mathbf{z}\right)d\mathbf{z} = O(N^{-K/2}e^{-N\delta}) \tag{A.29}$$

and

$$\left(\frac{2\pi}{1+\epsilon}\right)^{K/2} \geq \Psi_{\hat{\varepsilon}} \geq \left(\frac{2\pi}{1+\epsilon}\right)^{K/2}(1 - O(N^{-K/2}e^{-N\delta})) \quad \text{(A.30)}$$

for some $\delta > 0$.

Then from equation A.27, we get

$$\begin{aligned}
&\left\langle\left\langle\left(\int_{\mathcal{Z}_{\hat{\varepsilon}}}\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathcal{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)^T \mathbf{B}_0\left(\mathbf{z}+\frac{\tilde{\mathbf{v}}1_{\mathcal{R}_{\hat{\varepsilon}}}}{1+\epsilon}\right)\phi_{\hat{\varepsilon}}(\mathbf{z})d\mathbf{z}\right)\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\
&= \left(\frac{2\pi}{1+\epsilon}\right)^{K/2}\Psi_{\hat{\varepsilon}}^{-1}\left\langle\left\langle\left\langle \mathbf{z}^T\mathbf{B}_0\mathbf{z}1_{\mathcal{Z}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}} + \frac{\tilde{\mathbf{v}}^T\mathbf{B}_0^2\tilde{\mathbf{v}}1_{\mathcal{Z}_{\hat{\varepsilon}}}1_{\mathcal{R}_{\hat{\varepsilon}}}}{(1+\epsilon)^2}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \quad \text{(A.31)} \\
&\geq \left(\frac{2\pi}{1+\epsilon}\right)^{K/2}\Psi_{\hat{\varepsilon}}^{-1}\left\langle\left\langle\left\langle \mathbf{z}^T\mathbf{B}_0\mathbf{z}1_{\mathcal{Z}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \geq O(N^{-1}),
\end{aligned}$$

where

$$\begin{cases}
\left\langle\cdot\right\rangle_{\mathbf{z}} = \int_{\mathbb{R}^K}(\cdot)\phi_0(\mathbf{z})d\mathbf{z}, \\
\phi_0(\mathbf{z}) = \left(\frac{1+\epsilon}{2\pi}\right)^{K/2}\exp\left(-\frac{1+\epsilon}{2}\mathbf{z}^T\mathbf{z}\right).
\end{cases} \quad \text{(A.32)}$$

Here, notice that

$$\left(\frac{2\pi}{1+\epsilon}\right)^{K/2}\Psi_{\hat{\varepsilon}}^{-1} = 1 + O(N^{-K/2}e^{-N\alpha}) \quad \text{(A.33)}$$

and

$$\begin{aligned}
\left\langle\left\langle\left\langle \mathbf{z}^T\mathbf{B}_0\mathbf{z}1_{\bar{\mathcal{Z}}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} &= -\left\langle\left\langle\left\langle \mathbf{z}^T\mathbf{B}_0\mathbf{z}1_{\mathcal{Z}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}}\right\rangle_{\mathbf{r}\mid\mathbf{x}}\right\rangle_{\mathbf{x}} \\
&\geq -\left\langle\left\langle\left\|\mathbf{B}_0\right\|^2\right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2}\left\langle\left\langle\left\|\mathbf{z}\right\|^4 1_{\bar{\mathcal{Z}}_{\hat{\varepsilon}}}\right\rangle_{\mathbf{z}}\right\rangle_{\mathbf{r}\mid\mathbf{x}}^{1/2}\right\rangle_{\mathbf{x}} \quad \text{(A.34)} \\
&= O(N^{-1}).
\end{aligned}$$

Hence, from the consideration above, we find

$$\widehat{\Gamma}_{\omega} \geq \frac{K}{2} \ln\left(\frac{2\pi}{1+\epsilon}\right) + \frac{K}{2(1+\epsilon)^2} + O\left(N^{-1}\right). \quad \text{(A.35)}$$

Since ϵ is arbitrary, let it go to zero. Thus, combining equations A.24 and A.35 yields

$$\Gamma_{\omega} = -\left\langle \frac{1}{2} \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} + O\left(N^{-1}\right). \quad \text{(A.36)}$$

Considering

$$\left\langle \left\langle \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \geq \Gamma_{\omega}, \quad \text{(A.37)}$$

and combining equations 2.3 and A.36, we immediately get equation 2.53.

On the other hand, by conditions 2.54a and 2.54b, we have

$$\begin{cases} \left\langle \widetilde{\mathbf{v}}^T \widetilde{\mathbf{v}} 1_{\bar{\mathscr{R}}_{\hat{\varepsilon}}} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \leq \left\langle \dfrac{\|\widetilde{\mathbf{v}}\|_2^{2+2\tau}}{(\hat{\varepsilon}\sqrt{N})^{2\tau}} \right\rangle \leq N^{-\tau} (\hat{\varepsilon}_0)^{-2\tau} \left\langle \|\widetilde{\mathbf{v}}\|^{2+2\tau} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = o(1), \\ \left\langle \left\langle \left\langle \mathbf{z}^T \mathbf{B}_0 \mathbf{z} 1_{\bar{\mathscr{X}}_{\hat{\varepsilon}}} \right\rangle_{\mathbf{z}} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \geq -\left\langle \left\langle \|\mathbf{B}_0\|^2 \right\rangle_{\mathbf{r} \mid \mathbf{x}}^{1/2} \left\langle \left\langle \|\mathbf{z}\|^4 1_{\bar{\mathscr{X}}_{\hat{\varepsilon}}} \right\rangle_{\mathbf{z}} \right\rangle_{\mathbf{r} \mid \mathbf{x}}^{1/2} \right\rangle_{\mathbf{x}} = o(1). \end{cases} \quad \text{(A.38)}$$

Similarly we can get equation 2.55. This completes the proof of lemma 1. □

A.2 Proof of Lemma 2. Define the sets

$$\Omega_{\epsilon}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathscr{R} : \mathbf{y}^T \mathbf{B}_0 \mathbf{y} < \epsilon \|\mathbf{y}\|^2, \forall \mathbf{y} \in \mathbb{R}^K \right\} \quad \text{(A.39)}$$

and

$$\Theta_{\epsilon}(\mathbf{x}) = \left\{ \mathbf{r} \in \mathscr{R} : \int_{\bar{\mathscr{X}}_{\epsilon}(\mathbf{x})} \frac{p(\mathbf{r} \mid \hat{\mathbf{x}})\, p(\hat{\mathbf{x}})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} dx' < \epsilon \det\left(\mathbf{G}(\mathbf{x})\right)^{-1/2} \right\}, \quad \text{(A.40)}$$

where $\bar{\mathscr{X}}_{\epsilon}(\mathbf{x}) = \mathscr{X} - \mathscr{X}_{\epsilon}(\mathbf{x})$, assuming $\epsilon \in (0, 1/2)$ and $p(\mathbf{x}) > 0$.

Then by Markov's inequality, we have

$$\left\langle 1_{\bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \leq \mathbb{P}_{\mathbf{r}\mid\mathbf{x}}\left\{ \left\|\mathbf{B}_0\right\|^2 \geq \epsilon^2 \right\} \leq \epsilon^2 \left\langle \left\|\mathbf{B}_0\right\|^2 \right\rangle_{\mathbf{r}\mid\mathbf{x}} = O\left(N^{-1}\right), \quad \text{(A.41)}$$

and by equation 2.26b,

$$\begin{aligned} \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r}\mid\mathbf{x}} &= \mathbb{P}_{\mathbf{r}\mid\mathbf{x}}\left\{ \int_{\bar{\mathcal{X}}_\epsilon(\mathbf{x})} \frac{p(\mathbf{r}\mid\hat{\mathbf{x}})\, p(\hat{\mathbf{x}})}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x})} d\hat{\mathbf{x}} \geq \epsilon \det\left(\mathbf{G}(\mathbf{x})\right)^{-1/2} \right\} \\ &= \mathbb{P}_{\mathbf{r}\mid\mathbf{x}}\left\{ \det\left(\mathbf{G}(\mathbf{x})\right)^{1/2} \int_{\bar{\mathcal{X}}_{\hat{\omega}}(\mathbf{x})} p(\hat{\mathbf{x}}\mid\mathbf{r}) d\hat{\mathbf{x}} > \epsilon\, p(\mathbf{r}\mid\mathbf{x}) \right\} \\ &= O\left(N^{-\eta}\right). \end{aligned} \quad \text{(A.42)}$$

Consider the following equality:

$$\left\langle \ln \frac{p(\mathbf{r})}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r}\mid\mathbf{x}} = \left\langle 1_{\Theta_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r}\mid\mathbf{x}} + \left\langle 1_{\bar{\Theta}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r}\mid\mathbf{x}}. \quad \text{(A.43)}$$

For the last term in equation A.43, Jensen's inequality implies that

$$\left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \right\rangle_{\mathbf{x}} \leq \left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \right\rangle_{\mathbf{x}} \ln \frac{1}{\left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \right\rangle_{\mathbf{x}}} = o\left(N^{-1}\right). \quad \text{(A.44)}$$

For the first term in equation A.43, it follows from equations A.40 and A.9 that

$$\begin{aligned} &\left\langle 1_{\Theta_\epsilon} \ln \frac{p(\mathbf{r}}{p(\mathbf{r}\mid\mathbf{x})\, p(\mathbf{x}} \right\rangle_{\mathbf{r}\mid\mathbf{x}} \\ &\leq \left\langle 1_{\Theta_\epsilon} \ln \left(\int_{\mathcal{X}_\epsilon(\mathbf{x})} \exp\left(L(\mathbf{r}\mid\hat{\mathbf{x}}) - L(\mathbf{r}\mid\mathbf{x})\right) d\hat{\mathbf{x}} + \epsilon \det\left(\mathbf{G}(\mathbf{x})\right)^{-1/2} \right) \right\rangle_{\mathbf{r}\mid\mathbf{x}} \\ &\leq -\frac{K}{2} \ln\left(\det\left(\mathbf{G}(\mathbf{x})\right)\right) \\ &+ \left\langle 1_{\Theta_\epsilon} \ln \left(\int_{\mathcal{Y}_\epsilon} \exp\left(\mathbf{y}^T \tilde{\mathbf{v}} - \frac{1}{2}(1+\epsilon)\, \mathbf{y}^T \mathbf{y} + \frac{1}{2} \mathbf{y}^T \mathbf{B}_0 \mathbf{y} \right) d\mathbf{y} + \epsilon \right) \right\rangle_{\mathbf{r}\mid\mathbf{x}}. \end{aligned} \quad \text{(A.45)}$$

The last term, equation A.45, is upper-bounded by

$$\left\langle 1_{\Theta_\epsilon \cap \Omega_\epsilon} \ln\left(\int_{\mathbb{R}^K} \exp\left(\mathbf{y}^T \tilde{\mathbf{v}} - \frac{1}{2}(1-2\epsilon)\,\mathbf{y}^T\mathbf{y} \right) d\mathbf{y} + \epsilon \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}} \tag{A.46}$$

$$+ \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln\left(\int_{\mathbb{R}^K} \exp\left(\mathbf{y}^T \tilde{\mathbf{v}} - \frac{1}{2}(1-\epsilon)\,\mathbf{y}^T\mathbf{y} + \frac{1}{2}\mathbf{y}^T\mathbf{B}_0\mathbf{y} \right) d\mathbf{y} + \epsilon \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}}. \tag{A.47}$$

Equation A.46 is equal to

$$\begin{aligned} &\left\langle 1_{\Theta_\epsilon \cap \Omega_\epsilon} \ln\left(\left(\frac{2\pi}{1-2\epsilon} \right)^{K/2} \exp\left(\frac{\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}}{2(1-2\epsilon)} \right) + \epsilon \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}} \\ &\le \left\langle 1_{\Theta_\epsilon \cap \Omega_\epsilon} \left(\frac{\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}}{2(1-2\epsilon)} + \ln\left(\left(\frac{2\pi}{1-2\epsilon} \right)^{K/2} + \epsilon \right) \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}}. \end{aligned} \tag{A.48}$$

Equation A.47 is equal to

$$\begin{aligned} &\left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln\left(\left\langle \left(\frac{2\pi}{1-\epsilon} \right)^{K/2} \exp\left(\frac{1}{2}\left(\mathbf{z} + \frac{\tilde{\mathbf{v}}}{1-\epsilon} \right)^T \right.\right.\right. \\ &\left.\left.\left. \mathbf{B}_0\left(\mathbf{z} + \frac{\tilde{\mathbf{v}}}{1-\epsilon} \right) + \frac{\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}}{2(1-\epsilon)} \right) \right\rangle_{\mathbf{z}} + \epsilon \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}} \\ &\le \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \left(\frac{K}{2}\ln\left(\frac{2\pi}{1-\epsilon} \right) + \frac{\tilde{\mathbf{v}}^T\tilde{\mathbf{v}}}{2(1-\epsilon)} + \frac{\tilde{\mathbf{v}}^T\mathbf{B}_0^2\tilde{\mathbf{v}}}{2(1-\epsilon)^2} + \frac{\tilde{\mathbf{v}}^T\mathbf{B}_0^2\tilde{\mathbf{v}}}{(1-\epsilon)^3} \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}} \end{aligned} \tag{A.49a}$$

$$+ \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln\left(\left\langle \exp\left(\frac{1}{2}\mathbf{z}^T\mathbf{B}_0\mathbf{z} + \frac{\mathbf{z}^T\mathbf{B}_0\tilde{\mathbf{v}}}{1-\epsilon} - \frac{\tilde{\mathbf{v}}^T\mathbf{B}_0^2\tilde{\mathbf{v}}}{(1-\epsilon)^3} \right) \right\rangle_{\mathbf{z}} + \epsilon\left(\frac{1-\epsilon}{2\pi} \right)^{K/2} \right) \right\rangle_{\mathbf{r}\,|\,\mathbf{x}}, \tag{A.49b}$$

where

$$\begin{cases} \langle \cdot \rangle_{\mathbf{z}} = \displaystyle\int_{\mathbb{R}^K} (\cdot)\phi_1(\mathbf{z})d\mathbf{z} \\ \phi_1(\mathbf{z}) = \left(\dfrac{1-\epsilon}{2\pi} \right)^{K/2} \exp\left(-\dfrac{1-\epsilon}{2}\mathbf{z}^T\mathbf{z} \right) \end{cases}. \tag{A.50}$$

Notice that

$$\left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \leq \left\langle 1_{\bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(N^{-1}) \quad \text{(A.51)}$$

and

$$\left\langle 1_{\Theta_\epsilon \cap \Omega_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = 1 - \left\langle 1_{\bar{\Theta}_\epsilon \cup \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = 1 + O(N^{-1}). \quad \text{(A.52)}$$

Then by equation A.19, we get

$$\begin{aligned} &\left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \left(\left\langle \exp(\mathbf{z}^T \mathbf{B}_0 \mathbf{z}) \right\rangle_{\mathbf{z}}^{1/2} - 1 \right) \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &\leq \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \sum_{m=0}^{\infty} \frac{1}{m!} \left\langle (\mathbf{z}^T \mathbf{B}_0 \mathbf{z})^m \right\rangle_{\mathbf{z}} \right\rangle_{\mathbf{r} \mid \mathbf{x}}^{1/2} - \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(N^{-1}), \end{aligned} \quad \text{(A.53)}$$

$$\left\langle \tilde{\mathbf{v}}^T \tilde{\mathbf{v}} 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \leq \left\langle \|\tilde{\mathbf{v}}\|^4 \right\rangle_{\mathbf{r} \mid \mathbf{x}}^{1/2} \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}}^{1/2} = O(N^{-1}), \quad \text{(A.54)}$$

and by equation 2.51,

$$\begin{aligned} 0 \leq \left\langle \tilde{\mathbf{v}}^T \mathbf{B}_0^2 \tilde{\mathbf{v}} 1_{\Theta_\epsilon \cap \bar{\Omega}_\Omega} \right\rangle_{\mathbf{r} \mid \mathbf{x}} &\leq \left\langle \mathbf{v}^T \mathbf{B}_0^2 \mathbf{v} \right\rangle_{\mathbf{r} \mid \mathbf{x}} + O(N^{-1}) \\ &\leq \xi \left\| N \mathbf{G}^{-1}(\mathbf{x}) \right\| + O(N^{-1}) = O(N^{-1}). \end{aligned} \quad \text{(A.55)}$$

Hence, we have

$$\begin{aligned} &\left\langle 1_{\Theta_\epsilon} \left(\frac{K}{2} \ln \left(\frac{2\pi}{1-\epsilon} \right) + \frac{\tilde{\mathbf{v}}^T \tilde{\mathbf{v}}}{2(1-\epsilon)} \right) \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &= \left(\frac{K}{2} \ln \left(\frac{2\pi}{1-\epsilon} \right) + \frac{K+\zeta}{2(1-\epsilon)} \right) + O(N^{-1}), \end{aligned} \quad \text{(A.56)}$$

and by Cauchy-Schwarz inequality and equation A.53, the term A.49b is upper-bounded by

$$\begin{aligned}&\left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln \left(\left\langle \exp(\mathbf{z}^T \mathbf{B}_0 \mathbf{z}) \right\rangle_{\mathbf{z}}^{1/2} \left\langle \exp\left(\frac{2\mathbf{z}^T \mathbf{B}_0 \tilde{\mathbf{v}}}{1-\epsilon} - \frac{2\tilde{\mathbf{v}}^T \mathbf{B}_0^2 \tilde{\mathbf{v}}}{(1-\epsilon)^3} \right) \right\rangle_{\mathbf{z}}^{1/2} \right.\right. \\ &\left.\left. + \epsilon \left(\frac{1-\epsilon}{2\pi} \right)^{K/2} \right) \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &= \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln \left(\left\langle \exp(\mathbf{z}^T \mathbf{B}_0 \mathbf{z}) \right\rangle_{\mathbf{z}}^{1/2} + \epsilon \left(\frac{1-\epsilon}{2\pi} \right)^{K/2} \right) \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &\leq \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \left(\left\langle \exp(\mathbf{z}^T \mathbf{B}_0 \mathbf{z}) \right\rangle_{\mathbf{z}}^{1/2} + \epsilon \left(\frac{1-\epsilon}{2\pi} \right)^{K/2} - 1 \right) \right\rangle_{\mathbf{r} \mid \mathbf{x}} = O(N^{-1}) .\end{aligned} \tag{A.57}$$

Since ϵ is arbitrary, we can let it go to zero. Then, taking everything together, we get

$$\left\langle \left\langle \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{r})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \leq -\left\langle \frac{1}{2} \ln \left(\det \left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e} \right) \right) \right\rangle_{\mathbf{x}} + O(N^{-1}) . \tag{A.58}$$

Putting equation A.58 into 2.3 yields 2.56.

On the other hand, we have

$$\begin{aligned}&\left\langle \left\langle \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \\ &= \left\langle \left\langle 1_{\Theta_\epsilon \cap \Omega_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}}\end{aligned} \tag{A.59}$$

$$+ \left\langle \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} + \left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} . \tag{A.60}$$

For equation A.60, it fo11ows from Jensen's inequa1ity that

$$\left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \leq \left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \ln \frac{1}{\left\langle \left\langle 1_{\bar{\Theta}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}}} = o(1) \tag{A.61}$$

and

$$\left\langle \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \ln \frac{p(\mathbf{r})}{p(\mathbf{r} \mid \mathbf{x})\, p(\mathbf{x})} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \leq \left\langle \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}} \ln \frac{1}{\left\langle \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \right\rangle_{\mathbf{x}}} = o(1), \quad \text{(A.62)}$$

where

$$\begin{cases} \left\langle 1_{\bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \leq P(\left\| \mathbf{B}_0 \right\|^2 \geq \epsilon^2) \leq \epsilon^{-2} \left\langle \left\| \mathbf{B}_0 \right\|^2 \right\rangle_{\mathbf{r} \mid \mathbf{x}} = o(1), \\ \left\langle 1_{\Theta_\epsilon \cap \bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \leq \left\langle 1_{\bar{\Omega}_\epsilon} \right\rangle_{\mathbf{r} \mid \mathbf{x}} = o(1). \end{cases} \quad \text{(A.63)}$$

Similarly we can get equation 2.57. □

A.3 Proof of Theorem 1. By lemmas 1 and 2, we immediately get equation 2.58. The proof of equation 2.59 is similar. □

A.4 Proof of Theorem 2. First, we have

$$\mathbf{G}(\mathbf{x}) = \mathbf{J}^{1/2}(\mathbf{x})(\mathbf{I}_K + \boldsymbol{\Psi}(\mathbf{x}))\mathbf{J}^{1/2}(\mathbf{x}). \quad \text{(A.64)}$$

Since $\mathbf{J}(\mathbf{x})$ and $\mathbf{G}(\mathbf{x})$ are symmetric and positive-definite, $\mathbf{I}_K + \boldsymbol{\Psi}(\mathbf{x})$ is also symmetric and positive-definite. The eigendecompositon of $\boldsymbol{\Psi}(\mathbf{x})$ is given by

$$\boldsymbol{\Psi}(\mathbf{x}) = \mathbf{U}_\mathbf{x} \boldsymbol{\Lambda}_\mathbf{x} \mathbf{U}_\mathbf{x}^T, \quad \text{(A.65)}$$

where $\mathbf{U}_\mathbf{x} \in \mathbb{R}^{K \times K}$ is an orthogonal matrix and the matrix $\boldsymbol{\Lambda}_\mathbf{x} \in \mathbb{R}^{K \times K}$ is a $K \times K$ diagonal matrix with K nonnegative real numbers on the diagonal, $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_K > -1$. Then we have

$$\left\langle \mathrm{Tr}\,(\boldsymbol{\Lambda}_\mathbf{x}) \right\rangle_\mathbf{x} = \left\langle \mathrm{Tr}\,(\boldsymbol{\Psi}(\mathbf{x})) \right\rangle_\mathbf{x} = \left\langle \mathrm{Tr}\,(\mathbf{P}(\mathbf{x})\mathbf{J}^{-1}(\mathbf{x})) \right\rangle_\mathbf{x} = \varsigma \quad \text{(A.66)}$$

and

$$\left\langle \ln\,(\det\,(\mathbf{I}_K + \boldsymbol{\Psi}(\mathbf{x}))) \right\rangle_\mathbf{x} = \left\langle \mathrm{Tr}\,(\ln\,(\mathbf{I}_K + \boldsymbol{\Lambda}_\mathbf{x})) \right\rangle_\mathbf{x} \leq \left\langle \mathrm{Tr}\,(\boldsymbol{\Lambda}_\mathbf{x}) \right\rangle_\mathbf{x} = \varsigma. \quad \text{(A.67)}$$

Notice that $\ln(1 + x) \leq x$ for $\forall_x \in (-1, \infty)$. It follows from equations A.64 and A.67 that

$$\langle \ln(\det(\mathbf{G}(\mathbf{x})))\rangle_{\mathbf{x}} = \langle \ln(\det(\mathbf{J}(\mathbf{x})))\rangle_{\mathbf{x}} = \left\langle \ln(\det(\mathbf{I}_K + \mathbf{\Psi}(\mathbf{x})))\right\rangle_{\mathbf{x}} \leq \varsigma. \quad \text{(A.68)}$$

From equations 2.12, 2.11, and A.68, we obtain equation 2.62.

If $\mathbf{P}(\mathbf{x})$ is positive-semidefinite, then $\lambda_1 \geq \lambda_2 \geq, \ldots, \geq \lambda_K \geq 0$, $\varsigma \geq 0$ and $(\ln(\det\langle \mathbf{I}_K + \mathbf{\Psi}(\mathbf{x})))\rangle_{\mathbf{x}} \geq 0$. Hence we can get equation 2.63.

On the other hand, it follows from equations 2.64 and A.67 and the power mean inequality that

$$\begin{aligned} |\varsigma| &\leq \left\langle \sum_{k=1}^{K} |\lambda_k| \right\rangle_{\mathbf{x}} \leq \sqrt{K}\left\langle \left(\sum_{k=1}^{K} \lambda_k^2\right)^{1/2} \right\rangle_{\mathbf{x}} \\ &= \sqrt{K}\langle \|\mathbf{\Psi}(\mathbf{x})\| \rangle_{\mathbf{x}} = \sqrt{K}\varsigma_1 = O(N^{-\beta}). \end{aligned} \quad \text{(A.69)}$$

Let $\lambda_k^- = \min(0, \lambda_k)$ for $\forall k \in \{1, 2, \ldots, K\}$. Then

$$\left\langle \sum_{k=1}^{K} \ln(1+\lambda_k^-) \right\rangle_{\mathbf{x}} \leq \left\langle \ln(\det(\mathbf{I}_K + \mathbf{\Psi}(\mathbf{x}))) \right\rangle_{\mathbf{x}}. \quad \text{(A.70)}$$

Notice that $-1 < \lambda_k^- \leq 0$. Then by equation A.69, we have

$$\left\langle \sum_{k=1}^{K} \ln(1+\lambda_k^-) \right\rangle_{\mathbf{x}} = \left\langle \sum_{m=1}^{\infty} \frac{-1}{m} \sum_{k=1}^{K} (-\lambda_k^-)^m \right\rangle_{\mathbf{x}} = O(N^{-\beta}). \quad \text{(A.71)}$$

From equations 2.12, 2.11, A.68, A.70, and A.71, we immediately get equation 2.65. □

A.5 Proof of Theorem 3. Considering the change of variables theorem, for any real-valued function f and invertible transformation $\mathbf{T}$, we have

$$\int_{\widetilde{\mathcal{X}}} f(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}} = \int_{\mathcal{X}} f(\mathbf{T}(\mathbf{x})) \mid \det(D\mathbf{T}(\mathbf{x})) \mid d\mathbf{x}, \quad \text{(A.72)}$$

and for $p(\mathbf{x})$ and $p(\widetilde{\mathbf{x}})$,

$$p(\widetilde{\mathbf{x}})\,|_{\widetilde{\mathbf{x}} = T(\mathbf{x})} = \mid \det(D\mathbf{T}(\mathbf{x})) \mid^{-1} p(\mathbf{x}). \quad \text{(A.73)}$$

Then it follows from equations 4.2, A.72, and A.73 that

$$\begin{cases} p(\mathbf{r}) &= \int_{\mathcal{X}} p(\mathbf{r} \mid \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \int_{\widetilde{\mathcal{X}}} p(\mathbf{r} \mid \widetilde{\mathbf{x}}) p(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}}, \\ H(\widetilde{X}) &= -\int_{\widetilde{\mathcal{X}}} p(\widetilde{\mathbf{x}}) \ln p(\widetilde{\mathbf{x}}) d\mathbf{x} \\ &= -\int_{\mathcal{X}} p(\mathbf{x}) \ln \left(p(\mathbf{x}) \mid \det(D\mathbf{T}(\mathbf{x})) \mid^{-1} \right) d\mathbf{x} \\ &= H(X) + \int_{\mathcal{X}} p(\mathbf{x}) \ln \mid \det(D\mathbf{T}(\mathbf{x})) \mid d\mathbf{x}, \\ \mathbf{G}(\mathbf{x}) &= D\mathbf{T}(\mathbf{x})^T \mathbf{G}(\widetilde{\mathbf{x}}) D\mathbf{T}(\mathbf{x}) . \end{cases} \tag{A.74}$$

Substituting equations A.73 and A.74 into 2.1, we can directly obtain equation 4.3. Moreover, if $p(\widetilde{\mathbf{x}})$ and $p(\mathbf{r} \mid \widetilde{\mathbf{x}})$ fulfill conditions C1, C2 and $\xi = O(N^{-1})$, then by theorem 1, we immediately obtain equation 4.4. □

A.6 Proof of Corollary 1. It follows from equation 2.21 and theorem 3 that

$$\begin{aligned} I_G &= I_{G+} = I(X;R) = I(Y;R) \\ &= \frac{1}{2} \ln \left(\det \left(\frac{1}{2\pi e} (\mathbf{A}\mathbf{A}^T + \mathbf{\Sigma}_\mathbf{f}^{-1}) \right) \right) + H(Y) \end{aligned} \tag{A.75}$$

and

$$H(Y) = \frac{1}{2} \ln \left(\det \left(2\pi e \mathbf{\Sigma}_\mathbf{f} \right) \right) = H(X) + \langle \ln \mid \det \left(\mathbf{D}(\mathbf{x}) \right) \mid \rangle_\mathbf{x} . \tag{A.76}$$

Here notice that

$$\begin{aligned} \mathbf{J}(\mathbf{x}) &= \left\langle \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{x})}{\partial \mathbf{x}^T} \right\rangle_{\mathbf{r} \mid \mathbf{x}} \\ &= \left\langle \frac{\partial \mathbf{y}^T}{\partial \mathbf{x}} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}} \frac{\partial \ln p(\mathbf{r} \mid \mathbf{y})}{\partial \mathbf{y}^T} \frac{\partial \mathbf{y}}{\partial \mathbf{x}^T} \right\rangle_{\mathbf{r} \mid \mathbf{y}} \\ &= \mathbf{D}(\mathbf{x})^T \mathbf{A}\mathbf{A}^T \mathbf{D}(\mathbf{x}) \end{aligned} \tag{A.77}$$

and

$$\mathbf{P}(\mathbf{x}) = -\frac{\partial^2 \ln p(\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^T} = -\frac{\partial \mathbf{y}^T}{\partial \mathbf{x}} \frac{\partial^2 \ln p(\mathbf{y})}{\partial \mathbf{y} \partial \mathbf{y}^T} \frac{\partial \mathbf{y}}{\partial \mathbf{x}^T} = \mathbf{D}(\mathbf{x})^T \mathbf{\Sigma}_\mathbf{f}^{-1} \mathbf{D}(\mathbf{x}) . \tag{A.78}$$

Hence, combining equations A.75 to A.78, we can immediately obtain equation 4.9. □

A.7 Proof of Theorem 4. First, we have

$$\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} = \left\langle \ln\left(\det\left(\frac{\mathbf{G}_{1,1}(\mathbf{x})}{2\pi e}\right) \quad \det\left(\frac{1}{2\pi e}(\mathbf{G}_{2,2}(\mathbf{x}) - \mathbf{G}_{2,1}(\mathbf{x})\mathbf{G}_{1,1}^{-1}(\mathbf{x})\mathbf{G}_{1,2}(\mathbf{x}))\right)\right)\right\rangle_{\mathbf{x}}$$
$$= \left\langle \ln\left(\det\left(\frac{\mathbf{G}_{1,1}(\mathbf{x})}{2\pi e}\right)\right) + \ln\left(\det\left(\frac{\mathbf{G}_{2,2}(\mathbf{x})}{2\pi e}\right)\right) + \ln(\det(\mathbf{I}_{K_2} - \mathbf{A}_{\mathbf{x}}))\right\rangle_{\mathbf{x}}. \tag{A.79}$$

Then by the eigendecompositon of $\mathbf{A}_{\mathbf{x}}$, we have

$$\mathbf{A}_{\mathbf{x}} = \mathbf{U}_{\mathbf{x}}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{U}_{\mathbf{x}}^{T}, \tag{A.80}$$

where $\mathbf{U}_{\mathbf{x}}$ and $\mathbf{\Lambda}_{\mathbf{x}}$ are $K_2 \times K_2$ eigenvector matrix and eigenvalue matrix, respectively. Since $\mathbf{G}(\mathbf{x})$, $\mathbf{G}_{1,1}(\mathbf{x})$, and $\mathbf{G}_{2,2}(\mathbf{x})$ are positive-definite, then $\mathbf{I}_K - \mathbf{A}_{\mathbf{x}}$ is also positive-definite and $\mathbf{A}_{\mathbf{x}}$ is positive-semidefinite, with $0 \le (\mathbf{\Lambda}_{\mathbf{x}})_{k,k} = \lambda_k < 1$ for $\forall k \in \{1, 2, \ldots, K_2\}$. Moreover, it follows from equation 4.33 that

$$\begin{cases} 0 \le \left\langle \mathrm{Tr}(\mathbf{\Lambda}_{\mathbf{x}})\right\rangle_{\mathbf{x}} = \left\langle \mathrm{Tr}(\mathbf{A}_{\mathbf{x}})\right\rangle_{\mathbf{x}} \ll 1, \\ 0 \le \left\langle \mathrm{Tr}(\mathbf{\Lambda}_{\mathbf{x}}^{m})\right\rangle_{\mathbf{x}} = \left\langle \sum_{k=1}^{K_2} \lambda_k^m \right\rangle_{\mathbf{x}} \le \left\langle \mathrm{Tr}(\mathbf{\Lambda}_{\mathbf{x}})\right\rangle_{\mathbf{x}} \ll 1. \end{cases} \tag{A.81}$$

Then by equation A.81, we have

$$\begin{aligned} \left\langle \ln(\det(\mathbf{I}_{K_2} - \mathbf{A}_{\mathbf{x}}))\right\rangle_{\mathbf{x}} &= \left\langle \mathrm{Tr}(\ln(\mathbf{I}_{K_2} - \mathbf{\Lambda}_{\mathbf{x}}))\right\rangle_{\mathbf{x}} \\ &= \sum_{m=1}^{\infty} \frac{-1}{m} \left\langle \mathrm{Tr}(\mathbf{\Lambda}_{\mathbf{x}}^{m})\right\rangle_{\mathbf{x}} \simeq 0. \end{aligned} \tag{A.82}$$

Substituting equation A.82 into A.79 and then combining with equation 2.12, we get equation 4.35.

If equation 4.36 holds, then $\mathbf{A}_{\mathbf{x}} = \mathbf{0}$ and $I_G = I_{G_1}$. Conversely, if $I_G = I_{G_1}$, then

$$0 = \left\langle \ln(\det(\mathbf{I}_{K_2} - \mathbf{A}_{\mathbf{x}}))\right\rangle_{\mathbf{x}} \le -\left\langle \mathrm{Tr}(\mathbf{A}_{\mathbf{x}})\right\rangle_{\mathbf{x}} \le 0, \tag{A.83}$$

$\mathbf{A}_{\mathbf{x}} = \mathbf{0}$, and equation 4.36 holds. □

A.8 Proof of Theorem 5. Similar to equation A.79, we have

$$\left\langle \ln\left(\det\left(\frac{\mathbf{G}(\mathbf{x})}{2\pi e}\right)\right)\right\rangle_{\mathbf{x}} = \left\langle \ln\left(\det\left(\frac{\mathbf{G}_{1,1}(\mathbf{x})}{2\pi e}\right)\right) + \ln\left(\det\left(\frac{\mathbf{P}_{2,2}(\mathbf{x})}{2\pi e}\right)\right) + \ln\left(\det\left(\mathbf{I}_{K_2} - \mathbf{B}_{\mathbf{x}}\right)\right)\right\rangle_{\mathbf{x}}. \quad \text{(A.84)}$$

Similar to equation A.65, the eigendecompositon of $\mathbf{B}_{\mathbf{x}}$ is given by

$$\mathbf{B}_{\mathbf{x}} = \mathbf{U}_{\mathbf{x}}\mathbf{\Lambda}_{\mathbf{x}}\mathbf{U}_{\mathbf{x}}^{T}, \quad \text{(A.85)}$$

where $\mathbf{U}_{\mathbf{x}}$ and $\mathbf{\Lambda}_{\mathbf{x}}$ are $K_2 \times K_2$ eigenvector matrix and eigenvalue matrix, respectively. If the matrix $\mathbf{B}_{\mathbf{x}}$ is positive-semidefinite and satisfies equation 4.38, then $(\mathbf{\Lambda}_{\mathbf{x}})_{k,k} = \lambda_k \geq 0$ for $\forall k \in \{1, 2, \ldots, K_2\}$ and

$$0 \leq \left\langle \ln\left(\det\left(\mathbf{I}_{K_2} + \mathbf{B}_{\mathbf{x}}\right)\right)\right\rangle_{\mathbf{x}} = \left\langle \sum_{k=1}^{K_2} \ln\left(1 + \lambda_k\right)\right\rangle_{\mathbf{x}} \leq \left\langle \mathrm{Tr}\left(\mathbf{\Lambda}_{\mathbf{x}}\right)\right\rangle_{\mathbf{x}} = \mathrm{Tr}\left(\left\langle \mathbf{B}_{\mathbf{x}}\right\rangle_{\mathbf{x}}\right) \ll 1. \quad \text{(A.86)}$$

Substituting equation A.86 into A.84, we immediately get equation 4.41. If $\mathbf{C}_{\mathbf{x}} = \mathbf{0}$, then $\ln(\det(\mathbf{I}_{K_2} + \mathbf{B}_{\mathbf{x}})) = 0$ and $I_G = I_{G_2}$. And if $I_G = I_{G_2}$, then $\ln(\det(\mathbf{I}_{K_2} + \mathbf{B}_{\mathbf{x}})) = 0$, $\mathbf{B}_{\mathbf{x}} = \mathbf{0}$ and $\mathbf{C}_{\mathbf{x}} = \mathbf{0}$. □

A.9 Proof of Corollary 2. Notice that

$$\left\{H(X) = H(X_1) + H(X_2), H(X_2) = \frac{1}{2}\ln\left(\det\left(2\pi e\mathbf{\Sigma}_{\mathbf{x}_2}\right)\right), \mathbf{P}_{2,1}(\mathbf{x}) = \mathbf{P}_{1,2}(\mathbf{x}) = \mathbf{0}, \mathbf{P}_{2,2}(\mathbf{x}) = \mathbf{\Sigma}_{\mathbf{x}_2}^{-1}\right.,$$

(A.87)

and the matrices

$$\mathbf{C}_{\mathbf{x}} = \mathbf{J}_{2,2}(\mathbf{x}) - \mathbf{J}_{2,1}(\mathbf{x})\mathbf{G}_{1,1}^{-1}(\mathbf{x})\mathbf{J}_{1,2}(\mathbf{x}), \quad \text{(A.88)}$$

$$\mathbf{B}_{\mathbf{x}} = \mathbf{P}_{2,2}^{-1/2}(\mathbf{x})\mathbf{C}_{\mathbf{x}}\mathbf{P}_{2,2}^{-1/2}(\mathbf{x}) \quad \text{(A.89)}$$

are positive-semidefinite, and the proof is similar to equation 4.74. Then by theorem 5, we immediately get equation 4.41. Substituting equation A.87 into 4.41 yields equation 4.44 with strict equality if and only if $\mathbf{C}_{\mathbf{x}} = \mathbf{0}$. □

A.10 Proof of Proposition 2. By writing $p(\boldsymbol{\theta})$ as a sum of two density functions $p_1(\boldsymbol{\theta})$ and $p_2(\boldsymbol{\theta})$,

$$p(\boldsymbol{\theta}) = \alpha p_1(\boldsymbol{\theta}) + (1-\alpha)p_2(\boldsymbol{\theta}), \quad \text{(A.90)}$$

we have

$$\mathbf{G}(\mathbf{x}) = N\int_{\Theta} p(\boldsymbol{\theta})\mathbf{S}(\mathbf{x};\boldsymbol{\theta})d\boldsymbol{\theta} + \mathbf{P}(\mathbf{x}) = \alpha\mathbf{G}_1(\mathbf{x}) + (1-\alpha)\mathbf{G}_2(\mathbf{x}), \quad \text{(A.91)}$$

where $0 \le \alpha \le 1$ and

$$\mathbf{G}_1(\mathbf{x}) = N\int_{\Theta} p_1(\boldsymbol{\theta})\mathbf{S}(\mathbf{x};\boldsymbol{\theta})d\boldsymbol{\theta} + \mathbf{P}(\mathbf{x}), \quad \text{(A.92)}$$

$$\mathbf{G}_2(\mathbf{x}) = N\int_{\Theta} p_2(\boldsymbol{\theta})\mathbf{S}(\mathbf{x};\boldsymbol{\theta})d\boldsymbol{\theta} + \mathbf{P}(\mathbf{x}). \quad \text{(A.93)}$$

Using the Minkowski determinant inequality and the inequality of weighted arithmetic and geometric means, we find

$$\det(\mathbf{G}(\mathbf{x}))^{1/K} = \det(\alpha\mathbf{G}_1(\mathbf{x}) + (1-\alpha)\mathbf{G}_2(\mathbf{x}))^{1/K} \ge \alpha\det(\mathbf{G}_1(\mathbf{x}))^{1/K} + (1-\alpha)\det(\mathbf{G}_2(\mathbf{x}))^{1/K} \ge \left(\det(\mathbf{G}_1(\mathbf{x}))^{\alpha}\det(\mathbf{G}_2(\mathbf{x}))^{(1-\alpha)}\right)^{1/K}. \quad \text{(A.94)}$$

It follows from equations A.91 and A.94 that

$$\ln(\det(\alpha\mathbf{G}_1(\mathbf{x}) + (1-\alpha)\mathbf{G}_2(\mathbf{x}))) \ge \alpha\ln(\det(\mathbf{G}_1(\mathbf{x}))) + (1-\alpha)\ln(\det(\mathbf{G}_2(\mathbf{x}))), \quad \text{(A.95)}$$

where the equality holds if and only if $\mathbf{G}_1(\mathbf{x}) = \mathbf{G}_2(\mathbf{x})$. Thus $\ln(\det(\mathbf{G}(\mathbf{x})))$ is concave about $p(\boldsymbol{\theta})$. Therefore $I_G[p(\boldsymbol{\theta})]$ is a concave function about $p(\boldsymbol{\theta})$. Similarly, we can prove that $I_F[p(\boldsymbol{\theta})]$ is also a concave function about $p(\boldsymbol{\theta})$. □

References

Abbott LF, & Dayan P (1999). The effect of correlated variability on the accuracy of a population code. Neural Comput., 11(1), 91–101. [PubMed: 9950724]

Amari S, & Nakahara H (2005). Difficulty of singularity in population coding. Neural Comput., 17(4), 839–858. [PubMed: 15829091]

Atick JJ, Li ZP, & Redlich AN (1992). Understanding retinal color coding from first principles. Neural Comput., 4(4), 559–572.

Atick JJ, & Redlich AN (1990). Towards a theory of early visual processing. Neural Comput., 2(3), 308–320.

Becker S, & Hinton GE (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. Nature, 355(6356), 161–163. [PubMed: 1729650]

Bell AJ, & Sejnowski TJ (1997). The “independent components” of natural scenes are edge filters. Vision Res., 37(23), 3327–3338. [PubMed: 9425547]

Bethge M, Rotermund D, & Pawelzik K (2002). Optimal short-term population coding: When Fisher information fails. Neural Comput., 14(10), 2317–2351. [PubMed: 12396565]

Borst A, & Theunissen FE (1999). Information theory and neural coding. Nat. Neurosci, 2(11), 947–257. [PubMed: 10526332]

Boyd S, & Vandenberghe L (2004). Convex optimization. Cambridge: Cambridge University Press.

Brown EN, Kass RE, & Mitra PP (2004). Multiple neural spike train data analysis: State-of-the-art and future challenges. Nat. Neurosci, 7(5), 456–461. [PubMed: 15114358]

Brunel N, & Nadal JP (1998). Mutual information, Fisher information, and population coding. Neural Comput., 10(7), 1731–1757. [PubMed: 9744895]

Carlton A (1969). On the bias of information estimates. Psychological Bulletin, 71(2), 108.

Chase SM, & Young ED (2005). Limited segregation of different types of sound localization information among classes of units in the inferior colliculus. Journal of Neuroscience, 25(33), 7575–7585. [PubMed: 16107645]

Chechik G, Anderson MJ, Bar-Yosef O, Young ED, Tishby N, & Nelken I (2006). Reduction of information redundancy in the ascending auditory pathway. Neuron, 51(3), 359–368. [PubMed: 16880130]

Clarke BS, & Barron AR (1990). Information-theoretic asymptotics of Bayes methods. IEEE Trans. Inform. Theory, 36(3), 453–471.

Cover TM, & Thomas JA (2006). Elements of information (2nd ed.) New York: Wiley-Interscience.

Eckhorn R, & Pöpel B (1975). Rigorous and extended application of information theory to the afferent visual system of the cat. II. Experimental results. Biological Cybernetics, 17(1), 7–17.

Ganguli D, & Simoncelli EP (2014). Efficient sensory encoding and Bayesian inference with heterogeneous neural populations. Neural Comput., 26(10), 2103–2134. [PubMed: 25058702]

Gawne TJ, & Richmond BJ (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? Journal of Neuroscience, 13(7), 2758–2771. [PubMed: 8331371]

Gourévitch B, & Eggermont JJ (2007). Evaluating information transfer between auditory cortical neurons. Journal of Neurophysiology, 97(3), 2533–2543. [PubMed: 17202243]

Guo DN, Shamai S, & Verdu S (2005). Mutual information and minimum mean-square error in gaussian channels. IEEE Trans. Inform. Theory, 51(4), 1261–1282.

Harper NS, & McAlpine D (2004). Optimal neural population coding of an auditory spatial cue. Nature, 430(7000), 682–686. [PubMed: 15295602]

Huang W, Huang X, & Zhang K (2017). Information-theoretic interpretation of tuning curves for multiple motion directions. In Proceedings of the 51st Annual Conference on Information Sciences and Systems (pp. 1–4). Piscataway, NJ: IEEE.

Huang W, & Zhang K (2017). 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). arXiv:1611.01886.

Jeffreys H (1961). Theory of probability (3rd ed.). New York: Oxford University Press.

Kang K, & Sompolinsky H (2001). Mutual information of population codes and distance measures in probability space. Phys. Rev. Lett, 86(21), 4958–4961. [PubMed: 11384391]

Khan S, Bandyopadhyay S, Ganguly AR, Saigal S, Erickson DJ, III, Protopopescu V, & Ostrouchov G (2007). Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Physical Review E, 76(2), 026209.

Kraskov A, Stögbauer H, & Grassberger P (2004). Estimating mutual information. Physical Review E, 69(6), 066138.

Laughlin SB, & Sejnowski TJ (2003). Communication in neuronal networks. Science, 301(5641), 1870–1874. [PubMed: 14512617]

Lewis A, & Zhaoping L (2006). Are cone sensitivities determined by natural color statistics? J. Vis, 6(3), 285–302. [PubMed: 16643096]

MacKay DJC (2003). Information theory, inference and learning algorithms. Cambridge: Cambridge University Press.

McClurkin JW, Gawne TJ, Optican LM, & Richmond BJ (1991). Lateral geniculate neurons in behaving primates. II. Encoding of visual information in the temporal shape of the response. Journal of Neurophysiology, 66(3), 794–808. [PubMed: 1753288]

Miller GA (1955). Note on the bias of information estimates In Quastler H (Ed.), Information theory in psychology: Problems and methods II-B (pp. 95–100). Glencoe, IL: Free Press.

Olshausen BA, & Field DJ (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature, 381(6583), 607–609. [PubMed: 8637596]

Optican LM, & Richmond BJ (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. Journal of Neurophysiology, 57(1), 162–178. [PubMed: 3559670]

Paninski L (2003). Estimation of entropy and mutual information. Neural Comput., 15(6), 1191–1253.

Pouget A, Dayan P, & Zemel R (2000). Information processing with population codes. Nat. Rev. Neurosci, 1(2), 125–132. [PubMed: 11252775]

Quiroga R, & Panzeri S (2009). Extracting information from neuronal populations: Information theory and decoding approaches. Nat. Rev. Neurosci, 10(3), 173–185. [PubMed: 19229240]

Rao CR (1945). Information and accuracy attainable in the estimation of statistical parameters. Bulletin of the Calcutta Mathematical Society, 37(3), 81–91.

Rieke F, Warland D, de Ruyter van Steveninck R, & Bialek W (1997). Spikes: Exploring the neural code. Cambridge, MA: MIT Press.

Rissanen JJ (1996). Fisher information and stochastic complexity. IEEE Trans. Inform. Theory, 42(1), 40–47.

Shannon C (1948). A mathematical theory of communications. Bell System Technical Journal, 27, 379–423, 623–656.

Sompolinsky H, Yoon H, Kang KJ, & Shamir M (2001). Population coding in neuronal systems with correlated noise. Phys. Rev. E, 64(5), 051904.

Tovee MJ, Rolls ET, Treves A, & Bellis RP (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. Journal of Neurophysiology, 70(2), 640–654. [PubMed: 8410164]

Toyoizumi T, Aihara K, & Amari S (2006). Fisher information for spike-based population decoding. Phys. Rev. Lett, 97(9), 098102. [PubMed: 17026405]

Treves A, & Panzeri S (1995). The upward bias in measures of information derived from limited data samples. Neural Comput., 7(2), 399–407.

Van Hateren JH (1992). Real and optimal neural images in early vision. Nature, 360(6399), 68–70. [PubMed: 1436076]

Van Trees HL, & Bell KL (2007). Bayesian bounds for parameter estimation and nonlinear filtering/tracking. Hoboken, NJ: Wiley.

Verdu S (1986). Capacity region of gaussian CDMA channels: The symbol synchronous case. In Proc. 24th Allerton Conf. Communication, Control and Computing, (pp. 1025–1034). Piscataway, NJ: IEEE.

Victor JD (2000). Asymptotic bias in information estimates and the exponential (bell) polynomials. Neural Comput., 12(12), 2797–2804. [PubMed: 11112255]

Wei X-X, & Stocker AA (2015). Mutual information, Fisher information, and efficient coding. Neural Comput., 28, 305–326. [PubMed: 26654209]

Yarrow S, Challis E, & Series P (2012). Fisher and Shannon information in finite neural populations. Neural Comput., 24(7), 1740–1780. [PubMed: 22428594]

Zhang K, Ginzburg I, McNaughton BL, & Sejnowski TJ (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. J. Neurophysiol, 79(2), 1017–1044. [PubMed: 9463459]

Zhang K, & Sejnowski TJ (1999). Neuronal tuning: To sharpen or broaden? Neural Comput., 11(1), 75–84. [PubMed: 9950722]

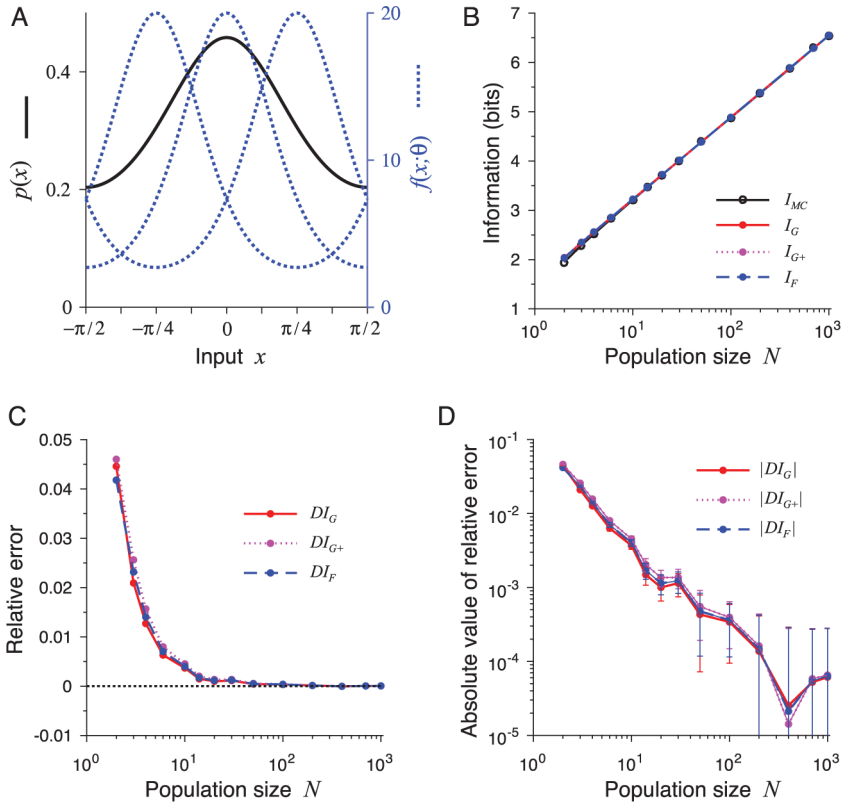

Figure 1:

A comparison of approximations I_{MC}, I_G, I_{G+}, and I_F for one-dimensional input stimuli. All of them were almost equally good, even for small population size N. (A) The stimulus distribution $p(x)$ and tuning curves $f(x; \theta)$ with different centers $\theta = -\pi/4, 0, \pi/4$. (B) The values of I_{MC}, I_G, I_{G+}, and I_F all increase with neuron number N. (C) The relative errors DI_G, DI_{G+}, and DI_F for the results in panel B. (D) The absolute values of the relative errors $|DI_G|$, $|DI_{G+}|$, and $|DI_F|$, with error bars showing standard deviations of repeated trials.

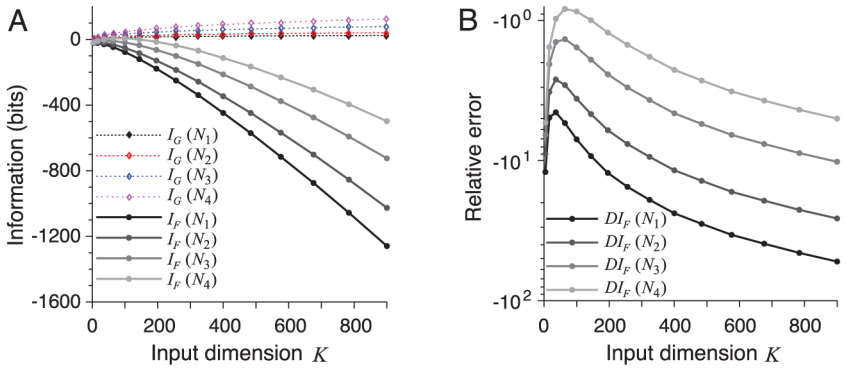

Figure 2:

A comparison of approximations I_G and I_F for different input dimensions. Here I_G is always equal to the true MI with $I_G = I_{G+} = I(X; R)$, whereas I_F always has nonzero errors. (A) The values I_G and I_F vary with input dimension $K = \omega^2$ with $\omega = 2, 4, 6, \ldots, 30$, and the number of neurons $N = N_i$ with $N_1 = 10^4$, $N_2 = 2 \times 10^4$, $N_3 = 5 \times 10^4$, $N_4 = 10^5$. (B) The relative error DI_F changes with input dimension K for different N.