Quantifying information conveyed by large neuronal populations

John A. Berkowitz

Department of Physics

University of California San Diego

San Diego, CA 92093

jaberkow@ucsd.edu

Tatyana O. Sharpee

Computational Neurobiology Laboratory

Salk Institute for Biological Studies, La Jolla, CA 92037

Department of Physics

University of California San Diego

San Diego, CA 92093

sharpee@salk.edu

February 15, 2019

Abstract

Quantifying mutual information between inputs and outputs of a large neural circuit is an important open problem in both machine learning and neuroscience. However, evaluation of the mutual information is known to be generally intractable for large systems due to the exponential growth in the number of terms that need to be evaluated. Here we show how information contained in the responses of large neural populations can be effectively computed provided the input-output functions of individual neurons can be measured and

approximated by a logistic function applied to a potentially nonlinear function of the stimulus. Neural responses in this model can remain sensitive to multiple stimulus components. We show that the mutual information in this model can be effectively approximated as a sum of lower-dimensional conditional mutual information terms. The approximations become exact in the limit of large neural populations and for certain conditions on the distribution of receptive fields across the neural population. We empirically find that these approximations continue to work well even when the conditions on the receptive field distributions are not fulfilled. The computing cost for the proposed methods grows linearly in the dimension of the input, and compares favorably with other approximations.

1 Introduction

Information theory has the potential of answering many important questions about how neurons communicate within the brain. In particular, it can help determine whether neural responses provide sufficient amounts of information about certain stimulus features, and in this way determine whether these features could possibly affect the animal's behavior (Rieke et al., 1997; Bialek, 2012). In addition, a number of previous studies have shown that one can understand many aspects of the neural circuit organization as those that provide maximal amounts of information under metabolic constraints (Laughlin et al., 1998; Bialek, 2012). Key to all of these analyses is the ability to compute the Shannon mutual information (Cover and Thomas, 2012). When estimating the information transmitted by neural populations from experimental recordings, all empirical methods produce biased estimates (Paninski, 2003). There are several approaches to trying to reduce or account for this bias (Nemenman et al., 2004; Strong et al., 1998; Brenner et al., 2000; Treves and Panzeri, 1995), but these approaches do not have finite-sample guarantees and are generally ineffective when the population response is high dimensional. In order to make progress on this problem, we consider the case where the response functions of individual neurons can be measured and where the stimulus-conditional ("noise") correlations between neural responses can be described by pairwise statistics (Schneidman et al., 2006). Historically, even with these assumptions the mutual information is notoriously difficult to compute in part due to the large number of possible responses that a set of neurons can jointly produce (Nemenman et al., 2004; Strong et al., 1998). The number of patterns grows exponentially with both the number of time points (Strong et al., 1998; Dettner

et al., 2016) and the number of neurons.

In this paper we will describe a set of approaches for computing information conveyed by responses of large neural populations. These methods build on recent advances for computing information based on linear combinations of neural responses across time (Dettner et al., 2016; Yu et al., 2010) and/or neurons (Berkowitz and Sharpee, 2018). We will show that when each individual neuron's firing probability depends monotonically on a (potentially nonlinear) function of the stimulus, the information contained in the full population response can be completely preserved by a linear transformation of the population output. This calculation still involves computing information between high dimensional vector variables. Therefore, we further show how the full information can be effectively approximated using a sum of conditional mutual information values between pairs of low-dimensional variables. The resulting approach makes it possible to avoid the "curse of dimensionality" with respect to the number of neurons when computing the mutual information from large neural populations.

2 Framework setup

Our analysis will target neural responses considered over sufficiently small time windows such that no more than one spike can be produced by any given neuron. We model the neural population as a set of binary neurons with sigmoidal tuning curves with response probability described by:

$$P(r_n = 1|\vec{s}) = \frac{1}{1 + e^{2f_n(\vec{s})}}, \tag{1}$$

where $\vec{s} \in \mathbb{R}^D$ is the input, $r_n \in \{-1, 1\}$ is the activity of the n^{th} neuron, and $f_n(\vec{s})$ is a scalar function of $\vec{s}$ representing the activation function of the n^{th} neuron. The population consists of N such neurons, and the population response is denoted as $\vec{r} = (r_1, ..., r_N)$. For clarity of the derivation, we will initially assume that neural responses are independent conditioned on $\vec{s}$:

$$P(\vec{r}|\vec{s}) = \prod_n P(r_n|\vec{s}), \tag{2}$$

and later discuss under what conditions our results generalize to the case where neural responses are correlated for a given stimulus $\vec{s}$. A few lines of algebra suffice to show that Eq. (2) can be expressed in

the following form:

$$\begin{aligned} P(\vec{r}\,|\vec{s}) &= \exp\left(\sum_n r_n f_n(\vec{s}) - A_n(\vec{s})\right), \\ A_n(\vec{s}) &= \log(2\cosh(f_n(\vec{s})). \end{aligned} \tag{3}$$

This formulation will assist all of the approaches described below for computing the mutual information.

3 An unbiased estimator of information for large neural populations

In order to test the approaches described in subsequent sections, we first developed a Monte-Carlo method for computing the "ground-truth" mutual information that works for large neural populations. The approach relies on the knowledge of neural response parameters $\{f_n(\vec{s})\}$ to produce unbiased estimates of mutual information between $\vec{R}$ and $\vec{S}$ for different choices of $\{f_n(\vec{s})\}$ or $P(\vec{s})$. Here and in what follows, upper case letters (e.g. $\vec{S}$) represent random variables, while lower case letters (e.g. $\vec{s}$) represent specific values of the associated random variables. The input distribution $P(\vec{s})$ is defined by drawing N_{stim} samples; we denote this set of samples as $\{\vec{s}_\mu\}$. Because of this approximation however, $I(\vec{R},\vec{S}\,)$ will be bounded above by $\log(N_{\text{stim}})$ (as will be any unbiased estimator of mutual information).

Although there are several formulations of the mutual information in terms of the entropies of $\vec{R}$ and $\vec{S}$ it serves to examine just one:

$$I(\vec{R},\vec{S}\,) = H(\vec{R}\,) - H(\vec{R}\,|\vec{S}\,). \tag{4}$$

Here, $H(\vec{R}\,)$ is the Shannon entropy of the marginal distribution of $\vec{R}$ and $H(\vec{R}\,|\vec{S}\,)$ is the conditional entropy of $\vec{R}$ given $\vec{S}$. Because we intend to use this estimator as a way to test the quality of other approximations, we will only consider here the case of conditionally independent neural responses $\vec{R}$. In this case, the noise entropy $H(\vec{R}\,|\vec{S}=\vec{S}\,)$ decomposes into a sum over neurons:

$$H(\vec{R}\,|\vec{S}=\vec{s}) = \sum_n \bar{r}_n(\vec{s}) f_n(\vec{s}) - A_n(\vec{s}), \tag{5}$$

where $\bar{r}_n(\vec{s})$ is the expected value of R_n given $\vec{s}$:

$$\bar{r}_n(\vec{s}) = \tanh(f_n(\vec{s})). \tag{6}$$

We denote $\hat{H}(\vec{R}\,|\vec{S}\,)$ as the finite sample approximation to $H(\vec{R}\,|\vec{S}\,)$:

$$\hat{H}(\vec{R}\,|\vec{S}\,) = -\frac{1}{N_{\text{stim}}} \sum_{\mu} \sum_{n} \bar{r}_n(\vec{s}_\mu) f_n(\vec{s}_\mu) - A_n(\vec{s}_\mu). \tag{7}$$

The conditional entropy $\hat{H}(\vec{R}\,|\vec{S}\,)$ can be evaluated in $O(N * N_{\text{stim}})$ time, not including the cost of evaluating $f_n(\vec{s})$. However, the marginal distribution of $\vec{r}$ will in general not factor. Thus evaluating $H(\vec{R}\,)$ requires computing the marginal $P(\vec{r})$ for all $\vec{r} \in \{-1, 1\}^N$. This computation grows like $O(N * N_{\text{stim}} * 2^N)$. Thus, evaluation of Eq. (4) is known to become intractable for realistic population sizes. To derive our estimator, we begin by rewriting $H(\vec{R}\,)$:

$$\begin{aligned} H(\vec{r}) &= -\sum_{\vec{r}} P(\vec{r}) \log(P(\vec{r})) = -\langle F(\vec{r}) \rangle_{\vec{r}}, \\ F(\vec{r}) &= \log\left(\langle P(\vec{r}\,|\vec{s}\,) \rangle_{\vec{s}} \right). \end{aligned} \tag{8}$$

We approximate the log-marginal $F(\vec{r})$ with an empirical average:

$$\hat{F}(\vec{r}) = \log\left(\frac{1}{N_{\text{stim}}} \sum_{\mu} P(\vec{r}\,|\vec{s}_\mu) \right) = \log\left(\frac{1}{N_{\text{stim}}} \sum_{\mu} \exp\left(\sum_{n} r_n f_n(\vec{s}_\mu) - A_n(\vec{s}_\mu) \right) \right) \tag{9}$$

In terms of numerical implementation, $\hat{F}(\vec{r})$ can be efficiently and stably evaluated in $O(N_{\text{stim}})$ time using the *logsumexp* function that is implemented in many numerical libraries. To approximate the averaging with respect to $P(\vec{r})$ we draw B samples of $\vec{r}$ for every $\vec{s}_\mu$, which is easily done with (1) and (2), and denote these samples as $\{\vec{r}_\nu\}$. We can thus produce an unbiased estimate of $-\langle \hat{F}(\vec{r}) \rangle_{\vec{r}}$:

$$\hat{H}(\vec{R}\,) = \frac{1}{B N_{\text{stim}}} \sum_{\nu} \log\left(\frac{1}{N_{\text{stim}}} \sum_{\mu} \exp\left(\sum_{n} r_{n\nu} f_n(\vec{s}_\mu) - A_n(\vec{s}_\mu) \right) \right) \tag{10}$$

Importantly, the response entropy $\hat{H}(\vec{R}\,)$ requires $O(N * B * N_{\text{stim}}^2)$ operations, a substantial improvement over exact evaluation of $H(\vec{r})$ when $B * N_{\text{stim}} \ll 2^N$. We note that even though we are able to produce

unbiased estimates of $\hat{H}(\vec{R})$, this estimator systematically underestimates the "infinite sample" entropy computed with respect to $P(\vec{s})$ explicitly, i.e. not defined by input samples (see Appendix A). Our Monte-Carlo estimator of $I(\vec{R},\vec{S})$ is the straightforward combination of $\hat{H}(\vec{R})$ and $\hat{H}(\vec{R}\,|\vec{S})$:

$$\hat{I}(\vec{R},\vec{S}) = \hat{H}(\vec{R}) - \hat{H}(\vec{R}\,|\vec{S}) \tag{11}$$

Although $\hat{I}(\vec{R},\vec{S})$ is an unbiased estimator of the mutual information (after accounting for the approximation of $P(\vec{s})$ by samples $\{\vec{s}_\mu\}$) the variance of $\hat{H}(\vec{r})$ and thus of $\hat{I}(\vec{R},\vec{S})$ can be difficult to quantify. However, $F(\vec{r})$ is a bounded function because $\vec{r}$ has finite support (or, more generally, $F(\vec{r})$ can be treated as a continuous function on the compact set $[-1,1]^N$). Thus, standard concentration bounds show that $\hat{H}(\vec{R})$ is a consistent estimator of $H(\vec{R})$.

In order to test our derivation that $\hat{I}(\vec{R},\vec{S})$ is an unbiased estimator of $I(\vec{R},\vec{S})$, we analyzed the statistics of $\hat{I}(\vec{R},\vec{S})$ on a tractable neural population where $I(\vec{R},\vec{S})$ can be computed exactly. We let $N = 10$ and $f_n(\vec{s}) = \vec{\phi}_n \cdot \vec{s}$ with $\vec{\phi}_n$ uniformly distributed along the unit circle. $P(\vec{s})$ is a spherical, two-dimensional Gaussian distribution and $N_{\text{stim}} = 8,000$. We evaluate $I(\vec{R},\vec{S})$ exactly, and get that $I(\vec{R},\vec{S}) = 1.3384$ nats. This is well below the upper bound of $\log(8,000) \approx 8.987$ nats. We computed $\hat{I}(\vec{R},\vec{S})$ 100 times for $B = 1$ and $B = 3$, with $\{s_\mu\}$ fixed. For each repetition we record the residual $\hat{I}(\vec{R},\vec{S}) - I(\vec{R},\vec{S})$. Distribution plots of the residuals are shown in Figure 1. For both distributions the sample mean is not significantly different from zero with $P = 0.848$ ($B = 1$) and $P = 0.851$ ($B = 3$) in a two-sided t-test. The simulation results therefore support the derivation of zero-bias in the proposed model-based Monte-Carlo estimator.

4 Simplifying the mutual information with sufficient statistics

4.1 A vector-valued sufficient statistic

The method introduced in Section 3 can be applied for very general formulations and parametrizations of the activation functions. However, when we constrain the activation functions to be affine we can show that $P(\vec{r}\,|\vec{s})$ has especially useful properties. Specifically, we assume the following parametrization of $f_n(\vec{s})$:

$$f_n(\vec{s}) = \vec{w}_n \cdot \vec{s} - \alpha_n,\ \forall n \tag{12}$$

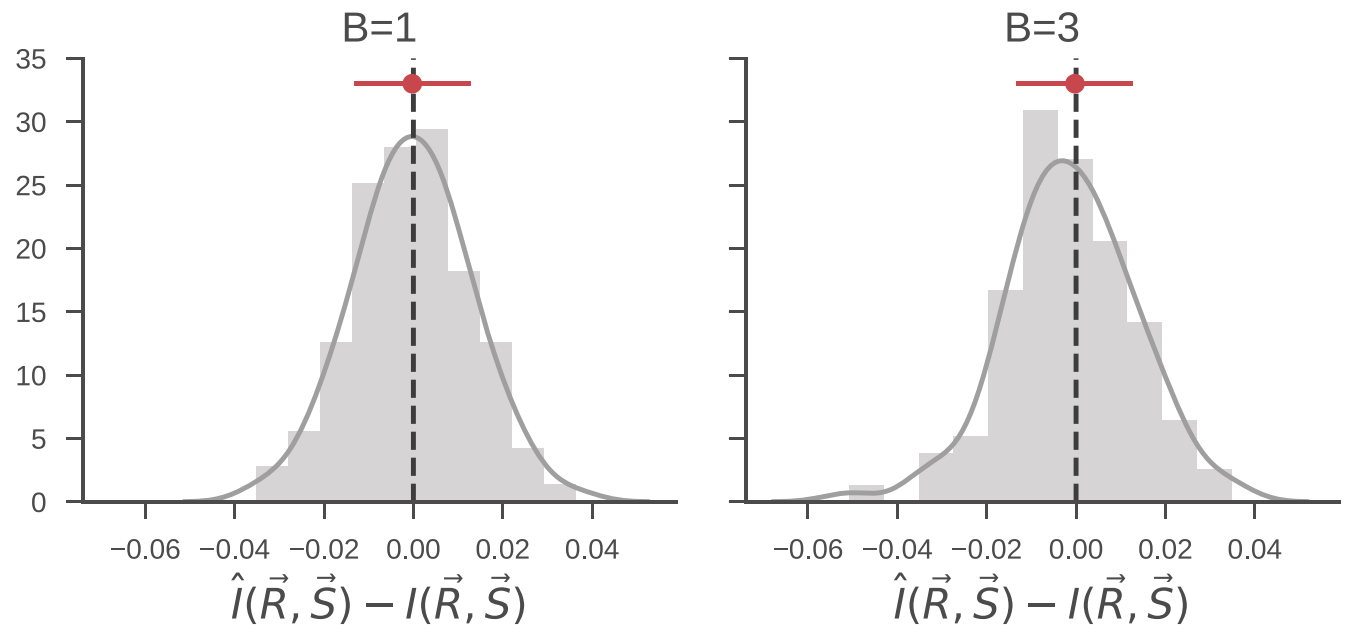

Figure 1: between exact calculation and the Monte Carlo results for the test neural population described in Section 3. Dashed black line indicates zero, while red marker and error bar are the sample mean and standard deviation.

While Eq. (12) implies a strong restriction on how stimuli drive the neural responses, some results of this section can be generalized to other activation functions. The reason for this is that even the general formulation of $P(\vec{r}|\vec{s})$ given in Eq. (3) can be viewed as an exponential family, with sufficient statistic $\vec{r}$ and natural parameter $\vec{f}(\vec{s})$. In particular, the framework can be extended to quadratic activation functions, which are an important model for describing neurons that are sensitive to multiple stimulus features. See Appendix E for further discussion.

If Eq. (12) holds, then Eq. (2) can be rewritten as follows:

$$P(\vec{r}|\vec{s}) = h(\vec{r}) \exp(\vec{s} \cdot \vec{t}(\vec{r}) - A(\vec{s})), \tag{13}$$

where,

$$\vec{t}(\vec{r}) = \mathbf{W} \cdot \vec{r}, \quad \mathbf{W} \equiv (\vec{w}_1^T, ..., \vec{w}_N^T) \tag{14}$$

$$h(\vec{r}) = e^{-\sum_n r_n \alpha_n}, \tag{15}$$

$$A(\vec{s}) = \sum_n \log\left(2\cosh(\vec{w}_n \cdot \vec{s} - \alpha_n)\right). \tag{16}$$

Equation (13) is an exponential family with sufficient statistic $\vec{t} \in \mathbb{R}^D$, natural parameter $\vec{s}$, base measure $h(\vec{r})$, and log-partition function $A(\vec{s})$ (Wainwright and Jordan, 2008).

The stimulus-conditional probability distribution $P(\vec{t}\,|\vec{s})$ can be defined by marginalizing over all $\vec{r}$ that map to the same $\vec{t}$:

$$\begin{aligned} P(\vec{t}\,|\vec{s}) &= \sum_{\vec{r}} \delta\left(\vec{t}, \vec{t}\,(\vec{r})\right) h(\vec{r}) \exp(\vec{s} \cdot \vec{t}\,(\vec{r}) - A(\vec{s})) \\ &= \exp(\vec{s} \cdot \vec{t} - A(\vec{s})) \sum_{\vec{r}} \delta\left(\vec{t}, \vec{t}\,(\vec{r})\right) h(\vec{r}) \\ &= \exp(\vec{s} \cdot \vec{t} - A(\vec{s})) h(\vec{t}). \end{aligned} \tag{17}$$

Note that $h(\vec{t}) = 0$ if there does not exist an $\vec{r}$ such that $\vec{t} = \vec{t}\,(\vec{r})$. An important property of sufficient statistics is the conservation of information (Cover and Thomas, 2012):

$$I(\vec{S}, \vec{R}\,) = I_{\text{vector}}(\vec{S}, \vec{T}\,) \tag{18}$$

with $\vec{T}$ defined by Eq. (14). Although $\vec{T}$ does not lose information relative to $\vec{r}$, it is worth making a few comments on $\vec{T}$ and Eq. (18). Because $\vec{R}$ is a discrete variable (with cardinality of at most 2^N) and $\vec{T}$ is a deterministic function of $\vec{R}$, then $\vec{T}$ is also a discrete variable with finite cardinality. Indeed, outside of cases of degeneracy between the columns of $\mathbf{W}$, there will generally be a one-to-one mapping between values of $\vec{R}$ and $\vec{T}$. Thus, even in cases where $D \ll N$, computing $H(\vec{T}\,)$ can be just as difficult as computing $H(\vec{R}\,)$. Furthermore, unlike $\vec{R}$, the components of $\vec{T}$ will generally not be conditionally independent. $H(\vec{T}\,|\vec{s}\,)$ will thus be similarly intractable. While it may seem that we have not gained any computational advantage by transforming from $\vec{R}$ to $\vec{T}$ we will now show that Eq. (18) can be expressed in a convenient form that facilitates several useful approximations.

4.2 Decomposition of mutual information based on sufficient statistics

We start by noting that the ordering of the components of $\vec{S}$ and $\vec{T}$ is arbitrary, because applying any matching permutation to the components of $\vec{S}$ and $\vec{T}$ does not affect $I(\vec{S}, \vec{R}\,)$. We will use the following notations for components of vectors $\vec{s} = (s_1, ..., s_D)$: $s_{\neg d} = (s_1, ..., s_{d-1}, s_{d+1}, ...s_D)$, $s_{<d} = (s_1, ..., s_{d-1})$,

and similarly for $s_{>d}$, $s_{\geq d}$, and $s_{\leq d}$. Note that $S_{\neg d} = (S_{<d}, S_{>d})$. Additionally, we will at times consider information theoretic quantities involving variables that are the concatenation of two other variables, such as X and Y. Such compound variables will be denoted as $\{X, Y\}$. Using these notations and applying the chain rule for mutual information to Eq. (18) yields: (Cover and Thomas, 2012):

$$I_{\text{vector}}(\vec{S}, \vec{T}\,) = \sum_{d=1}^{D} I(S_d, \vec{T}\,|S_{<d}). \tag{19}$$

In Eq. (19), $I(S_d, \vec{T}\,|S_{<d})$ is the mutual information between $\vec{T}$ and S_d conditioned on $S_{<d}$.

$$\begin{aligned} I(S_d, \vec{T}\,|S_{<d}) &= H(S_d, S_{<d}) + H(\vec{T}, S_{<d}) - H(S_d, \vec{T}, S_{<d}) - H(S_{<d}) \\ &= TC(S_{<d}, S_d, \vec{T}\,) - I(S_{<d}, S_d) - I(S_{<d}, \vec{T}\,) \\ &= I(\vec{T}, \{S_{<d}, S_d\}) - I(\vec{T}, S_{<d}) = I(\vec{T}, S_{\leq d}) - I(\vec{T}, S_{<d}) \\ &= \langle I(S_d, \vec{T}\,|s_{<d})\rangle_{s_{<d}}, \end{aligned} \tag{20}$$

where $TC(S_{<d}, S_d, \vec{T}\,)$ is the total correlation between $S_{<d}$, S_d, and $\vec{T}$. All four formulations of $I(S_d, \vec{T}\,|S_{<d})$ are equivalent provided they all exist. A notable situation is when there is functional dependence between S_d and $S_{<d}$, such as when the support of $S_{\leq d}$ lies on a manifold of intrinsic dimension $< d$. In this case $I(S_{<d}, S_d)$ diverges and the second line of Eq. (20) is ill-defined. However, it easy to show that $I(S_d, \vec{T}\,|S_{<d}) = 0$ if such a functional dependency exists using the fourth line of Eq. (20). Formally, let $s_d \equiv g(s_{<d})$ where $g(s_{<d}) : \mathcal{R}^{d-1} \to \mathcal{R}$ is a function defined explicitly or implicitly. Then we note that the mutual information between two variables is zero if at least one variable is constant:

$$I(S_d, \vec{T}\,|s_{<d}) = I(g(s_{<d}), \vec{T}\,|s_{<d}) = 0. \tag{21}$$

Thus, $I(S_d, \vec{T}\,|S_{<d})$ will also be zero:

$$I(S_d, \vec{T}\,|S_{<d}) = \langle I(S_d, \vec{T}\,|s_{<d})\rangle_{s_{<d}} = \langle I(g(s_{<d}), \vec{T}\,|s_{<d})\rangle_{s_{<d}} = 0. \tag{22}$$

Computing just $I(S_d, \vec{T}\,|s_{<d})$ remains challenging for the same reasons as computing $I_{\text{vector}}(\vec{S}, \vec{T}\,)$. However, we can achieve a further reduction by taking advantage of the fact that $P(\vec{t}\,|\vec{s}\,)$ is an exponential

family. To see this we first express two important marginalized forms of (14):

$$P(\vec{t}|s_d, s_{<d}) = h(\vec{t}) \exp(s_{<d} \cdot t_{<d} + s_d t_d) \langle \exp(s_{>d} \cdot t_{>d} - A(\vec{s})) \rangle_{s_{>d}|s_{\leq d}}. \tag{23}$$

$$P(\vec{t}|s_{<d}) = h(\vec{t}) \exp(s_{<d} \cdot t_{<d}) \langle \exp(s_{\geq d} \cdot t_{\geq d} - A(\vec{s})) \rangle_{s_{\geq d}|s_{<d}}. \tag{24}$$

The notation $\langle f(\vec{s}) \rangle_{s_{>d}|s_{\leq d}}$ denotes the expectation of $f(\vec{s})$ with respect to $P(s_{>d}|s_{\leq d})$, with analogous meanings for $\langle f(\vec{s}) \rangle_{s_{\geq d}|s_{<d}}$ and so forth. Marginalization over conditioned variables is expressed implicitly: $P(\vec{t}|s_{<d}) = \langle P(\vec{t}|\vec{s}) \rangle_{s_{\geq d}}$. The important consequence of (23) and (24) is that the log-likelihood ratio of $P(\vec{t}|s_d, s_{<d})$ and $P(\vec{t}|s_{<d})$ is independent of $t_{<d}$. From this we can show that $I(S_d, \vec{T}\,|s_{<d}) = I(S_d, T_{\geq d}|s_{<d})$:

$$\begin{aligned} I(S_d, \vec{T}\,|s_{<d}) &= \left\langle \sum_{\vec{t}} P(\vec{t}|s_d, s_{<d}) \log \left(\frac{P(\vec{t}|s_d, s_{<d})}{P(\vec{t}|s_{<d})} \right) \right\rangle_{s_d} \\ &= \left\langle \sum_{t_{\geq d}} P(t_{\geq d}|s_d, s_{<d}) \log \left(\frac{P(t_{\geq d}|s_d, s_{<d})}{P(t_{\geq d}|s_{<d})} \right) \right\rangle_{s_d}. \\ &= I(S_d, T_{\geq d}|s_{<d}) \end{aligned} \tag{25}$$

This leads to our next reduction:

$$I_{\text{vector}}(\vec{S}, \vec{T}\,) = \sum_{d=1}^{D} I(S_d, T_{\geq d}|S_{<d}). \tag{26}$$

We note that the d^{th} term in (19) has $D+d+1$ degrees of freedom, whereas the corresponding term in (26) has $D+1$ degrees of freedom. This effective dimension reduction has important algorithmic implications for the nonparametric estimators we use to compute the individual terms of (26) (c.f. Section 5). In section 5.1 and 5.2, we use Eq. (26) to evaluate $I_{\text{vector}}(\vec{S}, \vec{T}\,)$.

4.3 Lower bounds for the mutual information

While Eq. (26) represents a significant improvement in complexity over naive evaluation of $I(\vec{S}, \vec{T}\,)$, individual terms of $I_{\text{vector}}(\vec{S}, \vec{T}\,)$ may be still too high dimensional to reliably evaluate. In this section, we will present a series of lower bounds on $I_{\text{vector}}(\vec{S}, \vec{T}\,)$ that are more easily estimated. In particular we consider bounds that arise by replacing $T_{\geq d}$ in the d^{th} term of Eq (26) by a lower dimensional,

deterministic transformation of $T_{\geq d}$ denoted Z_d. Applying the Data Processing Inequality (DPI) to each term in Eq (26), will yield a lower bound for the mutual information. There are many possible lower bounds to $I_{\text{vector}}(\vec{S}, \vec{T})$ of this form. We focus on a variable $Z_d = \{T_d, |T_{>d}|\}$ where $|T_{>d}|$ is the L_2-norm of $T_{>d}$. This leads to the following lower bound approximation to $I_{\text{vector}}(\vec{S}, \vec{T})$

$$I_{\text{iso}}(\vec{S}, \vec{T}) = \sum_d I(S_d, \{T_d, |T_{>d}|\}\,|S_{<d}), \tag{27}$$

which we term isotropic. In Appendix B, we show that this approximation becomes exact in the asymptotic limit of large neural populations, meaning that $I_{\text{iso}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\vec{S}, \vec{T})$, when the stimulus distribution is isotropic and the distribution of receptive fields (RF) $\vec{w}$ across the population is such that $A(\vec{s}) = A(|\vec{s}|)$. Notably, this is achieved when RFs uniform cover the stimulus space, meaning that $P(\vec{w})$ is described by an uncorrelated Gaussian distribution. For finite number of neurons, $A(\vec{s})$ will never be perfectly isotropic. However, for large populations ($N \gg 1$) where the receptive fields $\vec{w}$ are drawn from an isotropic distribution and the distribution of α is independent of $\vec{w}$, $A(\vec{s})$ will become isotropic asymptotically as $N \to \infty$, cf. appendix B.1 for further details. The analogue of approximation Eq. (27) for the case where RFs and $\vec{s}$ are described by a matching correlated Gaussian distribution is described in appendix B.2.

The next reduction we consider is to drop $|T_{>d}|$ from each term of Eq. (27):

$$I_{\text{comp-cond}}(\vec{S}, \vec{T}) = \sum_{d=1}^{D} I(S_d, T_d|S_{<d}). \tag{28}$$

By the data-processing inequality, it again follows that $I_{\text{iso}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$. Overall, one obtains a series of bounds:

$$I_{\text{vector}}(\vec{S}, \vec{T}) \geq I_{\text{iso}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T}). \tag{29}$$

Our final, simplest approximation is to drop the conditioning on $S_{<d}$ in each term of Eq. (28).

$$I_{\text{comp-ind}}(\vec{S}, \vec{T}) = \sum_{d=1}^{D} I(S_d, T_d). \tag{30}$$

We show in Appendix C, that this last approximation becomes exact in the case where neural populations split into independent sub-populations with orthogonal RFs between sub-populations. Mathematically,

this corresponds to the case where both the stimulus distribution $P(\vec{s})$ and the function $A(\vec{s})$ factor in the same basis:

$$P(\vec{s}) = \prod_{k=1}^{D} P(s'_k), \quad A(\vec{s}) = \sum_{k=1}^{D} A(s'_k). \tag{31}$$

In general, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ may be greater or less than $I_{\text{comp-cond}}(\vec{S}, \vec{T})$ (or $I_{\text{vector}}(\vec{S}, \vec{T})$) (Renner and Maurer, 2002). However when $P(\vec{s}) = \prod_d P(s_d)$, the following additional inequality holds:

$$I_{\text{comp-cond}}(\vec{S}, \vec{T}) \geq I_{\text{comp-ind}}(\vec{S}, \vec{T}). \tag{32}$$

To derive (32) we first note that we can decompose $I(S_d, \{T_d, S_{<d}\})$ (for $d > 1$) in two different ways:

$$\begin{aligned} I(S_d, \{T_d, S_{<d}\}) &= I(S_d, T_d | S_{<d}) + I(S_d, S_{<d}) \\ &= I(S_d, S_{<d} | T_d) + I(S_d, T_d) \end{aligned} \tag{33}$$

Equating the first and second lines of (33) we can rewrite the residual $I(S_d, T_d|S_{<d}) - I(S_d, T_d)$:

$$I(S_d, T_d|S_{<d}) - I(S_d, T_d) = I(S_d, S_{<d}|T_d) - I(S_d, S_{<d}) \tag{34}$$

Though either side of (34) may be positive or negative in general, when we make the assumption that $P(\vec{s})$ factors across dimension, then $I(S_d, S_{<d}) = 0$. Thus (34) is nonnegative and $I(S_d, T_d|S_{<d}) \geq I(S_d, T_d)$, implying (32).

In the opposite extreme case where the value of S_d is a deterministic function of $S_{<d}$, Eq. (22) can be generalized to show that $I(S_d, T_d|S_{<d}) = 0$. Thus, in this case $I(S_d, T_d) \geq I(S_d, T_d|S_{<d})$, which in turn indicates that $I_{\text{comp-ind}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$. For example, when the support of $P(\vec{s})$ lies on a one-dimensional curve, e.g. $\vec{S}$ represents position along a one-dimensional nonlinear track, S_d is fully determined from the values of other variables $S_{<d}$ $\forall d$ regardless of component ordering. In this case, $I_{\text{comp-ind}}(\vec{S}, \vec{T}) \geq I_{\text{comp-cond}}(\vec{S}, \vec{T})$.

In the intermediate cases with some statistical dependencies between stimulus components, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ is not generally guaranteed to be a lower bound to either $I_{\text{comp-cond}}(\vec{S}, \vec{T})$ or $I_{\text{vector}}(\vec{S}, \vec{T})$. Nevertheless, we observed that even for some correlated $P(\vec{s})$ $I_{\text{comp-ind}}(\vec{S}, \vec{T}) < I_{\text{comp-cond}}(\vec{S}, \vec{T})$, c.f. section 5.2.

4.4 Alternative Approximations of $I(\vec{R}, \vec{S}\,)$

Previous authors have proposed other approximations to the mutual information. There exists a non-parametric *upper* bound to the mutual information computed in terms of pairwise relative entropies between $P(\vec{r}\,|\vec{s}\,)$ and $P(\vec{r}\,|\vec{s}\,')$ (Haussler et al., 1997; Kolchinsky et al., 2017):

$$I_{\text{k-w}}(\vec{R}, \vec{S}\,) = -\int d\vec{s} P(\vec{s}\,)\, \log\left(\int d\vec{s}\,' P(\vec{s}\,') \exp\left(-D_{KL}(P(\vec{r}\,|\vec{s}\,)||P(\vec{r}\,|\vec{s}\,'))\right)\right) \tag{35}$$

The model we consider for $P(\vec{r}\,|\vec{s}\,)$ is an exponential family and thus has a tractable relative entropy (Banerjee et al., 2005):

$$D_{KL}(P(\vec{r}\,|\vec{s}\,)||P(\vec{r}\,|\vec{s}\,') = \sum_n (\tanh(f_n(\vec{s}\,))f_n(\vec{s}\,) - A_n(\vec{s}\,)) - (\tanh(f_n(\vec{s}\,))f_n(\vec{s}\,') - A_n(\vec{s}\,')) \tag{36}$$

In Eq. (35) we have used the generalized definitions of Section 3. The evaluation of the upper bound (35) is quadratic in the sample size N_{stim} as opposed to $O(N_{\text{stim}} \log^2 N_{\text{stim}})$ for the estimator in section 5. In the limit where $N \gg D$, another popular approximation exist based on Fisher information (Brunel and Nadal, 1998); it can be computed with $O(N_{\text{stim}})$ operations. Recent work has shown that this approximation is valid only for certain classes of input distributions (Huang and Zhang, 2018). In appendix D we discuss the relationship between this approximation and $I_{\text{Fisher}}(\vec{R}, \vec{S}\,)$. We include numerical comparisons between $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$ and the methods proposed in this paper in sections 5.1 and 5.2. However, we found that the Fisher Information approximation drastically overestimated the true mutual information. Therefore to avoid obscuring differences between other results, the approximation based on Fisher Information is not included in Figures 2-3. Full plots including this approximation can be found in Appendix D.

We note that there are other variational approximations to mutual information (Belghazi et al., 2018; Barber and Agakov, 2003). However, because comparing the information for different choices of the parameters of $P(\vec{r}\,|\vec{s}\,)$ and $P(\vec{s}\,)$ requires training a different variational approximation each time, direct comparison requires substantial computational resources and we leave them for future work.

5 Numerical Simulations

We now test the performance of the above described bounds under several representative situations that include correlated and uncorrelated stimulus distributions, and isotropic and anisotropic receptive field distributions, including experimentally recorded receptive fields from the primary visual cortex, as well as the case where intrinsic "noise" correlations are present.

To empirically estimate the bounds on mutual information $[I(S_d, T_d)$, $I(S_d, T_d|S_{<d})$, $I(S_d, \{T_d, |T_{>d}|\}\,|S_{<d})$, and $I(S_d, T_{\geq d}|S_{<d})]$, we use the KSG estimator (Kraskov et al., 2004), a non-parametric method based on distributions of K nearest-neighbor distances. We chose to use the KSG estimator because even though we have reduced the mutual information between two high-dimensional variables into a sum over pairs of scalars, computing even just $I(S_d, T_d)$ can still be a daunting task, even more so for terms involving conditional informations. T_d may still have exponentially large cardinality, and complicated interdependencies between components of $\vec{T}$ present difficulties in forming explicit expressions for $P(t_d|s_d)$, so exact evaluation of $H(T_d)$ and $H(T_d|S_d)$ is not feasible at present. The KSG estimator requires only that we can draw N_{stim} samples of $\vec{S}$ and $\vec{T}$ from $P(\vec{s}, \vec{t})$, discarding the unused components. Sampling from $P(\vec{s}, \vec{t})$ is easily done given samples from $P(\vec{s})$. Given a sample $\vec{s}$, we draw $\vec{r}$ from $P(\vec{r}|\vec{s})$, which is easily done because of Eq. (2), and transform $\vec{r}$ into $\vec{t}$ using (14). This estimator has complexity $O(N_{\text{stim}} \log^2 N_{\text{stim}})$ when implemented with KD-Trees. For the case of two scalar variables the ℓ_2 error of the estimate decreases like $1/\sqrt{N_{\text{stim}}}$ (Gao et al., 2018), though if the true value of the mutual information is very high then the error may still be large (Gao et al., 2015). In order to partially alleviate this error, we use the PCA based Local Nonuniformity Correction of (Gao et al., 2015) (KSG-LNC). We extend this estimator to compute the conditional mutual information terms using a decomposition analogous to the second line of Eq. (20), and set the nonuniformity threshold hyperparameter according to the heuristics suggested in (Gao et al., 2015). Additionally, we assume that the distribution of $\vec{S}$ (and thus $S_{\leq d}$, $S_{<d}$, and S_d $\forall d$) is non-atomic. Thus, because $\vec{T}$ is discrete but real valued, the KSG estimator is applicable as neither $\vec{T}$ nor $\vec{S}$ is a mixed continuous-discrete variable (Gao et al., 2017).

5.1 Large populations responding to uncorrelated stimuli

We evaluated the performance of the bounds on information developed in section 4.3 for large populations ranging from $N \approx 100$ to $1,000$. Specifically, to test the performance of $I_{\text{iso}}(\vec{S}, \vec{T})$ we chose a highly

isotropic population and stimulus distribution. We set $D = 3$, $M = 8,000$, and let $P(\vec{s})$ be a zero mean gaussian with unit covariance matrix. For each value of N, the $\vec{w}_n$ were placed uniformly on the surface of the unit sphere, using the regular placement algorithm of (Deserno, 2004). Because N is too large for exact evaluation of $H(\vec{r})$ ground truth values were estimated using the Monte Carlo estimator $\hat{I}(\vec{R}, \vec{S})$ of section 3 with $B = 3$. Results are plotted in Figure 2. We find that for large N, $I_{\text{iso}}(\vec{S}, \vec{T})$ tightly approximates $I_{\text{vector}}(\vec{S}, \vec{T})$ and both are accurate approximations to $\hat{I}(\vec{R}, \vec{S})$, strongly outperforming $I_{\text{k-w}}(\vec{R}, \vec{S})$. We note that for this case the upper bound of $\log(N_{\text{stim}}) = \log(8,000) \approx 9$ (nats) is well above all of the curves other than $I_{\text{k-w}}(\vec{R}, \vec{S})$, which is already known to be an upper bound to $I(\vec{R}, \vec{S})$, demonstrating that we are in the well-sampled regime. Once again we see that inequalities (29) and (32) hold.

5.2 Correlated stimulus distributions

We now consider the case of correlated Gaussian stimuli. and model $P(\vec{s})$ as a zero-mean Gaussian with a full-rank non-diagonal covariance matrix $\mathbf{C}$. To better understand the effects of stimulus correlations we also perform computations in stimulus bases where components are independent. For this, we decompose $\mathbf{C}$ as $\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$ where $\mathbf{V}$ is an orthogonal matrix whose columns are the eigenvectors of $\mathbf{C}$.

$$\hat{S} = \mathbf{V}^T \vec{S}, \quad \hat{T} = \mathbf{V}^T \vec{T}. \tag{37}$$

Note that we have $\hat{t} \cdot \hat{s} = \vec{t} \cdot \vec{s}$. It is easy to see that $P(\hat{t}|\hat{s})$ is also an exponential family. Additionally because the mappings from $\vec{s}$ to $\hat{s}$ and $\vec{t}$ to $\hat{t}$ are diffeomorphisms the information is preserved:

$$I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\hat{S}, \hat{T}). \tag{38}$$

We note that while Eq. (38) holds in principle, in practice we may see variation as the KSG family of estimators is not invariant to under diffeomorphisms. Importantly we also note that Eq. (25) holds for $(\hat{S}, \hat{T})$. Given samples from $P(\vec{s}, \vec{t})$, we automatically have samples from $P(\hat{s}, \hat{t})$. We can straightforwardly generalize $I_{\text{vector}}(\vec{S}, \vec{T})$, $I_{\text{comp-cond}}(\vec{S}, \vec{T})$, $I_{\text{comp-ind}}(\vec{S}, \vec{T})$, and $I_{\text{iso}}(\vec{S}, \vec{T})$ to $I(\hat{S}, \hat{T})$, $I_{\text{comp-cond}}(\hat{S}, \hat{T})$, $I_{\text{comp-ind}}(\hat{S}, \hat{T})$, and $I_{\text{iso}}(\hat{S}, \hat{T})$ respectively. Eq. (38) does not generalize to I_{iso}, $I_{\text{comp-cond}}$, or $I_{\text{comp-ind}}$ as they are not expressible as mutual information quantities between two variables. We note that

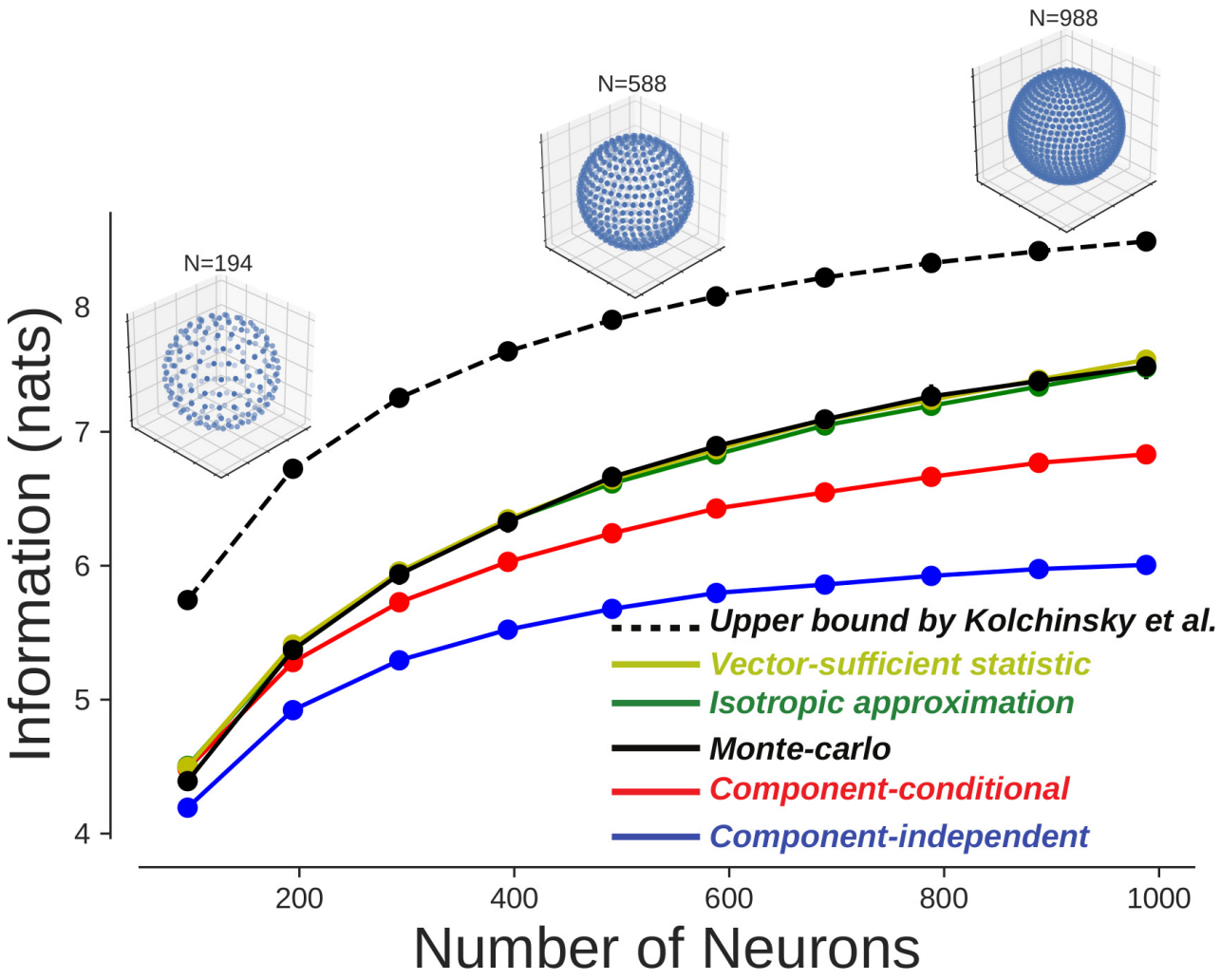

Figure 2: Information curves for neural populations with uncorrelated RF and stimulus distributions. Lines and error bars are mean and standard deviation over ten repeats of the estimator. Insets show RF distribution for several population sizes.

$I(\vec{R}, \vec{S}) = I(\vec{R}, \hat{S})$ so we do not modify $\hat{I}(\vec{R}, \vec{S})$.

Simulations in Figure 3 were done using the following stimulus covariance matrix

$$\mathbf{C} = \begin{pmatrix} 1.74716093 & 1.3103707 & 0.87358046 \\ 1.3103707 & 1.74716093 & 1.3103707 \\ 0.87358046 & 1.3103707 & 1.74716093 \end{pmatrix}. \tag{39}$$

For this choice of $\mathbf{C}$ $\rho_{1,2} = \rho_{2,3} = 0.75$, $\rho_{1,3} = 0.5$, and $|\mathbf{C}| = 1$. The covariance matrix of $\hat{S}$ is diagonal:

$$\hat{\mathbf{C}} = \begin{pmatrix} 0.283 & 0 & 0 \\ 0 & 0.868 & 0 \\ 0 & 0 & 4.032 \end{pmatrix}. \tag{40}$$

The receptive field configurations are the same as in Section 5.1. We note that in the $\vec{s}$ coordinates, all components have the same variance, whereas this symmetry is broken in the decorrelated components $\hat{s}$. We compared sorting the components of $\hat{s}$ in increasing and decreasing order of variance (triangles and squares respectively) in Figure 3A-C. Component order does not matter for $I_{\text{comp-ind}}(\hat{S}, \hat{T})$, Figure 3D. We find that for both I_{vector} and I_{iso} it is optimal to perform computation in the original basis, with both quantities accurately matching $\hat{I}(\vec{R}, \vec{S})$. For $I_{\text{comp-cond}}$ and $I_{\text{comp-ind}}$ accuracy is increased by using decorrelated components and, for $I_{\text{comp-cond}}$, sorting components by decreasing variance.

5.3 Highly asymmetric receptive field distributions

Next, we consider a small population ($N = 9$) with a highly asymmetric distribution of redundant receptive fields in two stimulus dimensions. In particular we are interested in a population where many different configurations of $\vec{R}$ map to the same configuration of $\vec{T}$, demonstrating the utility of using $\vec{T}$ as a non-trivial sufficient statistic of $\vec{R}$. With this in mind, we chose a heavily redundant configuration of $\{\vec{w}_n\}$: $\vec{w}_n = (0, 1)$ $(n = 1, 2, 3)$, $\vec{w}_n = (1, 0)$ $(n = 4, 5, 6)$, $\vec{w}_n = (1, 1)$ $(n = 7, 8, 9)$. The cardinalities of $\vec{R}$, $\vec{T}$, T_1 and T_2 are 512, 37, 7, and 7 respectively. Because N is small, ground truth values of $I(\vec{R}, \vec{S})$ were computed by exactly evaluating $P(\vec{r}\,|\vec{s})$ $\forall \vec{r} \in \{-1, 1\}^N$, for every sample of $\vec{s}$. Given $P(\vec{r}\,|\vec{s})$ we average across $\vec{s}$ to get $P(\vec{r})$ explicitly, and calculate $H(\vec{R})$, $H(\vec{R}\,|\vec{S})$, and $I(\vec{R}, \vec{S})$ from

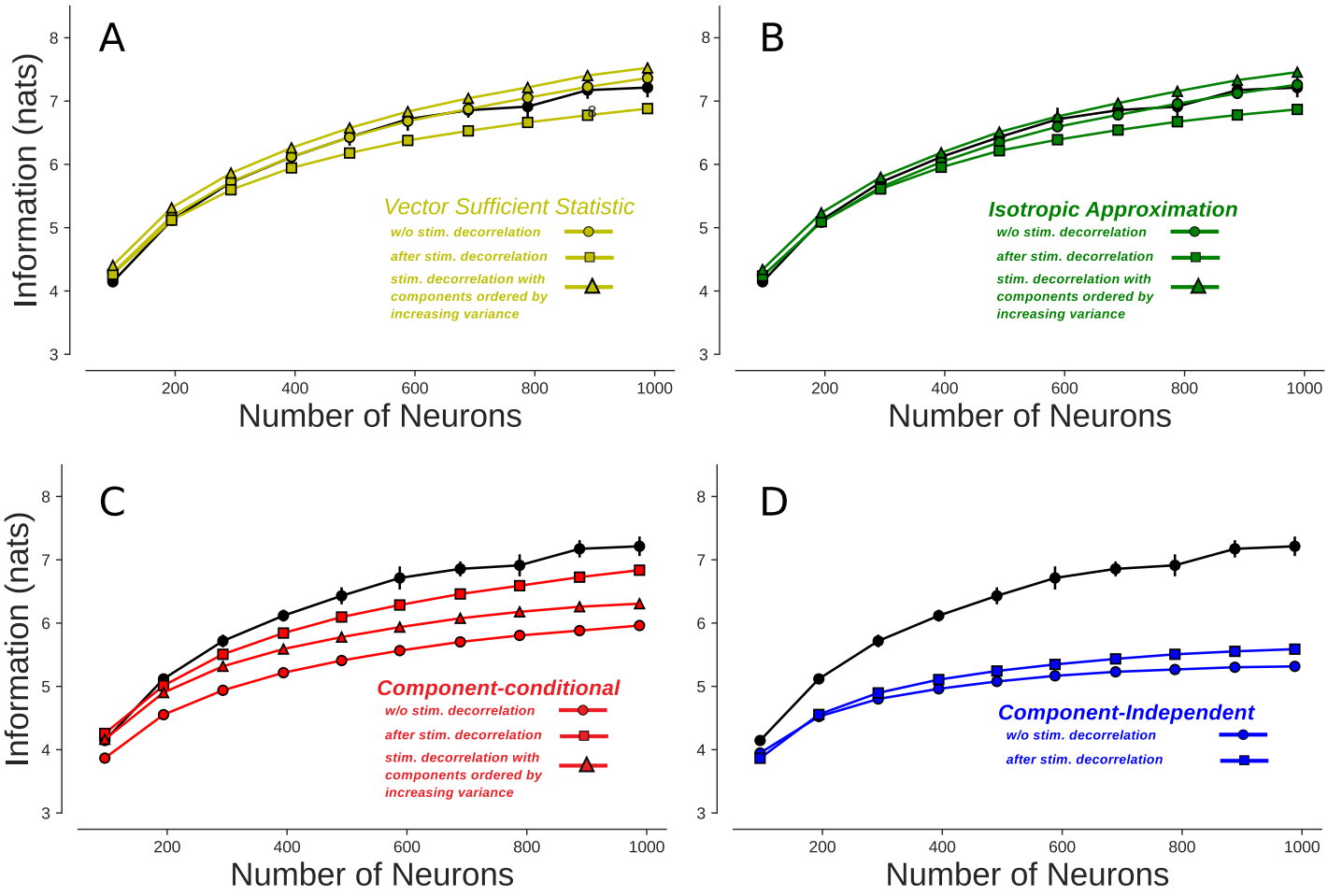

Figure 3: Information curves for neural populations with correlated RF distributions, cf. section 5.2. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. In all panels, circles represent information computed in the original basis, while squares and triangles are computations performed in decorrelated basis. I_{vector} (**A**) and I_{iso} (**B**) recover full information. These two computations do not benefit from working in the decorrelated stimulus basis. Stimulus decorrelation improves the performanceof $I_{\text{comp-cond}}$ (**C**) and $I_{\text{comp-ind}}$ (**D**)). In (**D**), computations are invariant to ordering of components.

these distributions. $P(\vec{s})$ is gaussianly distributed with diagonal covariance matrix, in accordance with (32). We set $N_{\text{stim}} = 10,000$. To test how the relative values of $I(S_1, T_1)$ and $I(S_2, T_2)$ impact the optimal ordering of components in $I_{\text{comp-cond}}(\vec{S}, \vec{T})$ we fixed $\sigma_2 = 1$ and varied σ_1 between 0.5 and 2.5. Results are plotted in Figure 4. As predicted, the hierarchy of bounds (29) and (32) holds for all σ_1. It is also notable that in this case, just like in the case of large neural populations, for computing $I_{\text{comp-cond}}(\vec{S}, \vec{T})$, it always seems best to start the information computation with the stimulus component that has the largest variance. As expected, the ordering or components does not strongly impact $I_{\text{vector}}(\vec{S}, \vec{T})$.

5.4 Experimental stimuli and receptive fields

In the previous three experimental sections we considered synthetic distributions of low-dimensional stimuli and artificial configurations of receptive fields. In this section we analyze a population of model neurons with receptive fields and α values that were fit to responses of primary visual cortex neurons (V1) elicited by natural stimuli (Sharpee et al., 2006). We use 147 pairs of $(\vec{w}_n, \alpha_n)$ values that were fit using the Maximally Informative Dimension (MID) algorithm as in (Sharpee et al., 2006). Stimuli are 10 pixel by 10 pixel patches extracted from the same set of images used to fit the model parameters. Receptive fields are normalized and centered on the patch, and we chose a 10×10 sub-patch of the original 32×32 shaped receptive fields so that all dimensions are well sampled by receptive fields. That is, for all pixels of the 10×10 patch, at least 115 of the 147 neurons have a nonzero value in the corresponding component of their receptive field. Additionally, the stimuli are z-scored by subtracting the mean and dividing by the standard deviation, with both quantities computed across all samples and pixels collectively.

Because of high stimulus dimensionality we could only compute the $I_{\text{comp-ind}}(\vec{S}, \vec{T})$ bound (via the KSG estimator) and the ground truth information (via the Monte Carlo method). Because the pixels of natural image patches are clearly not independent, we also computed $I_{\text{comp-ind}}$ in two additional coordinate systems. The first coordinate system is simply the linearly decorrelated components $\hat{S}$ and $\vec{T}$ defined in Eq. (37). The second coordinate systems uses independent components derived using independent component analysis on $\vec{S}$

$$\tilde{S} = \mathbf{U}\vec{S}, \quad , \tilde{T} = (\mathbf{U}^{-1})^T \vec{T} \tag{41}$$

Here, $\mathbf{U}$ is an unmixing matrix computed using Infomax ICA (Bell and Sejnowski, 1997) on the samples of $\vec{S}$. As is done in (Bell and Sejnowski, 1997), $\mathbf{U}$ includes a linear whitening matrix. We note that

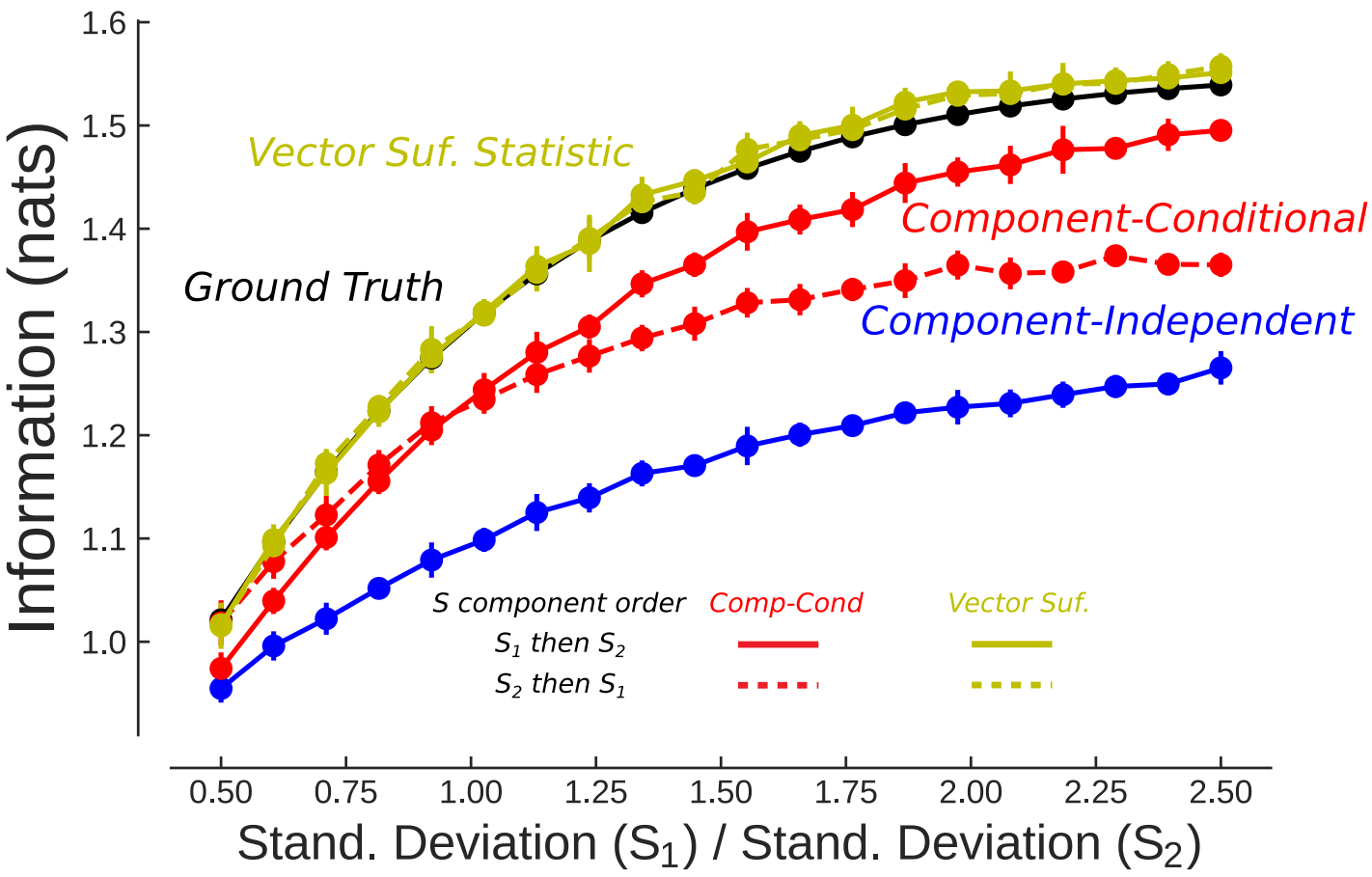

Figure 4: Information curves for the example population with highly redundant RFs from Sec. 5.3. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. Although neither the component-conditional nor the component information are guaranteed to tightly approximate the information, both provide good approximation to the full information (as estimated via unbiased Monte Carlo method), reaching values within $\geq 80\%$ of the maximum. For the vector-sufficient statistic, both component orderings accurately reproduced the full information. The next best approximation to the full information is provided by the component-conditional computation with components added in the order from largest to smallest variance. This approximation reaches accuracy within 95% of the full value over the range of neural population sizes.

in Eq. (41), $\vec{T}$ is multiplied by $(\mathbf{U}^{-1})^T$ and not $\mathbf{U}$ because the ICA unmixing matrix is generally not orthogonal and we require that $\vec{t}\cdot\vec{s} = \tilde{t}\cdot\tilde{s}$. As before $I_{\text{vector}}(\vec{S},\vec{T}) = I_{\text{vector}}(\hat{S},\hat{T}) = I_{\text{vector}}(\tilde{S},\tilde{T})$. However, the same cannot be said for $I_{\text{comp-ind}}$ expressed in different coordinate systems.

To evaluate the effect of using different coordinate systems to evaluate $I_{\text{comp-ind}}$ for different population sizes we first ranked the 147 neurons in descending order by the information each neuron carried about the stimulus. We computed $I(R_n,\vec{S})\ \forall n \in \{1,...,147\}$, which is easily done exactly since R_n is a binary variable, and then sorted neurons so that $I(R_n,\vec{S}) >= I(R_{n+1},\vec{S})\ \forall n$. We considered populations of size $N = 60, 70, ..., 140$, where each population contained the first N neurons under the aforementioned ordering. For each value of N we computed $\hat{I}(\vec{R},\vec{S})$ $(B = 3)$, $I_{\text{comp-ind}}(\vec{S},\vec{T})$, $I_{\text{comp-ind}}(\hat{S},\hat{T})$, and $I_{\text{comp-ind}}(\tilde{S},\tilde{T})$. We note that $\log(N_{\text{stim}}) = \log(49,152) \approx 10.8$ nats. Results are plotted in Figure 5.

We observe that both $I_{\text{comp-ind}}(\vec{S},\vec{T})$ and $I_{\text{comp-ind}}(\hat{S},\hat{T})$ overestimate the true information, especially $I_{\text{comp-ind}}(\vec{S},\vec{T})$. This overestimation occurs because stimulus components are not independent. By comparison, computation performed in the ICA basis, $I_{\text{comp-ind}}(\tilde{S},\tilde{T})$, lower bounds the mutual information for all N, achieving $\geq 75\%$ of the full information across the range of population sizes.

5.5 Handling intrinsically correlated neurons

In order to simplify derivations, we assumed that the neural responses were independent after conditioning on $\vec{s}$. However, all of the analytic results in Section 4.2 can be extended to specific forms of intrinsic interneuronal correlation to allow for the presence of correlations in neural responses for a given stimulus $\vec{s}$. Formally, we modify the base measure $h(\vec{r})$ to include a pairwise coupling term:

$$h(\vec{r},\mathbf{J}) = e^{\sum_{mn}\mathbf{J}_{mn}r_m r_n - \sum_n r_n\alpha_n}. \tag{42}$$

In Eq. 42, $\mathbf{J}$ is a symmetric $N \times N$ matrix where J_{mn} describes the intrinsic coupling between the m^{th} and n^{th} neurons. In this case $P(\vec{r}|\vec{s},\mathbf{J})$ can still be written as an exponential family in a canonical form:

$$\begin{aligned} P(\vec{r}|\vec{s}) &= h(\vec{r},\mathbf{J})\exp(\vec{s}\cdot\vec{t}(\vec{r}) - A(\vec{s},\mathbf{J})), \\ A(\vec{s},\mathbf{J}) &= \log\left(\sum_{\vec{r}} h(\vec{r},\mathbf{J})\exp(\vec{s}\cdot\vec{t}(\vec{r}))\right). \end{aligned} \tag{43}$$

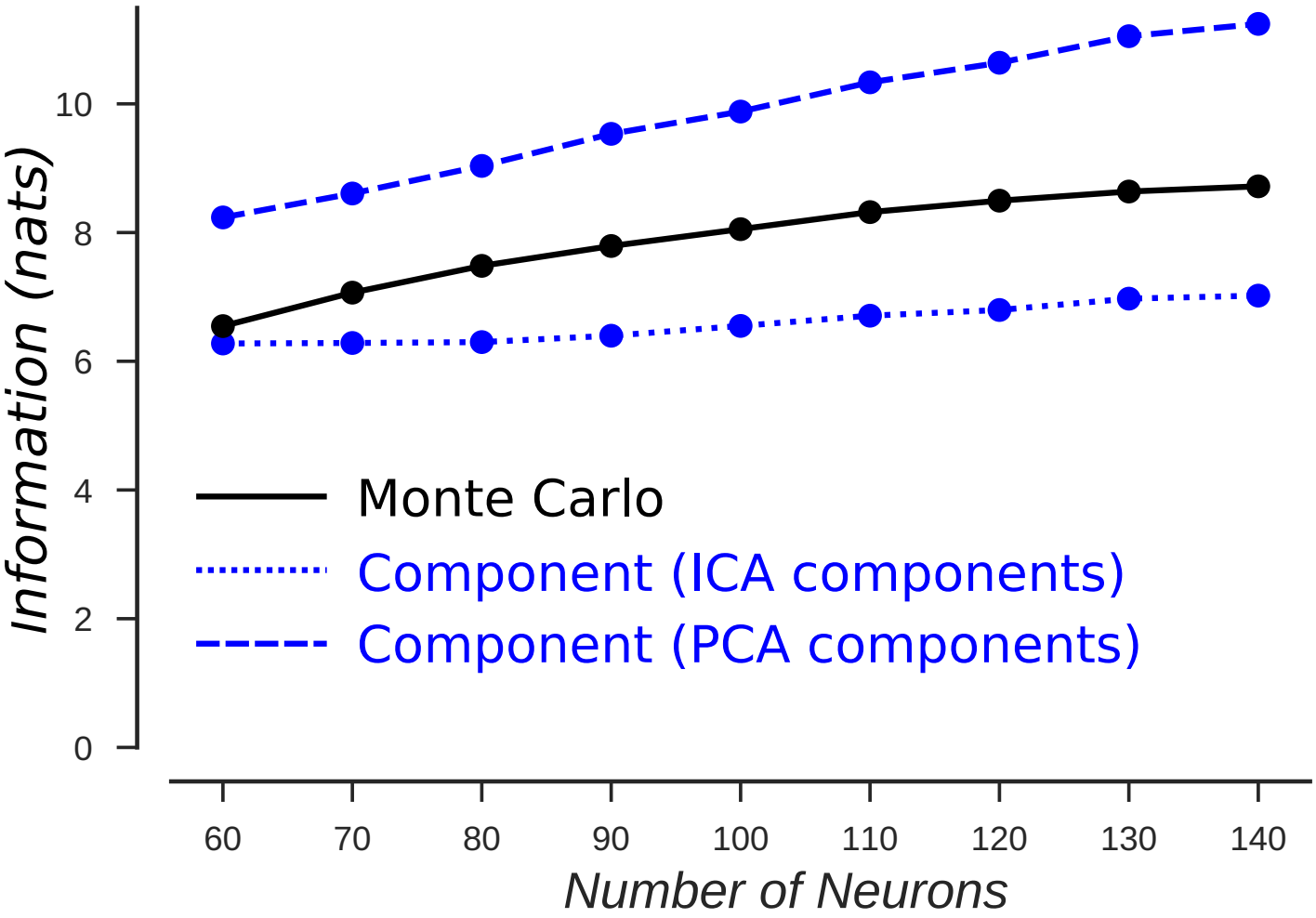

Figure 5: Information curves for populations based on experimentally recorded RFs and probed with $D = 100$ natural visual stimuli. The full information (solid black line) is computed via the Monte Carlo method and is compared to $I_{\text{comp-ind}}$ approximations computed in two different bases: the PCA basis (blue dashed line) and the ICA basis (blue dotted line). We do not show the calculation in the original pixel basis, because it$I_{\text{comp-ind}}(\vec{S}, \vec{T})$ is omitted as it was yielded values ~ 25 nats across the range of population sizes and obscured the other curves. Because of non-Gaussian statistics of natural stimuli, PCA components remain correlated, and as a result the approximation is no longer guaranteed to be a lower bound of the true information. Computation in the ICA basis respects the lower bound requirements, and achieves $\geq 75\%$ of the full information across the range of population sizes.

The form of the sufficient statistic remains unchanged, though $A(\vec{s}, \mathbf{J})$ generally lacks a closed form. Nevertheless, all of the decompositions, equalities, and inequalities in Section 4.2 require only for the exponential family to be in canonical form and remain valid.

We tested the accuracy of our approximations on a small ($N = 10$) population of intrinsically correlated neurons. Receptive fields are uniformly distributed on the unit circle, $\alpha_n = 0 \; \forall \, n$, and $P(\vec{s})$ is a standard two-dimensional Gaussian ($N_{\text{stim}} = 20,000$). Intrinsic coupling is set proportional to the overlap between receptive fields with a coupling strength J_0, the sign of which determines whether the intrinsic couplings perform stimulus decorrelation or error-correction (Tkačik et al., 2010).:

$$\mathbf{J}_{mn} = J_0 \, \vec{w}_m \cdot \vec{w}_n. \tag{44}$$

The algorithms of Sections 5 and 3 all depend on being able to sample easily from $P(\vec{r}\,|\vec{s})$. For large N and general $\mathbf{J}$ this is usually difficult, particularly for configurations of $\mathbf{J}$ that exhibit glassy dynamics. Additionally, evaluating Eq. (9) requires explicit knowledge of $A(\vec{s})$, though methods such as mean-field theory or the TAP approximation may be used to approximate $A(\vec{s})$ (Opper et al., 2001). Since this population is small, we evaluate the ground truth information exactly as in Section 5.3. Likewise, we sample $\vec{r}$ from $P(\vec{r}\,|\vec{s})$ exactly by computing all 2^N $(1,024)$ probabilities for every sample of $\vec{s}$. Analyzing the error introduced in using approximate sampling strategies such as Markov Chain Monte Carlo is left to future investigation. Results are plotted in Figure 6. As predicted, $I_{\text{vector}}(\vec{S}, \vec{T}\,)$ matches the ground truth values of the information. Similarly the hierarchy of bounds (29) and (32) is preserved, though for strongly negative couplings $I_{\text{comp-cond}}(\vec{S}, \vec{T}\,) \approx I_{\text{comp-ind}}(\vec{S}, \vec{T}\,)$. In sum, the presence of noise correlations does not invalidate the approximations and bounds that are derived above. However, numerical computation can become more difficult in the presence of noise correlations.

6 Conclusions and Future Work

We have presented three approximations that can be used to estimate the information transmitted by large neural populations. Each of these three approximations represents different trade-offs between accuracy and computational ease and feasibility. The best performance in terms of accuracy was provided by the isotropic approximation I_{iso}. This approximation worked well even in cases where is not

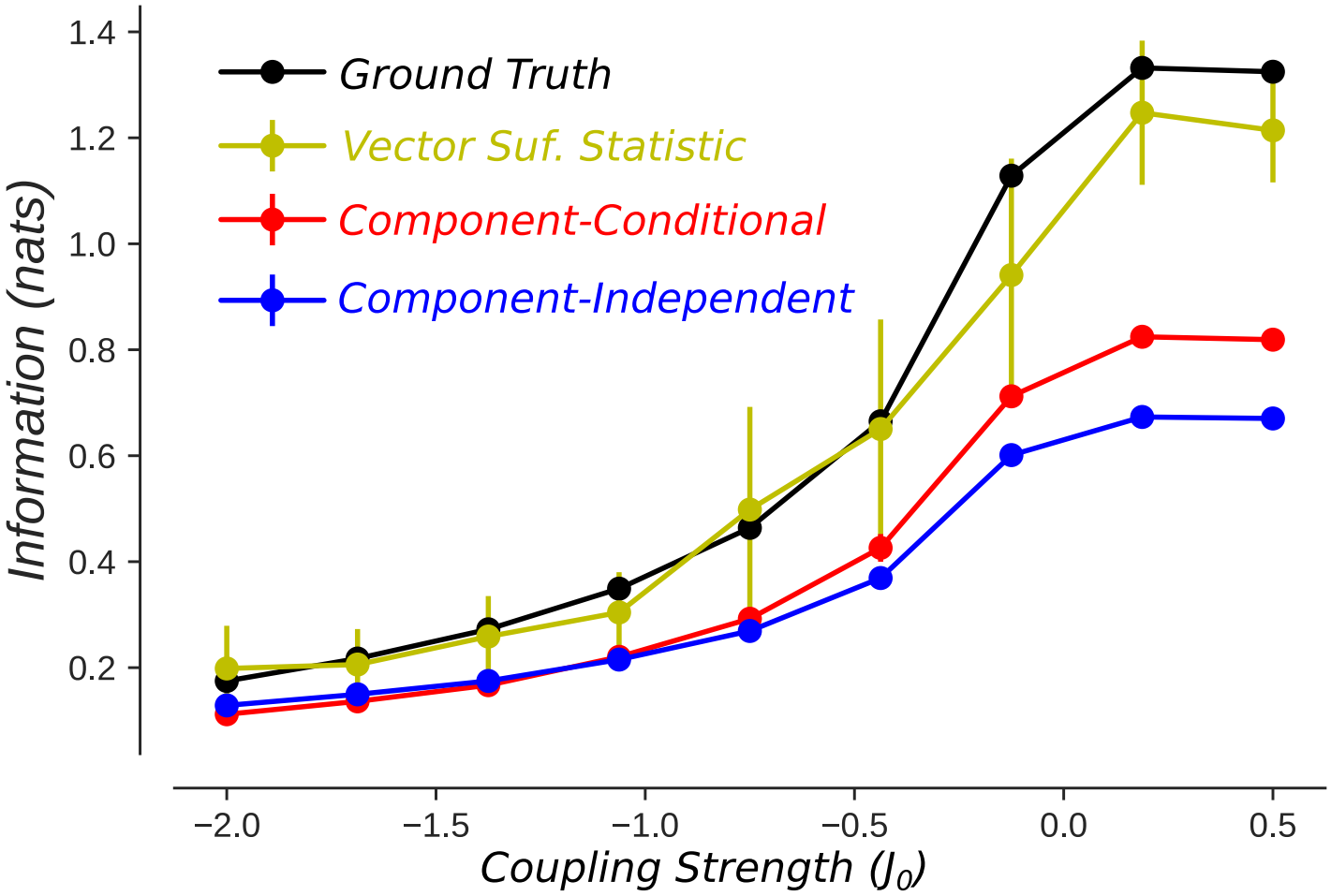

Figure 6: Information curves for populations with intrinsic correlations. Lines and errorbars are mean and standard deviation over ten repeats of the estimator.

guaranteed to become asymptotically exact with increasing population size. For example, the isotropic approximation was derived assuming a matching covariance matrix for both the stimulus and RF distributions, cf. Appendix B. Yet, this approximation matched the full ground-truth information values even for correlated stimuli with RFs remaining uncorrelated across the neural population (Fig. 3.) This approximation provided the best overall performance among the bounds tested, consistent with the theoretically expected inequalities between these bounds, cf. Eq. (29).

The component-conditional information $I_{\text{comp-cond}}$ offered the second-best performance. This approximation performed especially well when computed in the stimulus basis where stimulus components were not correlated. This approximation is less computationally difficult compared to I_{iso}, because each conditional information is evaluated between just two quantities S_d and T_d compared to S_d and a conjunction of T_d and $|T_{>d}|$ as in I_{iso}. For this reason, the finite-sample bias of $I_{\text{comp-ind}}$ can also be less than I_{iso}, because bias in the evaluation of the mutual information is usually larger for higher-dimensional calculations.

The last approximation $I_{\text{comp-ind}}$ is the least accurate of the three approximation but is computationally the easiest. It is the only approximation among the three we considered here that we were able to implement in conjunction with high dimensional stimuli. This approximation becomes most accurate in the stimulus bases where stimulus components are independent. There is strong evidence that neural receptive fields are organized along the ICA components of natural stimuli (Bell and Sejnowski, 1997; Olshausen and Field, 2004; Smith and Lewicki, 2006). This raises the possibility that the approaches proposed here will fair well when applied to recorded neural responses. Indeed, we found that $I_{\text{comp-ind}} \geq 75\%$ of the full information value for large neural populations constructed using experimentally recorded RFs and probed with natural stimuli.

At present, the main limitation for computing the conditional approximations $I_{\text{comp-cond}}$ and $I_{\text{comp-iso}}$ is not the number of neurons but rather the stimulus dimensionality. For stimulus distributions where $P(s_{\geq d}|s_{<d})$ can be easily sampled from, such as Gaussian distributions, we can take advantage of the fourth line of Eq. (20) to compute unbiased estimates of $I_{\text{comp-cond}}$ and I_{iso}, albeit with possibly high variance. Developing methods that can efficiently approximate these conditional computations represents an important opportunity for future research.

References

Banerjee, A., Merugu, S., Dhillon, I. S., and Ghosh, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749.

Barber, D. and Agakov, F. (2003). The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, pages 201–208. MIT Press.

Belghazi, I., Rajeswar, S., Baratin, A., Hjelm, R. D., and Courville, A. (2018). Mine: Mutual information neural estimation. *arXiv preprint arXiv:1801.04062*.

Bell, A. J. and Sejnowski, T. J. (1997). The "independent components" of natural scenes are edge filters. *Vision Res.*, 23:3327–3338.

Bender, C. and Orszag, S. (1999). *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. Advanced Mathematical Methods for Scientists and Engineers. Springer.

Berkowitz, J. and Sharpee, T. (2018). Decoding neural responses with minimal information loss. *bioRxiv*.

Bialek, W. (2012). *Biophysics: Searching for principles*. Princeton University Press, Princeton and Oxford.

Brenner, N., Strong, S. P., Koberle, R., Bialek, W., and de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comput*, 12:1531–1552.

Brunel, N. and Nadal, J. P. (1998). Mutual information, fisher information, and population coding. *Neural Comput.*, 10(7):1731–1757.

Cover, T. M. and Thomas, J. A. (2012). *Elements of information theory*. John Wiley and Sons.

Deserno, M. (2004). How to generate equidistributed points on the surface of a sphere. *P.-If Polymerforshung (Ed.)*, page 99.

Dettner, A., Münzberg, S., and Tchumatchenko, T. (2016). Temporal pairwise spike correlations fully capture single-neuron information. *Nature communications*, 7:13805.

Gao, S., Ver Steeg, G., and Galstyan, A. (2015). Efficient estimation of mutual information for strongly dependent variables. In *Artificial Intelligence and Statistics*, pages 277–286.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating mutual information for discrete-continuous mixtures. In *Advances in Neural Information Processing Systems*, pages 5986–5997.

Gao, W., Oh, S., and Viswanath, P. (2018). Demystifying fixed k-nearest neighbor information estimators. *IEEE Transactions on Information Theory*.

Haussler, D., Opper, M., et al. (1997). Mutual information, metric entropy and cumulative relative entropy risk. *The Annals of Statistics*, 25(6):2451–2492.

Huang, W. and Zhang, K. (2018). Information-theoretic bounds and approximations in neural population coding. *Neural computation*, (Early Access):1–60.

Kolchinsky, A., Tracey, B. D., and Wolpert, D. H. (2017). Nonlinear information bottleneck. *arXiv preprint arXiv:1705.02436*.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

Laughlin, S. B., de Ruyet van Steveninck, R. R., and Anderson, J. C. (1998). The metabolic cost of neural computation. *Nat. Neurosci.*, 41:36–41.

Nemenman, I., Bialek, W., and van Steveninck, R. d. R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111.

Olshausen, B. A. and Field, D. J. (2004). Sparse coding of sensory inputs. *Curr Opin Neurobiol*, 14(4):481–487.

Opper, M., Winther, O., et al. (2001). From naive mean field theory to the tap equations. *Advanced mean field methods: theory and practice*, pages 7–20.

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

Renner, R. and Maurer, U. (2002). About the mutual (conditional) information. In *Proc. IEEE ISIT*.

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

Schneidman, E., Berry II, M. J., Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440:1007=1012.

Sharpee, T. O., Sugihara, H., Kurgansky, A. V., Rebrik, S. P., Stryker, M. P., and Miller, K. D. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439(7079):936–942.

Smith, E. and Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, 439:978–82.

Strong, S. P., Koberle, R., van Steveninck, R. R. d. R., and Bialek, W. (1998). Entropy and information in neural spike trains. *Physical review letters*, 80(1):197.

Tkačik, G., Prentice, J. S., Balasubramanian, V., and Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences*, 107(32):14419–14424.

Treves, A. and Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Comp.*, 7:399–407.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Yu, Y., Crumiller, M., Knight, B., and Kaplan, E. (2010). Estimating the amount of information carried by a neuronal population. *Frontiers in computational neuroscience*, 4:10.

A Bias of $\hat{H}(\vec{R})$

In this section we give a self-contained proof that $\hat{H}(\vec{R})$ systematically underestimates the "true" entropy $H(\vec{R})$. We first assume that $P(\{\vec{s}_\mu\}) = \prod_\mu P(\vec{s}_\mu)$: The $\vec{s}_\mu$ are drawn independently from $P(\vec{s})$, whether $P(\vec{s})$ is a smooth density on $\mathcal{R}^D$ or some larger set of samples. We define a an empirical version of the marginal distribution on $P(\vec{r})$:

$$\hat{P}(\vec{r}|\{\vec{s}_\mu\}) = \frac{1}{M}\sum_\mu P(\vec{r}|\vec{s}_\mu) \tag{45}$$

The true marginal distribution is the expected value of $\hat{P}(\vec{r}|\{\vec{s}_\mu\})$: $\langle \hat{P}(\vec{r};\{\vec{s}_\mu\})\rangle_{P(\{\vec{s}_\mu\})} = P(\vec{r})$. The Shannon entropy is a concave function of $\hat{P}(\vec{r};\{\vec{s}_\mu\})$, which can be considered a random vector in the 2^N dimensional probability simplex. Thus, by Jensen's inequality we have the following:

$$\left\langle \hat{H}(\vec{R})\right\rangle_{P(\{\vec{s}_\mu\})} \leq H(\vec{R} : P = \langle \hat{P}\rangle) \equiv H(\vec{R}) \tag{46}$$

This bias holds even in the case of evaluating $\hat{H}(\vec{R})$ through exact enumeration. We note that we are able to produce unbiased estimates of $\hat{H}(\vec{R})$ because we have full access to $P(\vec{r}|\vec{s}_\mu)$: We can evaluate $P(\vec{r}|\vec{s}_\mu)$ explicitly and deterministically, and thus $\hat{F}(\vec{r})$ as well (up to factors of numerical precision). If we were always constrained to drawing samples from $P(\vec{r}|\vec{s}_\mu)$, then we would once again be limited to making biased estimates of $\hat{H}(\vec{R})$ (Paninski, 2003).

B On the asymptotic tightness of $I_{\mathbf{iso}}(\vec{S}, \vec{T})$

Consider a large population ($N \gg 1$) where the distribution of $\vec{w}$ and α is such that $A(\vec{s}) = A(|\vec{s}|)$ (in some sense to be made more precise later). Consider the likelihood ratio in the definition of (25):

$$\begin{aligned} \frac{P(t_{\geq d}|s_d, s_{<d})}{P(t_{\geq d}|s_{<d})} &= \frac{\langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(\vec{s}))\rangle_{s_{>d}|s_{\leq d}}}{\langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(\vec{s}))\rangle_{s_{\geq d}|s_{<d}}} \\ &= \frac{\langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\leq d}}}{\langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{\geq d}|s_{<d}}} \end{aligned} \tag{47}$$

Additionally we consider a stimulus distribution that is similarly isotropic so that the conditional distribution $P(s_{>d}|s_{\leq d})$ can be written in a convenient factored form:

$$\begin{aligned} P(\vec{s}) &= P(|\vec{s}|) = P\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right) \\ P(s_{>d}|s_{\leq d}) &= \frac{P(\vec{s})}{P(s_{\leq d})} = \frac{P\left(\sqrt{s_{<d}^2 + s_d^2 + s_{>d}^2}\right)}{P(s_{\leq d})} \end{aligned}$$

We will show that in this situation, we can replace $T_{\geq d}$ in (25) with the variable that is the concatenation of T_d and $|T_{>d}|$ without loss of information:

$$I(S_d, T_{\geq d}|s_{<d}) \approx I(S_d, \{T_d, |T_{>d}|\}\,|s_{<d}) \tag{48}$$

To show this using the Fisher-Neyman factorization theorem, it suffices to show that the numerator and denominator in (47) can be factored as follows:

$$\begin{aligned} \langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\leq d}} &= g_1(s_{<d}, s_d, t_d, |t_{>d}|) g_2(t_{>d}) \\ \langle \exp(s_d t_d + s_{>d} \cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{\geq d}|s_{<d}} &= f_1(s_{<d}, t_d, |t_{>d}|) f_2(t_{>d}) \end{aligned} \tag{49}$$

With the requirement that $f_2(t_{>d}) = g_2(t_{>d})$, so that dependence on $t_{>d}$ cancels out in (47). We note that the first line of (49) implies the second so we examine that term in more detail.

$$\begin{aligned}\langle \exp(s_d t_d + s_{>d}\cdot t_{>d} - A(|\vec{s}|))\rangle_{s_{>d}|s_{\le d}} &= \exp(s_d t_d)\int \exp\left(s_{>d}\cdot t_{>d} - A\left(\sqrt{s_{<d}^2+s_d^2+s_{>d}^2}\right)\right) P(s_{>d}|s_{\le d})ds_{>d} \\ &= \frac{\exp(s_d t_d)}{P(s_{\le d})}\int \exp\left(s_{>d}\cdot t_{>d} - A\left(\sqrt{s_{<d}^2+s_d^2+s_{>d}^2}\right)\right) P\left(\sqrt{s_{<d}^2+s_d^2+s_{>d}^2}\right) d\end{aligned}$$

We note that $s_{>d}$ is a $K = D - d$ dimensional vector. We assume that $d < D - 1$, so that $K > 1$, otherwise no further reduction of (50) is possible. We convert the integral over $\mathcal{R}^K$ in (50) into spherical coordinates and break it into three parts: Integration over $\rho \in [0,\infty)$ where $|s_{>d}| = \rho$; integration over $\theta \in [0,\pi]$ where $s_{>d}\cdot t_{>d} = \rho|t_{>d}|\cos(\theta)$, and $\varphi \in \Omega_{K-1}$ is the set of all directions in $\mathcal{R}^K$ with constant θ. The integrand of (50) doesn't depend on φ so we can integrate over it automatically, yielding a constant B_K that is a function only of K. We can now restate (50) in these coordinates:

$$\begin{aligned}\ldots &= \frac{B_K\exp(s_d t_d)}{P(s_{\le d})}\int_0^\infty\int_0^\pi d\rho d\theta \rho^{K-1}\sin^{K-2}(\theta)P\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\exp\left(\rho|t_{>d}|\cos(\theta) - A\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\right) \\ &= \frac{B_K\exp(s_d t_d)}{P(s_{\le d})}\int_0^\infty d\rho\rho^{K-1}P\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\exp\left(-A\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\right)\int_0^\pi d\theta\sin^{K-2}(\theta)\exp\left(\rho|t_{>d}|\cos(\theta)\right)\end{aligned} \tag{51}$$

We next evaluate the integral over θ in (51).

$$\int_0^\pi d\theta\sin^{K-2}(\theta)\exp(\rho|t_{>d}|\cos(\theta)) = \sqrt{\pi}\frac{\Gamma\left(\frac{K}{2}-\frac{1}{2}\right)}{\Gamma\left(\frac{K}{2}\right)}\,{}_0F_1\left(\frac{K}{2},\frac{\rho^2|t_{>d}|^2}{4}\right) \equiv F_K(\rho|t_{>d}|) \tag{52}$$

Where $\Gamma(x)$ is the Gamma function, and ${}_0F_1(a,z)$ is the confluent hypergeometric limit function. We have our final expression for the first term in (49):

$$\begin{aligned}\ldots &= \frac{B_K\exp(s_d t_d)}{P(s_{\le d})}\int_0^\infty d\rho\rho^{K-1}P\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\exp\left(-A\left(\sqrt{s_{<d}^2+s_d^2+\rho^2}\right)\right)F_K(\rho|t_{>d}|) \\ &= \frac{B_K\exp(s_d t_d)}{P(s_{\le d})}g(s_{<d},s_d,|t_{>d}|)\end{aligned} \tag{53}$$

By setting g_1 in (49) equal (53), and letting $g_2 = f_2 = 1$, we have established (48).

B.1 Approximating $A(\vec{s})$ for Gaussian $P(\vec{w})$

In the previous section we assumed that the distribution $P(\vec{w})$ and $P(\alpha)$ are such that $A(\vec{s}) = A(|\vec{s}|)$. In the special case when $N \gg 1$, $P(\alpha) = \delta(\alpha)$, and $\vec{w}$ are Gaussianly distributed we can approximate $A(\vec{s})$ in a semi-closed form. Let $P(\vec{w})$ be a zero-mean Gaussian with with positive-definite covariance matrix C:

$$\begin{aligned} A(\vec{s}) &= N \int d\vec{w} \frac{\exp\left(-\frac{1}{2}\vec{w}^T C^{-1} \vec{w}\right)}{\sqrt{\det(2\pi C)}} \log(2\cosh(\vec{w}\cdot\vec{s})) \\ &= N \int dx \frac{\exp\left(\frac{-x^2}{2\sigma_x^2}\right)}{\sqrt{2\pi\sigma_x^2}} \log(2\cosh(x)) \qquad (54) \\ \sigma_x &= \sqrt{\vec{s}^T C \vec{s}} \qquad (55) \end{aligned}$$

Where we have taken advantage of the fact that $\vec{w}\cdot\vec{s}$ is a scalar gaussian variable with standard deviation that depends on $\vec{s}$ and C. We next take an infinite series expansion of $\log(2\cosh(x))$.

$$\log(2\cosh(x)) = |x| + \log\left(1 + \exp(-2|x|)\right) = |x| + \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \exp(-2m|x|) \qquad (56)$$

As an aside, the first equality in (56) is a useful and numerically stable expression for $A(x)$. The "softplus" function $l(y) = \log(1 + \exp(y))$ is implemented in many scientific computing packages, and using this alternate form for $A(x)$ sidesteps computing the hyperbolic cosine. We next take the appropriate Gaussian average of each term in (56):

$$\begin{aligned} \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} dx \exp\left(\frac{-x^2}{2\sigma_x^2}\right) |x| &= \sqrt{\frac{2}{\pi}}\sigma_x \qquad (57) \\ \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{\infty} dx \exp\left(\frac{-x^2}{2\sigma_x^2} - 2m|x|\right) &= \mathrm{erfcx}(\sqrt{2}m\sigma_x) \qquad (58) \end{aligned}$$

Where $\mathrm{erfcx}(y)$ is the scaled complementary error function. Thus we have our final form for $A(\vec{s})$:

$$A(\vec{s}) = N\sqrt{\frac{2}{\pi}}\sigma_x + N\sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \mathrm{erfcx}(\sqrt{2}m\sigma_x) \qquad (59)$$

We note that $\mathrm{erfcx}(y)$ monotonically decreases to zero, so for large values of $\sqrt{\vec{s}^T C \vec{s}}$, $A(\vec{s})$ is well approximated by the first term in (59). Regardless we see that $A(\vec{s})$ depends on $\vec{s}$ only through $\sqrt{\vec{s}^T C \vec{s}}$:

$$A(\vec{s}) = A\left(\sqrt{\vec{s}^T C \vec{s}}\right) = A\left(|\mathbf{U}\vec{s}|\right) \qquad (60)$$

Where $\mathbf{U} = C^{\frac{1}{2}}$ is the cholesky decomposition of C.

B.2 Generalizing $I_{\text{iso}}(\vec{S}, \vec{T})$ for matched anisotropy

In this section we will show that a generalized form of $I_{\text{iso}}(\vec{S}, \vec{T})$ will also asymptotically converge to $I_{\text{vector}}(\vec{S}, \vec{T})$ when both $P(\vec{s})$ and $A(\vec{s})$ obey a certain form of "matched" anisotropy. Specifically we assume that $P(\vec{s})$ and $A(\vec{s})$ depend on on $\vec{s}$ through a quadratic function of $\vec{s}$ with positive-definite kernel $\mathbf{C}$.

$$P(\vec{s}) = P\left(\sqrt{\vec{s}^T C \vec{s}}\right) = P\left(|\mathbf{U}\vec{s}|\right) \tag{61}$$

$$A(\vec{s}) = A\left(\sqrt{\vec{s}^T C \vec{s}}\right) = A\left(|\mathbf{U}\vec{s}|\right) \tag{62}$$

Where, as in Section B.1, $\mathbf{U} = C^{\frac{1}{2}}$ is the cholesky decomposition of C. Let us transformed versions of $\vec{S}$ and $\vec{T}$:

$$\tilde{S} = \mathbf{U}\vec{S} \tag{63}$$

$$\tilde{T} = \mathbf{U}^{-T}\vec{T} \tag{64}$$

Where $\mathbf{U}^{-T}$ is the transpose of the inverse of $\mathbf{U}$, which is well defined since C was positive definite. We note that $\tilde{t} \cdot \tilde{s} = \vec{t} \cdot \vec{s}$ and $P(\tilde{t}|\tilde{s})$ can once again be written as an exponential family in canonical form:

$$P(\tilde{t}|\tilde{s}) = \exp(\tilde{s} \cdot \tilde{t} - A(\tilde{s}))h(\tilde{t}) \tag{65}$$

As the mappings from $\vec{S}$ to $\tilde{S}$ and $\vec{T}$ to $\tilde{T}$ are diffeomorphisms, we have that $I_{\text{vector}}(\vec{S}, \vec{T}) = I_{\text{vector}}(\tilde{S}, \tilde{T})$. Furthermore, equation (65) implies that an analogous form of equation (26) holds for $I_{\text{vector}}(\tilde{S}, \tilde{T})$:

$$I(\tilde{S}_d, \tilde{T}|\tilde{S}_{<d}) = I(\tilde{S}_d, \tilde{T}_{\geq d}|\tilde{S}_{<d}) \tag{66}$$

Additionally, $A(\tilde{s}) = A(|\tilde{s}|)$ and $P(\tilde{s}) = P(|\tilde{s}|)$. Thus, we may reuse the derivation of section 48 to derive the following analogy of equation (48):

$$I(\tilde{S}_d, \tilde{T}_{\geq d}|\tilde{s}_{<d}) \approx I\left(\tilde{S}_d, \left\{\tilde{T}_d, |\tilde{T}_{>d}|\right\} \middle| \tilde{s}_{<d}\right) \tag{67}$$

Therefore $I_{\text{iso}}(\tilde{S}, \tilde{T})$ is asymptotically equal to $I_{\text{vector}}(\vec{S}, \vec{T})$:

$$I_{\text{iso}}(\tilde{S}, \tilde{T}) = \sum_d I\left(\tilde{S}_d, \left\{\tilde{T}_d, |\tilde{T}_{>d}|\right\} \middle| \tilde{S}_{<d}\right) \approx I_{\text{vector}}(\vec{S}, \vec{T}) \tag{68}$$

We note that an example of such a matched isotropy situation would be where both the stimuli and receptive fields (for a large population) are distributed according to a Gaussian distribution with covariance matrix C (c.f. section B.1).

C Independent Sub-populations

In this section we present an example where one of our proposed approximations, $I_{\text{comp-ind}}(\vec{S},\vec{T}\,)$ in this case, is equal to $I_{\text{vector}}(\vec{S},\vec{T}\,)$. Let $(\hat{e}'_1,...,\hat{e}'_D)$ be an orthonormal basis for $\mathcal{R}^D$. Suppose that the distribution of $\vec{s}$ and $\vec{w}$ are such that both $P(\vec{s}\,)$ and $A(\vec{s}\,)$ factor when expressed in this basis ($s'_k = \vec{s}\cdot\hat{e}'_k$):

Similarly defining $t'_k = \vec{t}\cdot\hat{e}'_k$ we have that $\vec{t}\cdot\vec{s} = \vec{t}'\cdot\vec{s}'$. Because mutual information is invariant under bijective transformations of the variables (e.g. a change of basis) (Cover and Thomas, 2012) we have that $I_{\text{vector}}(\vec{S},\vec{T}\,) = I(\vec{S}',\vec{T}')$. It is easy to show that $P(\vec{t}'|\vec{s}')$ can be written as follows:

$$P(\vec{t}'|\vec{s}') = h(\vec{t}')\prod_k \exp(s'_k t'_k - A(s'_k)) \tag{69}$$

Eq (69) implies that the log-likelihood ratio of $P(\vec{t}'|\vec{s}')$ to $P(\vec{t}')$ decomposes across s'_k:

$$\log\left(\frac{P(\vec{t}'|\vec{s}')}{P(\vec{t}')}\right) = \sum_k \log\left(\frac{P(t'_k|s'_k)}{P(t'_k)}\right) \tag{70}$$

Thus we have the following reduction of $I_{\text{vector}}(\vec{S},\vec{T}\,)$:

$$I_{\text{vector}}(\vec{S},\vec{T}\,) = I_{\text{vector}}(\vec{S}',\vec{T}') = \sum_k I(S'_k,T'_k) = I_{\text{comp-ind}}(\vec{S}',\vec{T}') \tag{71}$$

We note that (31) includes the case where for some k, $\vec{w}_n\cdot\hat{e}'_k = 0\,\forall n$. In such a case $A(s'_k) = A(0) = \log(2)$, $t'_k = 0$ with probability one, and $I(S'_k,T'_k) = 0$. Thus, in the case of independent subpopulations, $I_{\text{vector}}(\vec{S},\vec{T}\,)$ can be reduced to computing $I_{\text{comp-ind}}(\vec{S},\vec{T}\,)$ following a change of basis.

D Relationship between $I_{\text{k-w}}(\vec{R},\vec{S}\,)$ and $I_{\text{Fisher}}(\vec{R},\vec{S}\,)$

In this appendix we relate $I_{\text{k-w}}(\vec{R},\vec{S}\,)$ to the Fisher Information based approximation of (Brunel and Nadal, 1998):

$$I_{\text{Fisher}}(\vec{R},\vec{S}\,) = H(\vec{S}\,) + \frac{1}{2}\int d\vec{s}P(\vec{s}\,)\log\left(\frac{|\mathbf{J}(\vec{s}\,)|}{(2\pi e)^D}\right) \tag{72}$$

Where $\mathbf{J}(\vec{s})$ is the Fisher Information Matrix of $P(\vec{r}|\vec{s})$:

$$\mathbf{J}_{ab}(\vec{s}) = \left\langle \frac{\partial^2}{\partial_a \partial_b} \log P(\vec{r}|\vec{s}) \right\rangle_{P(\vec{r}|\vec{s})} \tag{73}$$

We begin by considering the inner expectation over $\vec{s}'$ in $I_{\text{k-w}}(\vec{R}, \vec{S})$:

$$L(\vec{s}) = \int d\vec{s}' P(\vec{s}') \exp\left(-D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}')\right) \tag{74}$$

We next assume that the activation function $f_n(\vec{s})$ is affine (e.g. (13)), and thus in canonical form. We also assume that, $P(\vec{r}|\vec{s})$ is identifiable:

$$D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}') = 0 \Leftrightarrow \vec{s} = \vec{s}' \tag{75}$$

For (13) a necessary and sufficient condition for identifiability is that the matrix $\mathbf{W}$ has full rank, a reasonable assumption when $N \gg D$. We utilize the following properties of exponential families in canonical form:

1. $\frac{\partial^2}{\partial_a' \partial_b'} D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}') = \frac{\partial^2}{\partial_a' \partial_b'} A(\vec{s}') = \mathbf{J}(\vec{s}')$.
2. $A(\vec{s}')$ is convex and $\mathbf{J}(\vec{s}')$ is positive semi-definite. When $P(\vec{r}|\vec{s})$ is identifiable replace $A(\vec{s}')$ becomes strictly convex, $\mathbf{J}(\vec{s}')$ positive definite, and $D_{KL}(P(\vec{r}|\vec{s})||P(\vec{r}|\vec{s}')$ has a global minimum with respect to $\vec{s}'$ of 0 at $\vec{s}' = \vec{s}$.

In the limit $N \gg D$ we approximate $L(\vec{s})$ using Laplace's Method (Bender and Orszag, 1999), expanding around $\vec{s}' = \vec{s}$:

$$L(\vec{s}) \approx P(\vec{s}) \sqrt{\frac{(2\pi)^D}{|\mathbf{J}(\vec{s})|}} \tag{76}$$

Plugging (76) into the definition of $I_{\text{k-w}}(\vec{R}, \vec{S})$ we have the following asymptotic expression for $I_{\text{k-w}}(\vec{R}, \vec{S})$:

$$\begin{aligned} I_{\text{k-w}}(\vec{R}, \vec{S}) &\approx H(\vec{s}) + \frac{1}{2} \int d\vec{s} P(\vec{s}) \log\left(\frac{|\mathbf{J}(\vec{s})|}{(2\pi)^D}\right) \\ &= I_{\text{Fisher}}(\vec{R}, \vec{S}) - \frac{D}{2} \end{aligned} \tag{77}$$

For stimulus distributions where the entropy $H(\vec{S})$ is known *a priori*, such as the Gaussian distributions in sections 5.1 and 5.2, $I_{\text{Fisher}}(\vec{R}, \vec{S})$ can be computed in $O(M)$ time. If not, then $H(\vec{S})$ must be estimated, a challenging task in high dimensions. In figure 7, we replot the results of sections 5.1 with the inclusion of $I_{\text{Fisher}}(\vec{R}, \vec{S})$. We see that $I_{\text{Fisher}}(\vec{R}, \vec{S})$ is a very loose upper bound of $I(\vec{R}, \vec{S})$ and of $I_{\text{k-w}}(\vec{R}, \vec{S})$, indicating

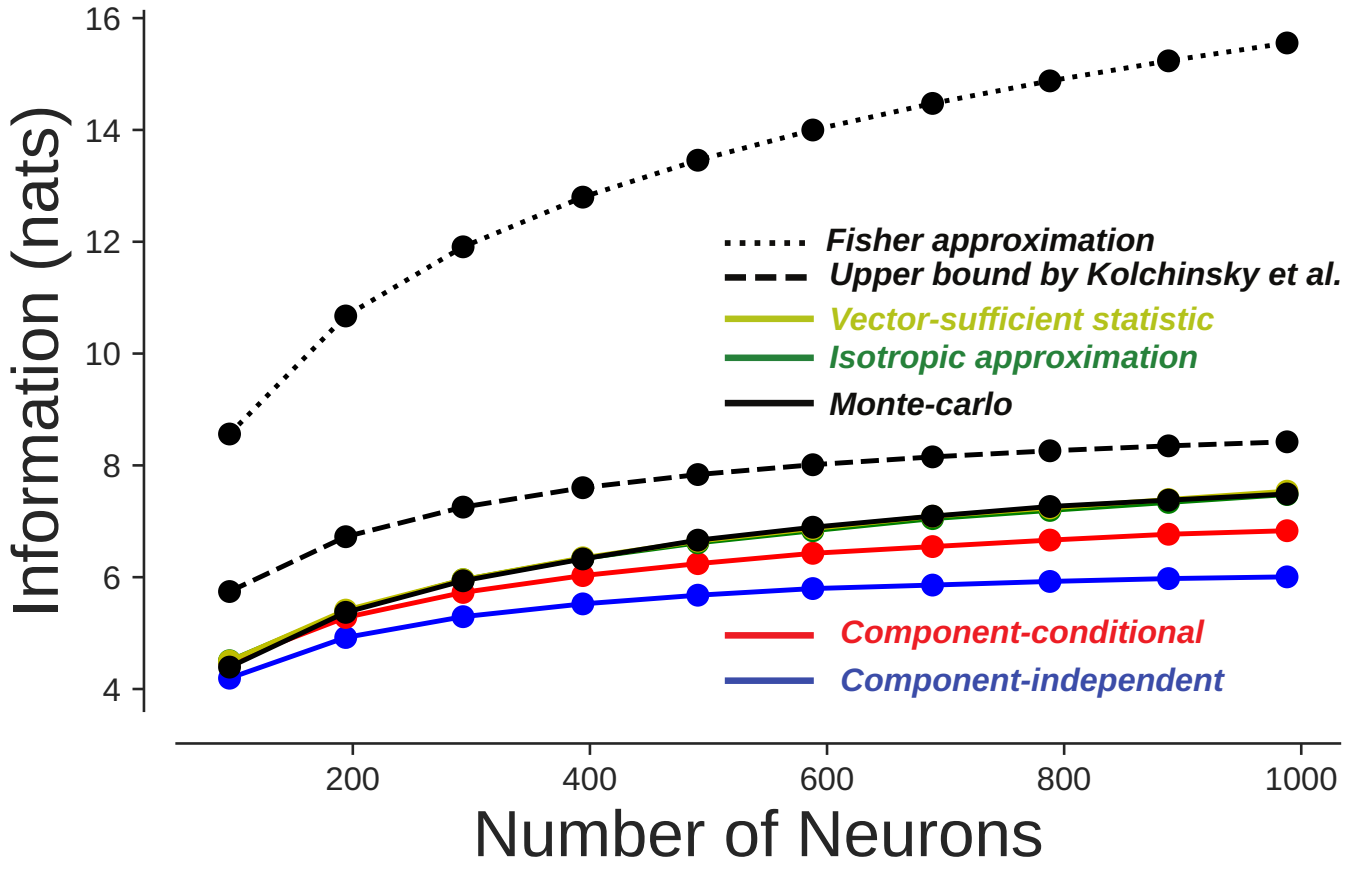

Figure 7: Information curves for the population of section 5.1 compared to Fisher approximation. Lines and errorbars are mean and standard deviation over ten repeats of the estimator. $\hat{I}(\vec{R}, \vec{S}\,)$ (solid black line), $I_{\text{k-w}}(\vec{R}, \vec{S}\,)$ (dashed black line), $I_{\text{vector}}(\vec{S}, \vec{T}\,)$ (solid yellow line), $I_{\text{Fisher}}(\vec{R}, \vec{S}\,)$ (dotted black line), $I_{\text{comp-cond}}(\vec{S}, \vec{T}\,)$ (solid red line), $I_{\text{comp-ind}}(\vec{S}, \vec{T}\,)$ (solid blue line), $I_{\text{iso}}(\vec{S}, \vec{T}\,)$ (solid green line)

that the convergence of Laplace's Method may be very slow in this situation.

E Extension to polynomial activation functions

In section 4.1 we assumed that the activation functions were affine functions of the stimulus vector $\vec{s}$. In this appendix we will show how to generalize some of the results of section 4.1 to polynomial activation functions. For clarity of exposition we will demonstrate this generalization for quadratic functions as the procedure for higher order polynomials follows quickly. To begin, we add a quadratic term to Eq (12):

$$f_n(\vec{s}) = \vec{s}^{\,T} \gamma_n \vec{s} + \vec{w}_n \cdot \vec{s} - \alpha_n \tag{78}$$

$\gamma_n \in \mathcal{R}^{D\times D}$ is a symmetric $D \times D$ matrix representing the quadratic kernel of the n^{th} neuron's activation function. We note that $\vec{s}^T\gamma_n\vec{s} = \vec{s}\,\vec{s}^T \circ \gamma_n$ where $A \circ B$ is the hadamard product between equally shaped matrices A and B and $\vec{s}\,\vec{s}^T \in \mathcal{R}^{D\times D}$ is the outer product of $\vec{s}$ with itself. We define a vector embedding of $\vec{s}$ into $\mathcal{R}^{D+D^2}$, $\vec{\psi}(\vec{s})$:

$$\psi(\vec{s})_d = \begin{cases} s_d & \text{if } d \leq D \\ s_a * s_b & \text{if } d > D \end{cases}$$

Where $a = d \mod D$ and $b = \lfloor d/D \rfloor$ are index mappings that map a $D \times D$ matrix into a vector of length D^2. We define a similar vector embedding of $\vec{w}_n$ and γ_n, $\vec{\tau}_n(\vec{w}_n, \gamma_n)$:

$$\tau_n(\vec{w}_n, \gamma_n)_d = \begin{cases} w_{n,d} & \text{if } d \leq D \\ \gamma_{n,ab} & \text{if } d > D \end{cases}$$

For the sake of brevity we henceforth omit the dependence of $\vec{\tau}_n$ on $\vec{w}_n$ and γ_n. By construction we have the following equivalence:

$$\vec{s}^T\gamma_n\vec{s} + \vec{w}_n \cdot \vec{s} = \vec{\tau}_n \cdot \vec{\psi}(\vec{s}) \tag{79}$$

If all neurons have activation functions of the form in (78) then $P(\vec{r}|\vec{s})$ may once again be written as an exponential family

$$\begin{aligned} P(\vec{r}|\vec{s}) &= h(\vec{r}) \exp(\vec{t}^{\,\text{quad}}(\vec{r}) \cdot \vec{\psi}(\vec{s}) - A(\vec{s})) \\ A(\vec{s}) &= \sum_n \log(2\cosh(\vec{\tau}_n \cdot \vec{\psi}(\vec{s}) - \alpha_n)) \\ \vec{t}^{\,\text{quad}}(\vec{r}) &= \sum_n \vec{\tau}_n r_n \end{aligned} \tag{80}$$

As $\vec{t}^{\,\text{quad}}(\vec{r}) \in \mathcal{R}^{D+D^2}$ is the sufficient statistic for this family, and $\vec{\psi}(\vec{s})$ is the natural parameter. As before, $I(\vec{R}, \vec{S}) = I(\vec{T}^{\,\text{quad}}, \vec{S})$. However, we note that $P(\vec{r}\,|\vec{S})$ can be written entirely in terms of $\vec{\psi}$. Additionally, we note that the support of $\vec{\psi}$ lies on a D-dimensional manifold in $\mathcal{R}^{D+D^2}$ and $\vec{s}$ maps injectively into this manifold. Thus $I(\vec{T}^{\,\text{quad}}, \vec{S}) = I(\vec{T}^{\,\text{quad}}, \vec{\Psi})$.

We note several properties of $I(\vec{T}^{\,\text{quad}}, \vec{\Psi})$. First, we can in principle expand $I(\vec{T}^{\,\text{quad}}, \vec{\Psi})$ like Eq. (19):

$$I(\vec{T}^{\,\text{quad}}, \vec{\Psi}) = \sum_{d=1}^{d=D+D^2} I(\vec{T}^{\,\text{quad}}, \Psi_d|\Psi_{<d}) \tag{81}$$

Secondly, the same reduction as Eq. (26) holds for $I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d})$:

$$I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d}) = I(T^{\mathrm{quad}}_{\geq d}, \Psi_d|\Psi_{<d}) \tag{82}$$

Most notably however, is that $I(\vec{T}^{\,\mathrm{quad}}, \Psi_d|\Psi_{<d}) = I(T^{\mathrm{quad}}_{\geq d}, \Psi_d|\Psi_{<d}) = 0$ for $d > D$. This holds because $\psi_d = g(\psi_{\leq D})$ for $d > D$, where $g(\psi_{\leq D})$ is just the product of the two relevant components of $\psi_{\leq D}$. Because of this functional dependence we can just apply the generalization of Eq. (22). Therefore the expansion of $I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi}\,)$ can be truncated after D terms.

$$I(\vec{T}^{\,\mathrm{quad}}, \vec{\Psi}\,) = \sum_{d=1}^{d=D} I(T^2_{\geq d}, \Psi_d|\Psi_{<d}) = \sum_{d=1}^{d=D} I(T^{\mathrm{quad}}_{\geq d}, S_d|S_{<d}) \tag{83}$$

In fact, we can make an even stronger reduction by noting that conditioning on components of $\vec{S} = \psi_{\leq D}$ effectively also conditions on elements of $\psi_{>D}$. For clarity of exposition we break down $\vec{T}^{\,\mathrm{quad}}$ into vector and matrix valued components:

$$\begin{aligned}
\vec{T}^{\,\mathrm{quad}} &\equiv \left\{T_{\leq D}, T^{\leq D}_{\leq D}\right\} \\
T^{\mathrm{quad}}_{\leq D} &\rightarrow T_{\leq D} \in \mathcal{R}^D \\
T^{\mathrm{quad}}_{>D} &\rightarrow T^{\leq D}_{\leq D} \in \mathcal{R}^{D\times D}
\end{aligned}$$

We note that conditioning on S_d conditions on the components of $\vec{\psi}$ corresponding to T_d and T^d_d. Additionally conditioning on $S_{<d}$ conditions on the components of $T_{<d}$ *and* on the components of $T^{d_2}_{d_1}$ for all indices d_1 and d_2 such that $1 \leq d_1, d_2 < d$. Thus Eq. (83) can be further generalized:

$$I(\vec{T}^{\,\mathrm{quad}}, \vec{S}\,) = \sum_{d=1}^{d=D} I\left(\left\{T_{\geq d}, T^{\geq d}_{\geq d}\right\}, S_d \middle| S_{<d}\right) \tag{84}$$

The d^{th} term in Eq (83) has D^2+D+1 degrees of freedom while the d^{th} term in Eq (84) has $D^2+D+1-(d-1)^2$ degrees of freedom. The above procedure can be generalized to polynomial activation functions of arbitrarily high but finite order, though the dimensionality of the sufficient statistic and natural parameter grow exponentially with the order. However, Eq. (84) holds for any order of polynomial, so that one one needs only compute the first D terms of the expansion of the mutual information between the sufficient statistic and natural parameter.