Inferring Mutual Information from Classification Accuracy

Charles Zheng and Yuval Benjamini

November 4, 2016

Abstract

Multivariate pattern analyses approaches in neuroimaging are fundamentally concerned with investigating the quantity and type of information processed by various regions of the human brain; typically, estimates of classification accuracy are used to quantify information. While a extensive and powerful library of methods can be applied to train and assess classifiers, it is not always clear how to use the resulting measures of classification performance to draw scientific conclusions: e.g. for the purpose of locating locally informative areas in the brain. An additional confound for interpreting classification accuracy is the dependence of the error rate on the number and choice of distinct classes obtained for the classification task. In contrast, mutual information is a quantity which is in principle defined independently of the experimental design, and has ideal properties for comparative analyses. One obstacle to wider use of mutual information is the difficulty of obtaining either estimates or lower confidence bounds for mutual information in high-dimensional neuroimaging data. We propose a new lower cofidence bound for mutual information based on the concept of "average Bayes error." Since the error of a classifier can be used to infer an upper bound on the average Bayes error, we develop a lower confidence bounds for mutual information based on a new theoretical lower bound on mutual information as a function of average Bayes error. We demonstrate the utility of our approach in simulated and real data examples.

1 Introduction

A fundamental challenge of computational neuroscience is to understand how information about the external world is processed and represented in the brain. Each individual neuron aggregates the incoming information into a single sequence of spikes–an output which is too simplistic by itself to capture the full complexity of sensory input. Only by combining the signals from massive ensembles of neurons is it possible to reconstruct our complex representation of the world. Nevertheless, neurons form hierarchies of specialization within neural circuits, which are further organized in various specialized regions of the brain. At the lowest level of the hierarchy–individual neurons, it is possible to infer and interpret the functional relationship between a neuron and stimulus features of interest using single-cell recording technologies. Due to the inherent stochasticity of the neural output, it is natural to view the neuron as a noisy channel, and use mutual information to quantify how much of the stimulus information is encoded by the neuron. Moving up the hierarchy to the the macroscale level of organization in the brain requires both different experimental methodologies and new approaches for summarizing and inferring measures of information in the brain.

Shannon's mutual information $I(X;Y)$ is fundamentally a measure of dependence between random variables X and Y, and is defined as

$$I(X;Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy.$$

Various properties of $I(X;Y)$ make it ideal for quantifying the information between a random stimulus X and the signaling behavior of an ensembles of neurons, Y (Borst 1999). A leading metaphor is that of a noisy communications channel; the mutual information describes the rate at which Y can communicate bits from X. This framework is well-suited for summarizing the properties of a single neuron coding external stimulus information; indeed, experiments studying the properties of a single or a small number of neurons often make use of the concept of mutual information in summarizing or interpreting their results (Quiroga 2009). However, estimating mutual information for multiple channels require large and over-parameterized generative models. For instance, one can tractably estimate mutual information by assuming a multivariate Gaussian model: however, this approach essentially assumes a linear relationship between the input and output, and hence fails to quantify nonlinear dependencies. As the complexity of stimuli and

the number of output channels increases, these models are hard or impossible to estimate without gross over-fitting. As new technologies for simultaneous measurement of multiple brain regions developed, such as functional MRI, it became increasingly difficult to quantify information at such scales under the classical approach.

Machine learning algorithms showed a way forward: a seminal work by Haxby (2001) proposed to quantify the information in multiple channels by measuring how well the stimulus can be identified from the brain responses, in what is known as "multivariate pattern analysis" (MVPA). To demonstrate that a particular brain region responds to a certain type of sensory information, one employs supervised learning to build a classifier that classifies the stimulus class from the brain activation in that region. Classifiers that achieve above-chance classification accuracy indicate that information from the stimulus is represented in the brain region. In principle, one could just as well test the statistical hypothesis that the Fisher information or mutual information between the stimulus and the activation patterns is nonzero. But in practice, the machine learning approach enjoys several advantages: First, it is invariant to the parametric representation of the stimulus space, and is opportunistic in the parameterization of the response space. This is an important quality for naturalistic stimulus-spaces, such as faces or natural images. Second, it scales better with the dimensionality of both the stimulus space and the responses space, because a slimmer discriminative model can be used rather than a fully generative model.

Nevertheless, classification accuracy is problematic for quantifying the strength of the relation between stimulus and outputs due to its arbitrary scale and strong dependence on experimental choices. Classification accuracy depends on the particular choice of stimuli exemplars employed in the study and the number of partitions used to define the classes for the classification task. The difficulty of the classification task depends on the number of classes defined: high classification accuracy can be achieved relatively easily by using a coarse partition of stimuli exemplars into classes. In a meta-analysis on visual decoding, Coutanche et al (2016) quantified the strength of a classification study using the formula

$$\text{decoding strength} = \frac{\text{accuracy} - \text{chance}}{\text{chance}}.$$

Such an approach may compensate for the differences in accuracy due purely to choice of number of classes defined; however, no theory is provided to

justify the formula. In contrast, mutual information has ideal properties for quantitatively comparing information between different studies, or between different brain regions, subjects, featurization models, or modalities. Not only is the mutual information defined independently of the arbitrary definition of stimulus classes (albeit still dependent on an implied distribution over stimuli), it is even meaningful to discuss the difference between the mutual information measured for one system and the mutual information for a second system.

Hence, a popular approach which combines the strengths of the machine learning approach and the advantages of the information theoretic approach is to obtain an probabilistic lower bound on the mutual information by using the confusion matrix of a classifier. This is the most popular approach for estimating mutual information in neuroimaging studies, but suffers from known shortcomings (Gastpar 2010, Quiroga 2009). The idea of linking classification performance to mutual information dates back to the beginnings of information theory: Shannon's original motivation was to characterize the minimum achievable error probability of a noisy communication channel. More explicitly, Fano's inequality provides a lower bound on mutual information in relation to the optimal prediction error, or Bayes error. Fano's inequality can be further refined to obtain a tighter lower bound on mutual information (Tebbe and Dwyer 1968.) However, the statistical properties of these procedures have not been analyzed, and no coverage guarantees can be stated for the obtained.

In this paper, we propose a new method for constructing a *lower confidence bound* on mutual information based on classification performance. That is, we can construct a data-dependent probabilistic lower bound $\underline{I}$ such that the probability of type I error is controlled,

$$\Pr[\underline{I} > I(X;Y)] \leq \alpha$$

for user-specified $\alpha \in [0, 1]$–and that this coverage guarantee holds regardless of the underlying distribution. To accomplish we need to overcome the arbitrary choice of exemplars, and the arbitrary number of classes K. Towards this end, we define a notion of k-class *average Bayes accuracy* which is uniquely defined for any given stimulus distribution and stochastic mapping from stimulus to response. The k-class average Bayes error is the expectation of the Bayes error (the classification error of the optimal classifier) when k stimuli exemplars are drawn i.i.d. from the stimulus distribution,

and treated as distinct classes. Hence the average Bayes error can in principle be estimated if the appropriate randomization is employed for designing the experiment.

Our main theoretical contributions are (i) the derivation of a variance bound on the k-class Bayes accuracy, and (ii) the derivation of a tight lower bound on mutual information as a function of k-class average Bayes error. The variance bound (i) is needed in order for be able to construct a lower confidence bound on k-class average Bayes accuracy based on data; the lower bound (ii) is then employed to convert a lower confidence bound for k-class average Bayes accuracy into a lower bound for mutual information. Although the k-class average Bayes accuracy is defined independently of the particular choice of stimuli, the quantity still depends on the choice of number of classes, k. Mutual information provides a quantification of information that does not depend on k, allowing more flexible comparisons and easier interpretation.

2 Framework

2.1 Notation

Functionals of distributions, such as expectation $\mathbf{E}[\cdot]$ or mutual information $\mathrm{I}[\cdot]$, can be written as functionals of density functions:

$$\mathbf{E}[g] \stackrel{D}{=} \mathbf{E}[g(x)] \stackrel{D}{=} \int g(x)dx$$

$$\mathrm{H}[g] \stackrel{D}{=} \int g(x) \log g(x)dx$$

$$\mathrm{I}[p] \stackrel{D}{=} \mathrm{I}[p(x,y)] = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)}dxdy$$

or as functionals of random variables: $\mathbf{E}(X)$, $\mathrm{H}(X)$, $\mathrm{I}(X,Y)$. Square brackets are used when the arguments are density functions, and parentheses are used when the arguments are functions of random variables. $|\cdot|$ denotes set cardinality.

2.2 Model

The following simplified model captures the essence of many neuroimaging studies. Let $\mathcal{X}$ be a space of stimuli, represented by q-dimensional vec-

tors. In the design stage of the experiment, a set of k stimuli exemplars $\{x^{(1)}, \ldots, x^{(k)}\}$ are selected, and assigned into a T-length sequence of stimuli $(x^{(i_1)}, \ldots, x^{(i_T)})$ to be presented to the subject. In the execution of the experiment, an activation pattern or *response* y^t is obtained for each of the stimuli presentations $x^{(i_t)}$. Generally y^t is a p-dimensional vector, representing activity levels of p disjoint brain regions. To simplify, assume that each of the k stimuli is presented a total of r times in the sequence; further assume that the responses to each stimulus presentation are conditionally independent, hence the ordering of the sequence does not matter. Henceforth we can let $y^{(i),j}$ denote the response to the jth presentation of the stimulus $x^{(i)}$.

While in many studies, the stimuli exemplars $x^{(1)}, \ldots, x^{(k)}$ are chosen somewhat arbitrarily, more types of inferences can be drawn if randomization is employed to choose the stimuli. Based on this distinction, we discuss two types of experimental designs:

1. *Fixed stimuli design.* The exemplars $x^{(1)}, \ldots, x^{(k)}$ are treated as fixed.

2. *Randomized stimuli design.* The experimenter first specifies a space of stimuli $\mathcal{X}$ and a probability distribution $p(x)$ over that space. Then, k exemplars are drawn i.i.d. from that space,

$$X^{(1)}, \ldots, X^{(k)} \overset{iid}{\sim} p(x).$$

As we will discuss, the randomized stimuli design allows the possibility of *generalizing* beyond the chosen exemplars; i.e. drawing inferences about the entire joint distribution $p(x, y)$.

And while we draw a clear-cut distinction between the fixed design setting and randomized design setting, in practice, the boundaries are frequently blurred. Indeed, few researchers in the field actually employ an explicit randomization mechanism to choose the stimuli, yet they may think that their stimuli are "effectively random". For stimulus categories such as faces or natural images, it is not clear how to specify a probability distribution $p(x)$ in the first place. Yet, a practical approach is to choose stimuli in an *ad-hoc* manner with the intent of obtaining a representative set, and to treat the stimuli as having been randomly sampled from some implicitly defined $p(x)$ *after the fact.* In such studies, the convenient fiction of a randomized design is employed to formalize the process of generalizing experimental results. Such a cavalier attitude towards randomization may seem concerning by

the standards of clinical trials, but it is appropriate for the goals of typical imaging studies, where the ultimate objective to discover some qualitative properties of human perception and cognition.

2.3 Fixed stimuli design

We begin by describing a fixed stimuli design, and the types of inferences that can be made using machine learning tools. The general approach we describe was first introduced to the neuroimaging community by Haxby (1999), but its foundational concepts belong to the area of statistical learning (Hastie et al. 2008).

Let $\mathcal{X} \subset \mathbb{R}^q$ and $\mathcal{Y} \subset \mathbb{R}^p$; For every $x \in \mathcal{X}$, let $p_x(y)$ be a probability density on $\mathcal{Y}$. Take a subset $\{x^{(1)}, \ldots, x^{(k)}\} \subset \mathcal{X}$. Let Y_i^j be a random vector distributed according to density $p_{x^{(i)}}(y)$; and let $Y^{(i),j}$ be independent for $i = 1, \ldots, k$ and $j = 1, \ldots, r$.

In MVPA, one carries out a classification task to assess whether y contains information about x. Formally, a classification rule is any (possibly stochastic) mapping $f : \mathcal{Y} \to \{1, \ldots, k\}$. The *generalization accuracy* of the classification rule for classes $x^{(1)}, \ldots, x^{(k)}$ is

$$\mathrm{GA}(f) = \frac{1}{k} \sum_{i=1}^{k} \Pr[f(Y) = i | X = x^{(i)}].$$

A trivial classification rule which outputs the result of a k-sided die roll for all inputs y would achieve a generalization accuracy of $\mathrm{GA} = \frac{1}{k}$. Conversely, even a single counterexample with $\mathrm{GA} > \frac{1}{k}$ is indicative that y contains nonzero information about x. Hence, in order to demonstrate that y is informative of x, one tests the null hypothesis

$$H_0 : \mathrm{GA}(f) = \frac{1}{k}$$

versus the alternative

$$H_1 : \mathrm{GA}(f) > \frac{1}{k}.$$

Rejecting the null hypothesis for a given classification rule f can be taken as evidence that y is informative of x.

We have not yet specified how any classification rule f is to be obtained. Unless one has strong prior knowledge about the nature of the brain encoding,

it is necessary to choose the function f in a data-dependent way in order to obtain a reasonable classification rule. A wide variety of machine learning algorithms exist for "learning" good classification rules f from data. We use the terminology *classifier* to refer to any algorithm which takes data as input, and produces a classification rule f as output. The following discussion makes it necessary for us to make a precise distinction between the *classifier* and the *classification rule* it produces, and our usage of the terms may differ from the standard in the literature. Mathematically speaking, the classifier is a functional which maps a set of observations to a classification rule,

$$\mathcal{F} : \{(x^1, y^1), \ldots, (x^m, y^m)\} \mapsto f(\cdot).$$

The data $(x^1, y^1), \ldots, (x^m, y^m)$ used to obtain the classification rule is called *training data.* When the objective is to obtain the best possible classification rule, as is the case in diagnostic settings, it is optimal to use all of the availible data to train the classifier. However, when the goal is to obtain *inference* about the performance of the classification rule, it beomes necessary to split the data into two independent sets: one set to train the classifier, and one to evaluate the performance. The reason that such a splitting is necessary is because using the same data to test and train a classifier introduces significant bias into the empirical classification error.

In *data-splitting*, one creates a *training set* consisting of r_1 repeats per class,

$$\{(x^{(1)}, y^{(1),1}), \ldots, (x^1, y^{(1),r_1}), \ldots, (x^{(k)}, y^{(k),1}), \ldots, (x^{(m)}, y^{(m),r_1})\}$$

and a *test set* consisting of the remaining $r_2 = r - r_1$ repeats.

$$\{(x^{(1)}, y^{(1),r_1+1}), \ldots, (x^{(1)}, y^{(1),r}), \ldots, (x^{(k)}, y^{(k),r_1+1}), \ldots, (x^{(k)}, y^{(k),r_1})\}.$$

One inputs the training data into the classifier to obtain the classification rule f,

$$f = \mathcal{F}(\{(x^{(1)}, y^{(1),1}), \ldots, (x^{(1)}, y^{(1),r_1}), \ldots, (x^{(k)}, y^{(k),1}, \ldots, (x^{(k)}, y^{(k),r_1})\}).$$

The test statistic of interest is the test error, defined as

$$\widehat{\mathrm{GA}} = \frac{1}{kr_2} \sum_{i=1}^{k} \sum_{j=r_1+1}^{r} \mathrm{I}(f(y^{(i),j}) \neq i).$$

Due to the conditional independence of the training set and test set, $\widehat{\mathrm{GA}}$ is an unbiased estimate of GA. Hence various approaches can be used to obtain a threshold c_α such that $\Pr[\widehat{\mathrm{GA}} > c_\alpha] \leq \alpha$ holds (approximately) under the null hypothesis. These approaches include permutation tests, tests based on a universal variance bound, or the generalized likelihood ratio test. The hypothesis H_0 is then rejected at level α if $\widehat{\mathrm{GA}} > c_\alpha$.

There is also a way to improve on data-splitting, by using *cross-validation*. Essentially, one applies the data-splitting approach multiple times, with different training and test partitions, and aggregates the results. Cross-validation can be used to reduce the variance of the estimated GA and hence improve the power of the test, and it can also be used to obtain *confidence intervals* for the estimate. We discuss such a cross-validation in greater detail in section 5.

While tests of the generalization accuracy suffice to establish the presence of information, the generalization accuracy is less satisfactory as a measure of the information between X and Y, because GA depends on the classification rule f obtained–but since the performance of the classifier may vary depending on the choice of model (k-nearest neighbors, SVM, etc.) and the choice of tuning parameters, the quantity GA is therefore not uniquely defined. To resolve the ill-definedness of the generalization accuracy, we define the Bayes accuracy, which is simply the *optimal* generalization accuracy

$$\mathrm{BA} = \sup_f \mathrm{GA}(f).$$

Due to Bayes' theorem, the optimal classification rule f^* which achieves the Bayes error can be given explicitly: it is the maximum a posteriori (MAP) rule

$$f^*(y) = \mathrm{argmax}_{i=1}^k \; p(y|x^{(i)}).$$

Of course, it is not possible to construct this rule in practice since the joint distribution is unknown. Instead, a reasonable approach is to try a variety of classifiers, producing rules $f_1, \ldots, f_m$, and taking the best generalization accuracy as an estimate of the Bayes accuracy. We give more details of this approach in the Discussion.

This way one can draw inferences about information between the fixed ensemble $x_1, ..., x_k$ and the brain regions being examined. But in many cases it is desired to generalize the results to other stimuli that were not included in the experiment. In order to do this, we need to be able to assume a randomized design.

2.4 Randomized stimuli design

Under the assumption of randomized design, we treat the stimuli $x^{(1)}, \ldots, x^{(k)}$ as independent draws from some distribution $p(x)$. Defining the joint distribution of the stimulus X and the response Y as

$$p(x, y) = p(x)p(y|x),$$

one can also define the mutual information

$$I(X;Y) = \int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxy.$$

In contrast the case of fixed design, the randomized design framework provides a principled way of making inferences about the population of stimuli exemplars, beyond the particular exemplars that were chosen for the study. This is particularly relevant for complex stimuli, for which the number of distinct stimuli species (e.g. distinct faces) could be astronomically large (i.e. the number of faces that a human could distinguish.) The concept of information $I(X;Y)$ captures the notion of the "complexity" of the stimuli representation. By inferring the information $I(X;Y)$, we make inferences about the complexity of the representation in the given brain region Y.

Methodologies for estimating mutual information under the assumption of randomization have been studied in (Borst 1999), (Nelken 2005) and most notably (Gastpar 2010); the latter studies the problem under an identical setup to ours. In addition to these approaches, one can obtain a lower bound on the information using the mutual information of the confusion matrix (Treves et al), (Quiroga 2009). In Section 3, we present our methodology for inferring mutual information under the randomized stimuli design setting.

Before presenting our approach, we first define the notion of average Bayes accuracy, which is fundamental to our approach. The motivation for defining average Bayes accuracy is the fact that the quantity BA is not a parameter of the joint distribution $p(x, y)$, but rather depends on the specific exemplars $x_1, \ldots, x_k$ selected; hence, one may write $\mathrm{BA}(x_1, \ldots, x_k)$ to emphasize this dependence. Explicitly,

$$\mathrm{BA}(x^{(1)}, \ldots, x^{(k)}) = \sup_f \sum_{z=1}^{k} \frac{1}{k} \int_y I\{f(x^{(1)}, ..., x^{(k)}, y) = z\} p(y|x^{(z)}) dy.$$

The k-class *average* Bayes accuracy, on the other hand, is defined uniquely for any joint distribution,

$$\mathrm{ABA}_k[p(x,y)] = \sup_f \Pr[f(X^{(1)},...,X^{(k)},Y) = Z] = \mathbf{E}[\mathrm{BA}(X_1,\ldots,X_k)], \quad (1)$$

where $X_1,\ldots,X_K$ are drawn i.i.d. from $p(x)$.

Our theoretical contributions will be to obtain (i) an upper bound on $\mathrm{Var}[\mathrm{BA}(X_1,\ldots,X_k)]$, enabling inference of ABA_k, and (ii) a lower bound on mutual information $\mathrm{I}[p(x,y)]$ as a function of average Bayes accuracy $\mathrm{ABA}_k[p(x,y)]$ for any continuous joint density function $p(x,y)$, and (iii) derivation of data-based lower confidence bounds based on these results.

3 Variance of Average Bayes Accuracy

3.1 Identities

The following closed-form expressions of $\mathrm{BA}(X^{(1)},\ldots,X^{(k)})$ and ABA_k are extremely useful for deriving theoretical results.

Theorem 3.1 *For joint density $p(x,y)$, we have*

$$BA(x^{(1)},\ldots,x^{(k)}) = \frac{1}{k}\int dy \max_i p(y|x_i).$$

and

$$ABA_k = \frac{1}{k}\int \left[\prod_{i=1}^{k} p(x^{(i)})dx^{(i)}\right] \int dy \max_i p(y|x^{(i)}).$$

Proof. First, we claim that the supremum in (1) is attained by choosing

$$f(x^{(1)},...,x^{(k)},y) = \mathrm{argmax}_{z\in\{1,\ldots,k\}} p(y|x_z).$$

To show this claim, write

$$\sup_f \Pr[f(X^{(1)},...,X^{(k)},Y) = Z] = \sup_f \frac{1}{k}\int p_X(x^{(1)})\ldots p_X(x^{(k)})p(y|x_{f(x^{(1)},\ldots,x^{(k)},y)})dx^{(1)}\ldots dx^{(k)}dy$$

We see that maximizing $\Pr[f(X^{(1)},...,X^{(k)},Y) = Z]$ over functions f additively decomposes into infinitely many subproblems, where in each subproblem we are given $\{x^{(1)},...,x^{(k)},y\} \in \mathcal{X}^k \times \mathcal{Y}$, and our goal is to choose

$f(x^{(1)}, ..., x^{(k)}, y)$ from the set $\{1, ..., k\}$ in order to maximize the quantity $p(y|x_{f(x^{(1)},...,x^{(k)},y)})$. In each subproblem, the maximum is attained by setting $f(x^{(1)}, ..., x^{(k)}, y) = \text{argmax}_z p(y|x^{(z)})$–and the resulting function f attains the supremum to the functional optimization problem. This proves the claim.

We therefore have

$$p(y|x_{f(x^{(1)},...,x^{(k)},y)}) = \max_{i=1}^{k} p(y|x^{(i)}).$$

Therefore, we can write

$$\begin{aligned}
\text{ABA}_k[p(x,y)] &= \sup_f \Pr[f(X^{(1)}, ..., X^{(k)}, Y) = Z] \\
&= \frac{1}{k} \int p_X(x^{(1)}) \dots p_X(x^{(k)}) p(y|x_{f(x^{(1)},...,x^{(k)},y)}) dx^{(1)} \dots dx^{(k)} dy. \\
&= \frac{1}{k} \int p_X(x^{(1)}) \dots p_X(x^{(k)}) \max_{i=1}^{k} p(y|x^{(i)}) dx^{(1)} \dots dx^{(k)} dy.
\end{aligned}$$

□.

3.2 Variability of Bayes Accuracy

We have

$$\text{ABA}_k = \mathbf{E}[\text{BA}(X^{(1)}, ..., X^{(k)})]$$

where the expectation is over the independent sampling of $X^{(1)}, ..., X^{(k)}$ from $p(x)$.

Therefore, $\text{BA}_k = \text{BA}(X^{(1)}, ..., X^{(k)})$ is already an unbiased estimator of ABA_k. However, to get confidence intervals for ABA_k, we also need to know the variability.

We have the following upper bound on the variability.

Theorem 3.2 *Given joint density $p(x, y)$, for $X^{(1)}, ..., X^{(k)} \overset{iid}{\sim} p(x)$, we have*

$$Var[BA(X^{(1)}, ..., X^{(k)})] \leq \frac{1}{4k}.$$

Proof. According to the Efron-Stein lemma,

$$\text{Var}[\text{BA}(X_1, ..., X_k)] \leq \sum_{i=1}^{k} \mathbf{E}[\text{Var}[\text{BA}|X_1, ..., X^{(i-1)}, X^{(i+1)}, ..., X_k]].$$

which is the same as

$$\mathrm{Var}[\mathrm{BA}(X_1, ..., X_k)] \leq k\mathbf{E}[\mathrm{Var}[\mathrm{BA}|X_1, ..., X^{(k-1)}]].$$

The term $\mathrm{Var}[\mathrm{BA}|X_1, ..., X_{k-1}]$ is the variance of $\mathrm{BA}(X_1, ..., X_k)$ conditional on fixing the first $k-1$ curves $p(y|x_1), ..., p(y|x_{k-1})$ and allowing the final curve $p(y|X_k)$ to vary randomly.

Note the following trivial results

$$-p(y|x_k) + \max_{i=1}^{k} p(y|x_i) \leq \max_{i=1}^{k-1} p(y|x_i) \leq \max_{i=1}^{k} p(y|x_i).$$

This implies

$$\mathrm{BA}(X_1, ..., X_k) - \frac{1}{k} \leq \frac{k-1}{k}\mathrm{BA}(X_1, ..., X_{k-1}) \leq \mathrm{BA}(X_1, ..., X_k).$$

i.e. conditional on $(X_1, ..., X_{k-1})$, BA_k is supported on an interval of size $1/k$. Therefore,

$$\mathrm{Var}[\mathrm{BA}|X_1, ..., X_{k-1}] \leq \frac{1}{4k^2}$$

since $\frac{1}{4c^2}$ is the maximal variance for any r.v. with support of length c. $\square$

4 Mutual information lower bound

4.1 Problem formulation

Let $\mathcal{P}$ denote the collection of all joint densities $p(x, y)$ on finite-dimensional Euclidean space. For $\iota \in [0, \infty)$ define $C_k(\iota)$ to be the largest k-class average Bayes error attained by any distribution $p(x, y)$ with mutual information not exceeding ι:

$$C_k(\iota) = \sup_{p \in \mathcal{P}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x, y)].$$

A priori, $C_k(\iota)$ exists since ABA_k is bounded between 0 and 1. Furthermore, C_k is nondecreasing since the domain of the supremum is monotonically increasing with ι.

It follows that for any density $p(x, y)$, we have

$$\mathrm{ABA}_k[p(x, y)] \leq C_k(\mathrm{I}[p(x, y)]).$$

Hence C_k provides an upper bound for average Bayes error in terms of mutual information.

Conversely we have

$$\mathrm{I}[p(x,y)] \geq C_k^{-1}(\mathrm{ABA}_k[p(x,y)])$$

so that C_k^{-1} provides a lower bound for mutual information in terms of average Bayes error.

On the other hand, there is no nontrivial *lower* bound for average Bayes error in terms of mutual information, nor upper bound for mutual information in terms of average Bayes error, since

$$\inf_{p\in\mathcal{P}:\mathrm{I}[p(x,y)]\leq\iota} \mathrm{ABA}_k[p(x,y)] = \frac{1}{k}.$$

regardless of ι.

The goal of this work is to attempt to compute or approximate the functions C_k and C_k^{-1}.

In the following sections we determine the value of $C_k(\iota)$, leading to the following result.

Theorem 4.1 *For any $\iota > 0$, there exists $c_\iota \geq 0$ such that defining*

$$Q_c(t) = \frac{\exp[ct^{k-1}]}{\int_0^1 \exp[ct^{k-1}]},$$

we have

$$\int_0^1 Q_{c_\iota}(t) \log Q_{c_\iota}(t) dt = \iota.$$

Then,

$$C_k(\iota) = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

We obtain this result by first reducing the problem to the case of densities with uniform marginals, then doing the optimization over the reduced space.

4.2 Reduction

Let $p(x, y)$ be a density supported on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is a subset of $\mathbb{R}^{d_1}$ and $\mathcal{Y}$ is a subset of $\mathbb{R}^{d_2}$, and such that $p(x)$ is uniform on $\mathcal{X}$ and $p(y)$ is uniform on $\mathcal{Y}$.

Now let $\mathcal{P}^{unif}$ denote the set of such distributions: in other words, $\mathcal{P}^{unif}$ is the space of joint densities in Euclidean space with uniform marginals over the marginal supports. In this section, we prove that

$$C_k(\iota) = \inf_{p \in \mathcal{P}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x, y)] = \inf_{p \in \mathcal{P}^{unif}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x, y)],$$

thus reducing the problem of optimizing over the space of all densities to the problem of optimizing over densities with uniform marginals.

Also define $\mathcal{P}^{bounded}$ to be the space of all densities $p(x, y)$ with finite-volume support. Since uniform distributions can only be defined over sets of finite volume, we have

$$\mathcal{P}^{unif} \subset \mathcal{P}^{bounded} \subset \mathcal{P}.$$

Therefore, it is necessary to first show that

$$\inf_{p \in \mathcal{P}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x, y)] = \inf_{p \in \mathcal{P}^{bounded}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x, y)].$$

This is accomplished via the following lemma.

Lemma 4.1 *(Truncation). Let $p(x, y)$ be a density on $\mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$. For all $\epsilon > 0$, there exists a subset $\mathcal{X} \subset \mathbb{R}^{d_x}$ with finite volume with respect to d_x-dimensional Lesbegue measure, and a subset $\mathcal{Y} \subset \mathbb{R}^{d_y}$ with finite volume with respect to d_y-dimensional Lesbegue measure, such that defining*

$$\tilde{p}(x, y) = \frac{I\{(x, y) \in \mathcal{X} \times \mathcal{Y}\}}{\int_{\mathcal{X} \times \mathcal{Y}} p(x, y) dx dy} p(x, y),$$

we have

$$|I[p] - I[\tilde{p}]| < \epsilon$$

and

$$|ABA_k[p] - ABA_k[\tilde{p}]| < \epsilon.$$

Proof. Recall the definition of the Shannon entropy H:

$$\mathrm{H}[p(x)] = -\int p(x) \log p(x) dx.$$

It is a well-known in information theory that

$$\mathrm{I}[p(x,y)] = \mathrm{H}[p(x)] + \mathrm{H}[p(y)] - \mathrm{H}[p(x,y)].$$

There exists a sequence $(\mathcal{X}_i, \mathcal{Y}_i)_{i=1}^{\infty}$ where $(\mathcal{X}_i)_{i=1}^{\infty}$ is an increasing sequence of finite-volume subsets of $\mathbb{R}^{d_x}$ and $(\mathcal{Y}_i)_{i=1}^{\infty}$ is an increasing sequence of finite-volume subsets of $\mathbb{R}^{d_y}$, and $\lim_{i\to\infty} \mathcal{X}_i = \mathbb{R}^{d_x}$, $\lim_{i\to\infty} \mathcal{Y}_j$. Define

$$\tilde{p}_i(x,y) = \frac{I\{(x,y) \in \mathcal{X}_i \times \mathcal{Y}_i\}}{\int_{\mathcal{X}_i \times \mathcal{Y}_i} p(x,y)dxdy} p(x,y)$$

Note that $\tilde{p}_i$ gives the conditional distribution of (X,Y) conditional on $(X,Y) \in \mathcal{X}_i \times \mathcal{Y}_i$. Furthermore, it is convenient to define $\tilde{p}_\infty = p$. We can find some i_1, such that for all $i \geq i_1$, we have

$$\left| \int_{x \notin \mathcal{X}_i} p(x) \log p(x) dx \right| < \frac{\epsilon}{6}$$

$$\left| \int_{y \notin \mathcal{Y}_i} p(y) \log p(y) dy \right| < \frac{\epsilon}{6}$$

$$\left| \int_{(x,y) \notin \mathcal{X}_i \times \mathcal{Y}_i} p(x,y) \log p(x,y) dxdy \right| < \frac{\epsilon}{6}$$

and also such that

$$-\log \left[\int_{x,y \in \mathcal{X}_i \times \mathcal{Y}_i} p(x,y) dxdy \right] < \frac{\epsilon}{2}$$

Then, it follows that

$$|\mathrm{I}[p] - \mathrm{I}[\tilde{p}_i]| < \epsilon$$

for all $i \geq i_1$.

Now we turn to the analysis of average Bayes error. Let f_i denote the Bayes k-class classifier for $\tilde{p}_i(x,y)$ and f_∞ the Bayes k-class classifier for $p(x,y)$: recall that by definition,

$$\mathrm{ABA}_k[\tilde{p}_i] = \Pr_{\tilde{p}_i}[f_i(X_1, ..., X_k, Y) = Z]$$

Define

$$\epsilon_i = \Pr_p[(X_1, ..., X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i];$$

by continuity of probability we have $\lim_i \epsilon_i \to 0$. We claim that

$$|\text{ABA}_k[\tilde{p}_i] - \text{ABA}_k[p]| \leq \epsilon_i.$$

Given the claim, the proof is completed by finding $i > i_1$ such that $\epsilon_i < \epsilon$, and defining $\mathcal{X} = \mathcal{X}_i$, $\mathcal{Y} = \mathcal{Y}_i$.

Consider using f_i to obtain a classification rule for $p(x, y)$: define

$$\tilde{f}_i = \begin{cases} f_i(x_1, ..., x_k, y) & \text{when } (x_1, ..., x_k, y) \in \mathcal{X}_i^k \times \mathcal{Y} \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$\begin{aligned}
\text{ABA}_k[p] =& \sup_f \Pr_p[f(X_1, ..., X_k, Y) = Z] \\
\geq & \\
=& (1 - \epsilon_i) \Pr_p[f_i(X_1, ..., X_k, Y) = Z | (X_1, ..., X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i] \\
& + \epsilon_i \Pr_p[f_i(X_1, ..., X_k, Y) = Z | (X_1, ..., X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i] \\
=& (1 - \epsilon_i) \Pr_{\tilde{p}}[f_i(X_1, ..., X_k, Y) = Z] + \epsilon_i 0 \\
=& (1 - \epsilon_i) \text{ABA}_k[\tilde{p}_i] \geq \text{ABA}_k[\tilde{p}_i] - \epsilon_i.
\end{aligned}$$

In other words, when $\tilde{p}_i$ is close to p, the Bayes classification rule for $\tilde{p}_i$ obtains close to the Bayes rate when the data is generated under p.

Now consider the reverse scenario of using f_p to perform classification under $\tilde{p}_i$. This is equivalent to generating data under $p(x, y)$, performing classification using f, then only evaluating classification accuracy conditional on $(X_1, ..., X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i$. Therefore,

$$
\begin{aligned}
\mathrm{ABA}_k[\tilde{p}_i] =& \sup_f \Pr_{\tilde{p}_i}[f(X_1, ..., X_k, Y) = Z] \\
\geq& \Pr_{\tilde{p}_i}[f_p(X_1, ..., X_k, Y) = Z] \\
=& \Pr_p[f_p(X_1, ..., X_k, Y) = Z | (X_1, ..., X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i] \\
=& \frac{1}{1-\epsilon_i} \Pr_p[I\{(X_1, ..., X_k, Y) \in \mathcal{X}_i^k \times \mathcal{Y}_i\} \text{ and } f_p(X_1, ..., X_k, Y) = Z] \\
\geq& \frac{1}{1-\epsilon_i} \left(1 - \Pr_p[I\{(X_1, ..., X_k, Y) \notin \mathcal{X}_i^k \times \mathcal{Y}_i\}] - \Pr_p[f_p(X_1, ..., X_k, Y) \neq Z]] \right) \\
=& \frac{\mathrm{ABA}_k[p] - \epsilon_i}{1-\epsilon_i} \geq \mathrm{ABA}_k[p] - \epsilon_i.
\end{aligned}
$$

In other words, when $\tilde{p}_i$ is close to p, the Bayes classification rule for p obtains close to the Bayes rate when the data is generated under $\tilde{p}_i$.

Combining the two directions gives $|\mathrm{ABA}_k[\tilde{p}_i] - \mathrm{ABA}_k[p]| \leq \epsilon_i$, as claimed. □

One can go from bounded-volume sets to uniform distributions by adding auxillary variables. To illustrate the intution, consider a density $p(x)$ on a set of bounded volume, $\mathcal{X}$. Introduce a variable W such that conditional on $X = x$, we have w uniform on $[0, p(x)]$. It follows that the joint density $p(x, w) = 1$ and is supported on a set $\mathcal{X}' = \mathcal{X} \times [0, \infty]$. Furthermore, $\mathcal{X}'$ is of bounded volume (in fact, of volume 1) since

$$\int_{\mathcal{X}'} dx = \int_{\mathcal{X}'} p(x, w) dx = 1.$$

Therefore, to accomplish the reduction from $\mathcal{P}$ to $\mathcal{P}^{unif}$, we start with a density $p(x, y) \in \mathcal{P}$, and using Lemma 4.1, find a suitable finite-volume truncation $\tilde{p}(x, y)$. Finally, we introduce auxillary variables w and z so that the expanded joint distribution $p(x, w, y, z)$ has uniform marginals $p(x, w)$ and $p(y, z)$. However, we still need to check that the introduction of auxillary variables preserves the mutual information and average Bayes error; this is the content of the next lemma.

Lemma 4.2 *Suppose X, Y, W, Z are continuous random variables, and that $W \perp Y | Z$, $Z \perp X | Y$, and $W \perp Z | (X, Y)$. Then,*

$$I[p(x, y)] = I[p((x, w), (y, z))]$$

Proof. Due to conditional independence relationships, we have

$$p((x,w),(y,z)) = p(x,y)p(w|x)p(z|y).$$

It follows that

$$\begin{aligned}
\mathrm{I}[p((x,w),(y,z))] &= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p((x,w),(y,z))}{p(x,w)p(y,z)} \\
&= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p(x,y)p(w|x)p(z|y)}{p(x)p(y)p(w|x)p(z|y)} \\
&= \int dxdwdydz\ p(x,y)p(w|x)p(z|w)\log\frac{p(x,y)}{p(x)p(y)} \\
&= \int dxdy\ p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = \mathrm{I}[p(x,y)].
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{ABA}_k[p((x,w),(y,z))] &= \int\left[\prod_{i=1}^{k} p(x_i,w_i)dx_idw_i\right]\int dydz\ \max_i p(y,z|x_i,w_i). \\
&= \int\left[\prod_{i=1}^{k} p(x_i,w_i)dx_idw_i\right]\int dy\ \max_i p(y|x_i)\int dz\ p(z|y). \\
&= \int\left[\prod_{i=1}^{k} p(x_i)dx_i\right]\left[\prod_{i=1}^{k}\int dw_i p(w_i|x_i)\right]\int dy\ \max_i p(y|x_i) \\
&= \mathrm{ABA}_k[p(x,y)].
\end{aligned}$$

□

Combining these lemmas gives the needed reduction, given by the following theorem.

Theorem 4.2 *(Reduction.)*

$$\inf_{p\in\mathcal{P}:I[p(x,y)]\leq\iota} ABA_k[p(x,y)] = \inf_{p\in\mathcal{P}^{unif}:I[p(x,y)]\leq\iota} ABA_k[p(x,y)].$$

The proof is trivial given the previous two lemmas.

4.3 Proof of theorem

Proof of theorem 4.1

Using Theorem 4.2, we have

$$C_k(\iota) = \inf_{p \in \mathcal{P}^{unif}: \mathrm{I}[p(x,y)] \leq \iota} \mathrm{ABA}_k[p(x,y)].$$

Define $f(\iota) = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt$: our goal is to establish that $C_k(\iota) = f(\iota)$. Note that $f(\iota)$ is the same function which appears in Lemma A.5 and the same bound as established in Lemma A.4.

Define the density $p_\iota(x, y)$ where

$$p_\iota(x,y) = \begin{cases} g_\iota(y-x) & \text{for } x \geq y \\ g_\iota(1+y-x) & \text{for } x < y \end{cases}$$

where

$$g_\iota(x) = \frac{d}{dx} G_\iota(x)$$

and G_ι is the inverse of Q_c.

One can verify that $\mathrm{I}[p_\iota] = \iota$, and

$$\mathrm{ABA}_k[p] = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

This establishes that

$$C_k(\iota) \geq \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

It remains to show that for all $p \in \mathcal{P}^{unif}$ with $\mathrm{I}[p] \leq \iota$, that $\mathrm{ABA}_k[p] \leq \mathrm{ABA}_k[p_\iota]$.

Take $p \in \mathcal{P}^{unif}$ such that $\mathrm{I}[p] \leq \iota$. Letting $X_1, ..., X_k \sim \mathrm{Unif}[0,1]$, and $Y \sim \mathrm{Unif}[0,1]$ define $Z_i(y) = p(y|X_i)$. We have $\mathbf{E}(Z(y)) = 1$ and,

$$\mathrm{I}[p(x,y)] = \mathbf{E}(Z(Y) \log Z(Y))$$

while

$$\mathrm{ABA}_k[p(x,y)] = k^{-1} \mathbf{E}(\max_i Z_i(Y)).$$

Letting G_y be the distribution of $Z(y)$, we have

$$E[G_y] = 1$$

$$\mathrm{I}[p(x,y)] = \mathbf{E}(I[G_Y])$$
$$\mathrm{ABA}_k[p(x,y)] = \mathbf{E}(\psi_k[G_Y])$$

where the expectation is taken over $Y \sim \mathrm{Unif}[0,1]$ and where $E[G]$, $I[G]$, and $\psi_k[G]$ are defined as in Lemma A.4.

Define the random variable $J = I[G_Y]$. We have

$$\begin{aligned}
\mathrm{ABA}_k[p(x,y)] &= \mathbf{E}(\psi_k[G_Y]) \\
&= \int_0^1 \psi_k[G_y]dy \\
&\leq \int_0^1 \left(\sup_{G:I[G]\leq I[G_y]} \psi_k[G] \right) dy \\
&= \int_0^1 f(I[G_y])dy = \mathbf{E}[f(J)].
\end{aligned}$$

Now, since f is concave by Lemma A.5, we can apply Jensen's inequality to conclude that

$$\mathrm{ABA}_k[p(x,y)] = \mathbf{E}[f(J)] \leq f(\mathbf{E}[J]) = f(\iota),$$

which completes the proof. □

5 Lower confidence bounds

Having established the theoretical basis of our method, the current section derives methods for lower confidence bounds in various settings.

Within the framework of randomized designs, a number of machine-learning-based modeling approaches can be applied to assess informativity. The most common approaches are based on *classification* and *regression*, but a third approach, *identification*, combines the strengths of both classification and regression.

5.1 Classification

Recall the notation used in section 2.1: the k stimuli exemplars are denoted $\{x^{(1)}, \ldots, x^{(k)}\}$ and the r responses for the ith class are given by $y^{(i),1}, \ldots, y^{(i),r}$.

Recall that *data-splitting*, one creates a *training set* consisting of r_1 repeats per class,

$$\{(x^{(1)}, y^{(1),1}), \ldots, (x^1, y^{(1),r_1}), \ldots, (x^{(k)}, y^{(k),1}), \ldots, (x^{(m)}, y^{(m),r_1})\}$$

and a *test set* consisting of the remaining $r_2 = r - r_1$ repeats.

$$\{(x^{(1)}, y^{(1),r_1+1}), \ldots, (x^{(1)}, y^{(1),r}), \ldots, (x^{(k)}, y^{(k),r_1+1}), \ldots, (x^{(k)}, y^{(k),r_1})\}.$$

One inputs the training data into the classifier to obtain the classification rule f,

$$f = \mathcal{F}(\{(x^{(1)}, y^{(1),1}), \ldots, (x^{(1)}, y^{(1),r_1}), \ldots, (x^{(k)}, y^{(k),1}, \ldots, (x^{(k)}, y^{(k),r_1})\}).$$

The test statistic of interest is the test error, defined as

$$\widehat{\mathrm{GA}} = \frac{1}{kr_2} \sum_{i=1}^{k} \sum_{j=r_1+1}^{r} \mathrm{I}(f(y^{(i),j}) \neq i).$$

It follows that $kr_2\widehat{\mathrm{GA}}$ is a sum of Bernoulli random variables. Approximately,

$$kr_2\widehat{\mathrm{GA}} \sim \mathrm{Binomial}(\mathrm{GA}(f), kr_2)$$

5.2 Identification setting

The different machine-learning-based modeling approaches excel in different experimental settings. Classification-based approaches require large amounts of training data and test data for a small number of stimuli categories. Therefore, an investigator intending to use classification will design the experiment to have a large number of repeats relative to the number of stimulus categories.

In contrast, in the identification-based approach, it is not necessary to have multiple repeats for each stimulus exemplar. On the other hand, it *is* necessary to have a parameterization of the stimulus exemplars. The success of the identification approach depends heavily on having a parameterization which captures the relevant features of the stimulus–this can be a difficult task when studying complex stimuli such as faces or natural images.

While we described the classification approach in section 2, we describe the application of our method using classifiers in section 4.1. Then, in section 4.2, we will introduce the identification task, and how our method can also be combined with the identification-based approach.

6 Results

6.1 Simulation

6.2 Application to data

Kay et al. employed a randomized stimuli design in their 2008 paper, "Identifying Natural Images from Human Brain Activity." The experiment was designed in order to investigate how visual information from natural images is encoded in the V1 through V3 brain regions. The stimulus space, $\mathcal{X}$, consists of 128×128-pixel grayscale photographs. The response data consists of BOLD response in regions V1, V2, and V3 from a single subject. The raw time series were processed to yield a single averaged response vector $y^{(i)}$ for each stimulus $x^{(i)}$, for $i = 1, \ldots, 1870$. The dimensionality of $y^{(i)}$ varies depending on which regions of interest we are discussing, and whether we consider a subset of the voxels in those regions. Let v denote the dimensionality (number of voxels) of y.

Let $x^{(i)}$ denote the native 128×128-pixel representation of the image (i.e. a 16384-dimensional vector with entries between 0 and 1.) One of the goals of the Kay et al. paper is to evaluate competing encoding models. In the context of the study, an *encoding model* is a vector-valued function from the stimulus to a p-dimensional space,

$$\vec{g}(x) = (g_1(x), \ldots, g_p(x)).$$

One of the encoding models studied by Kay et al. is the Gabor wavelet pyramid, $\vec{g}_{Gabor}$, with $p = 10921$. Using a *training* subset of the stimulus-response pairs (x_i, y_i), $i = 1, \ldots, 1750$, Kay et al. fit a regularized linear model

$$y^{(i)} \approx B^T \vec{g}(x^{(i)})$$

where B is a $10921 \times v$-dimensional matrix, which is to be fit to the data. The resulting model is used to obtain predictions $\hat{y}^{(i)}$ for the remaining validation subset, $i = 1751, \ldots, 1870$, by

$$\hat{y}^{(i)} = B^T \vec{g}(x^{(i)}).$$

Kay et al. introduced the novel idea of assessing the accuracy of these predictions using a *classification*-based metric, rather than a usual metric for prediction error such as mean-squared error. One treats the 120 validation

stimuli $x_{1751}, \ldots, x_{1870}$ as distinct classes. Next, for each $i = 1751, \ldots, 1870$, one predicts the class of $y^{(i)}$, by correlating the observed $y^{(i)}$ with the predictions $\hat{y}^{(j)}$. In other words, let f be the mapping from the observed response to the predicted class: f is given by

$$f(y) = \text{argmax}_{j=1751}^{1870} \text{Cor}(y, \hat{y}^{(j)}).$$

The test error is

$$e_{test} = \frac{1}{120} \sum_{i=1751}^{1870} I(i \neq f(y^{(i)})).$$

This test error can be computed using different encoding models $\vec{g}$ and different regions of the brain, or different-sized subsets of voxels. One can therefore get a sense of the differences between encoding models or between brain regions by comparing the resulting test errors.

While differing from a conventional classification task, the Kay et al. *identification* task can also be treated using our framework. An *identification rule* (as opposed to a classification rule) is defined by a function f which takes arguments $\{y, x^{(1)}, \ldots, x^{(k)}\}$ where k can be any positive integer. The function f returns an integer from 1 to k,

$$f(y, x^{(1)}, \ldots, x^{(k)}) \in \{1, \ldots, k\}.$$

One can study competing encoding models by constraining identification rules to using features provided by the encoding, i.e. only considering functions

$$f(y, \vec{g}(x^{(1)}), \ldots, \vec{g}(x^{(k)})) \in \{1, \ldots, k\}.$$

The generalization error for the identification task is defined as

$$e_{gen}(f) = \Pr[f(Y, \vec{g}(x^{(1)}), \ldots, \vec{g}(x^{(k)})) \neq i | Y \sim p(y|x^{(i)})],$$

Now, the Bayes rule will be specific to the encoding model

$$e_{Bayes}(\vec{g}(x^{(1)}), \ldots, \vec{g}(x^{(k)})) = \inf_f \Pr[f(Y, \vec{g}(x^{(1)}), \ldots, \vec{g}(x^{(k)})) \neq i | Y \sim p(y|x^{(i)})].$$

As before, the Bayes identification rule which attains the minimal error can be given explicitly,

$$f_{Bayes}(y, \vec{g}(x^{(1)}), \ldots, \vec{g}(x^{(k)})) = \text{argmin}_{i=1}^{k} p(y|\vec{g}(x^{(i)})).$$

The average Bayes error can be defined as before, but is specific to the encoding model:

$$e_{ABE}(\vec{g}) = \mathbf{E}[e_{Bayes}(\vec{g}(X^{(1)}), \ldots, \vec{g}(X^{(k)}))]$$

Our asymptotic theory links average Bayes error to information and is *not specific* to the whether the task is based on identification or classification. Hence, the approximation

$$e_{ABE,k}(\vec{g}) \approx \pi_k(\sqrt{2I(\vec{g}(X); Y)})$$

still holds in the appropriate regime.

However, one adjustment that needs to be made is how to construct point estimates or confidence intervals for e_{ABE}. Here we will limit our discussion to confidence intervals. The approach taken is to use cross-validation to construct the confidence interval.

Let N be the total number of stimuli exemplars. Choose $T < N$ and $k < N - T$; T denotes the number of *training* exemplars and k controls the number of *test* exemplars.

Set B, the number of Monte Carlo simulations, to be a large number. Let $\beta \in [0, 1]$ be a Bayesian smoothing parameter (typically β is close to 0). For each $b = 1, \ldots, B$, define the bth sampled test error e_b by:

1. Sample T stimuli exemplars without replacement from the library of N exemplars.

2. Train an identification rule f_b using the training set pairs $(\vec{g}(x^{(i)}), y^{(i)})$.

3. Let $S_1, \ldots, S_M$ denote all possible subsets of size k from the remaining $N - T$ exemplars.

4. For each subset $S_j = \{x^{(i_{j,1})}, \ldots, x^{(i_{j,k})}\}$, compute the test error $e_{test,b,j}$ via
$$e_{test,b,j} = \frac{1}{k} \sum_{\ell=1}^{k} I(i_{j,\ell} \neq f(y^{(i_{j,\ell})}, \vec{g}(x^{(i_{j,1})}), \ldots, \vec{g}(x^{(i_{j,k})})))$$

5. Set $e_b = \beta \frac{k-1}{k} + (1 - \beta) \frac{1}{M} \sum_{j=1}^{M} e_{test,b,j}$.

For most identification rules, e_b can be computed in $O(N^2)$ time by using the hypergeometric sampling formula; e.g., see Nishimoto (2011) for details.

Figure 1: Confidence intervals for $I(\vec{g}_{Gabor}(X); Y)$, where Y are subsamples of size $\{100, 200, \ldots, 600\}$ voxels from V1.

Figure 2: Confidence intervals for $I(\vec{g}_{reduced}(X); Y)$, where Y are subsamples of size $\{100, 200, \ldots, 600\}$ voxels from V1.

Setting the parameter $\beta > 0$ is typically recommended in order to reduce variance in the procedure.

Now, choosing a level α, let $\bar{e}(\vec{g})$ be the $(1 - \alpha/2)$th quantile and $\underline{e}(\vec{g})$ be the $(\alpha/2)$th quantile of $\{e_b\}_{b=1}^{B}$. Then one constructs the confidence interval

$$[\frac{1}{2}(\pi_k^{-1}(\bar{e}))^2, \frac{1}{2}(\pi_k^{-1}(\underline{e}))^2].$$

for $I(\vec{g}(X); Y)$.

We computed such confidence intervals for $\vec{g}_{Gabor}$, taking Y to be random subsamples of size $\{100, 200, \ldots, 600\}$ voxels from V1, and also varying the parameter k within the procedure. The confidence intervals are shown in Figure ?.

Using the same voxel subsamples and parameters k, but now using a reduced encoding model $\vec{g}_{reduced}$ with $p = 681$, we computed confidence intervals for $I(\vec{g}_{reduced}(X); Y)$.

7 Discussion

Since the Bayes error is the large-sample limit of the achieved classification error, a promising approach is to perform classification using differently sized subsamples of the training data, producing a plot of classification error versus sample size–a "learning curve." One can then extrapolate the learning curve to estimate the Bayes error (Cortes et al. 1994.)However, much work remains to develop rigorous methodology for estimating Bayes error, and so we leave this first issue for future work.

We should keep in mind that the definition of mutual information depends on the specific class of stimulus that is of interest in the experiment. The Bayes error itself depends on the choices made by the experimenter in regards to the stimuli exemplar chosen in the experiment, and the decision of how

to partition those exemplars in the classification task. For example, Nishimoto et al classified segments of a movie clips based on activation patterns, but the definition of the classification task, and the achievable classification performance, depends not only on the particular move clips used in the experiment, but also the choice of time interval used to define discrete classes: defining each class to be a 1sec segment of movie results in more distinct classes and lower classification accuracy than defining each class to be a 4sec segment of movie. The Bayes error, and any estimate of mutual information based on the Bayes error, would therefore be necessarily dependent on the experimental parameters.

Estimating e_{Bayes} is much more difficult in practice than estimating e_{gen}; however, nonparametric estimators of e_{Bayes} have been proposed in the literature. Cover (1969) first noted that the leave-one-out error of 1-nearest neighbor, e_{1nn}, asymptotically bounds the Bayes error in the sense that

$$\frac{1}{2}\mathbf{E}[e_{1nn}] \leq e_{Bayes} \leq \mathbf{E}[e_{1nn}].$$

Later, Fukunaga and Kessel (1971) and Fralick and Scott (1971) both proposed a consistent estimator of e_{Bayes} using the test error from kernel density estimation-based classifiers. Fukunaga and Hostetler 1975 improved on the density estimation approach by estimating the generalization error directly, rather than using empirical test error (hence reducing variance.)

Nevertheless, the convergence rates of nonparametric methods of Bayes risk estimation are too slow to provide much assurance for their use in high-dimensional problems. On the other hand, the empirical success of supervised learning approaches seems to justify a certain optimism that at least one of the popular black-box methods (random forests, neural networks, kernel SVM) can come close to the Bayes error rate within realistic sample sizes. Methods such as random forests and neural networks are known to have *universal approximation* properties that guarantee consistent recovery of the classification rule, given that the model complexity is scaled at an appropriate rate as the sample size increases; but methods such as kernel SVM have much more restrictive function classes and therefore have no guarantee of being able to recover the Bayes error. Yet, it is seen in many learning tasks that the test error is very similar for very different classification methods, some universal and some non-universal, suggesting that the true signal lies in a low-complexity function class which can be adequately approximated using a variety of methods. Empirical evidence for these phenomenon are collected

in studies on the learning curves of classifers, beginning with Cortes 1994, and applied to specific applications by Figueroa et al. 2012, Beleites et al. 2013. We leave further discussion of Bayes error estimation for future work; in the theoretical portion of the current paper, we will simply leave unspecified the method used to estimate Bayes error, and in our simulation and real data examples, we will use an ad hoc approach for estimating Bayes error from test error for the purpose of demonstrating our method.

8 Conclusions

- We derive a relationship between average Bayes error (ABE) and mutual information (MI), motivating a novel estimator $\hat{I}_{LS}$.
- Theory based on high dimensional, low SNR limit, where

$$\text{ABE} \leftrightarrow \text{MI}.$$

- In ideal settings for supervised learning, ABE can be estimated effectively and $\hat{I}_{LS}$ can recover MI at much lower sample sizes than nonparametric methods.
- In simulations, $\hat{I}_{LS}$ works better than Fano's inequality or the confusion matrix approach.

A Appendix

Lemma A.1 *Let $f(t)$ be an increasing function from $[a, b] \to \mathbb{R}$, where $a < b$, and let $g(t)$ be a bounded continuous function from $[a, b] \to \mathbb{R}$. Define the set*

$$A = \{t : f(t) \neq g(t)\}.$$

Then, we can write A as a countable union of intervals

$$A = \bigcup_{i=1}^{\infty} A_i$$

where A_i are mutually disjoint intervals, with $\inf A_i < \sup A_i$, and for each i, either $f(t) > g(t)$ for all $t \in A_i$ or $f(t) < g(t)$ for all $t \in A_i$.

Proof of Lemma A.1. (This will appear in the appendix of the paper.)

The function $h(t) = f(t) - g(t)$ is measurable, since all increasing functions are measurable. Define $A^+ = \{t : f(t) > g(t)\}$ and $A^- = \{t : f(t) < g(t)\}$. Since A^+ and A^- are measurable subsets of $\mathbb{R}$, they both admit countable partitions consisting of open, closed, or half-open intervals. Let $\mathcal{H}^+$ be the collection of all partitions of A^+ consisting of such intervals. There exists a least refined partition $\mathcal{A}^+$ within $\mathcal{H}^+$. Define $\mathcal{A}^-$ analogously, and let

$$\mathcal{A} = \mathcal{A}^+ \cup \mathcal{A}^-$$

and enumerate the elements

$$\mathcal{A} = \{A_i\}_{i=1}^{\infty}.$$

We claim that the partitions $\mathcal{A}^+$ and $\mathcal{A}^-$ have the the property that for all $t \in A^{\pm}$, ther interval $I \in \mathcal{A}^{\pm}$ containing t has endpoints $l \leq u$ defined by

$$l = \inf_{x \in [a,b]} \{x : \text{Sign}(h([x,t])) = \{\text{Sign}(h(t))\}\}$$

and

$$u = \sup_{x \in a[,b]} \{x : \text{Sign}(h([t,x])) = \{\text{Sign}(h(t))\}\}.$$

We prove the claim for the partition $\mathcal{A}^+$. Take $t \in A^+$ and define l and u as above. It is clear that $(l, u) \in A^+$, and furthermore, there is no $l' < l$ and $u' > u$ such that $(l', x) \in A^+$ or $(x, u') \in A^+$ for any $x \in I$. Let $\mathcal{H}$ be any other partition of A^+. Some disjoint union of intervals $H_i \in \mathcal{H}$ necessarily covers I for $i = 1, ...,$ and we can further require that none of the H_i are disjoint with I. Since each H_i has nonempty intersection with I, and I is an interval, this implies that $\cup_i H_i$ is also an interval. Let $l'' \leq u''$ be the endpoints of $\cup_i H_i$ Since $I \subseteq \cup_i H_i$, we have $l'' \leq l \leq u \leq u''$. However, since also $I \in A^+$, we must have $l \leq l'' \leq u'' \leq u$. This implies that $l'' = l$ and $u'' = u$. Since $\cup_i H_i = I$, and this holds for any $I \in \mathcal{A}^+$, we conclude that $\mathcal{H}$ is a refinement of $\mathcal{A}^+$. The proof of the claim for $\mathcal{A}^-$ is similar.

It remains to show that there are not isolated points in $\mathcal{A}$, i.e. that for all $I \in \mathcal{A}$ with endpoints $l \leq u$, we have $l < u$. Take $I \in \mathcal{A}$ with endpoints $l \leq u$ and let $t = \frac{l+u}{2}$. By definition, we have $h(t) \neq 0$. Consider the two cases $h(t) > 0$ and $h(t) < 0$.

If $h(t) > 0$, then $t' = g^{-1}(h(t)) > t$, and for all $x \in [t, t']$ we have $h(x) > 0$. Therefore, it follows from definition that $[t, t'] \in I$, and since $l \leq t < t' \leq u$, this implies that $l < u$. The case $h(t) < 0$ is handled similarly. $\square$

Lemma A.2 *Let $f(t)$ be a measurable function from $[a,b] \to \mathbb{R}$, where $a < b$. Then there exists sets $\mathcal{B}_0$ and $\mathcal{B}_1$, satisfying the following properties:*

- *$\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ is countable partition of $[a,b]$,*
- *$f(t)$ is constant on all $B \in \mathcal{B}_0$, but not constant on any proper superinterval $B' \supset B$, and*
- *$B \in \mathcal{B}_1$ contains no positive-length subinterval where $f(t)$ is constant.*

Proof of Lemma A.2. (This will appear in the appendix of the paper.) To construct the interval, define

$$l(t) = \inf\{x \in [0,1] : f([x,t]) = \{f(t)\}\}$$

$$u(t) = \sup\{x \in [0,1] : f([t,x]) = \{f(t)\}\},$$

Let B_0 be the set of all t such that $l(t) < u(t)$, and let B_1 be the set of all t such that $l(t) = t = u(t)$. For all $t \in B_0$, define

$$I(t) = (l(t), u(t)) \cup \{x \in \{l(t), u(t)\} : f(x) = f(t)\}.$$

Then we claim

$$\mathcal{B}_0 = \{I(t) : t \in B_0\}$$

is a countable partition of B_0. The claim follows since the members of $\mathcal{B}_0$ are disjoint intervals of nonzero length, and B_0 has finite length. It follows from definition that for any $B \in B_0$, that f is not constant on any proper superinterval $B' \supset B$.

Meanwhile, let $\mathcal{B}_1$ be a countable partition of B_1 into intervals.

Next, we show that for all $I \in \mathcal{B}_1$, I does not contain a subinterval I' of nonzero length such that f is constant on I'. Suppose to the contrary, we could find such an interval I and subinterval I'. Then for any $t \in I'$, we have $t \in B_0$. However, this implies that $t \notin B_1$, a contradiction.

Since $t \in [a,b]$ belongs to either B_0 or B_1, letting $\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ yields the desired partition of $[a,b]$. $\square$.

Lemma A.3 *Define an exponential family on $[0,1]$ by the density function*

$$q_\beta(t) = \exp[\beta t^{k-1} - \log Z(\beta)]$$

where

$$Z(\beta) = \int_0^1 \exp[\beta t^{k-1}]dt.$$

Then, the negative entropy

$$I(\beta) = \int_0^1 q_\beta(t) \log q_\beta(t) dt$$

is decreasing in β on the interval $(-\infty, 0]$. and increasing on the interval $[0, \infty)$.

Furthermore, for any $\iota \in (0, \infty)$, there exist two solutions to $I(\beta) = \iota$: one positive and one negative.

Proof of Lemma A.3.

Define $\beta(\mu)$ as the solution to

$$\mu = \int_0^1 t q_\beta(t) dt.$$

By [Wainwright and Jordan 2008], the function $\beta(\mu)$ is well-defined. Furthermore, since the sufficient statistic t^{k-1} is increasing in t, it follows that $\beta(\mu)$ is increasing.

Define the negative entropy as a function of μ,

$$N(\mu) = \int_0^1 q_{\beta(\mu)}(t) \log q_{\beta(\mu)}(t) dt.$$

By Theorem 3.4 of [Wainwright and Jordan 2008], $N(\mu)$ is convex in μ. We claim that the derivative of $N(\mu) = 0$ at $\mu = \frac{1}{2}$. This implies that $N(\mu)$ is decreasing in μ for $\mu \leq \frac{1}{2}$ and increasing for $\mu \geq \frac{1}{2}$. Since $I(\beta(\mu)) = N(\mu)$, β is increasing in μ, and $\beta(\frac{1}{2}) = 0$, this implies that $I(\beta)$ is decreasing in β for $\beta \leq 0$ and increasing for $\beta \geq 0$.

We will now prove the claim. Write

$$\frac{d}{d\mu} N(\mu)\Big|_{\mu=1/2} = \frac{d}{d\beta} I(\beta(\mu))\Big|_{\beta=0} \frac{d\beta}{d\mu}\Big|_{\mu=1/2}.$$

We have

$$\frac{d}{d\beta} I(\beta) = \beta \int q_\beta t^{k-1} dt - \log Z(\beta).$$

Meanwhile, $Z(0) = 1$ so $\log Z(0) = 0$. Therefore,

$$\frac{d}{d\beta} I(\beta)\bigg|_{\beta=0} = 0.$$

This implies that $\frac{d}{d\mu} N(\mu)|_{\mu=1/2} = 0$, as needed.

For the final statement of the lemma, note that $I(0) = 0$ since q_0 is the uniform distribution. Meanwhile, since q_β tends to a point mass as either $\beta \to \infty$ or $\beta \to -\infty$, we have

$$\lim_{\beta \to \infty} I(\beta) = \lim_{\beta \to -\infty} I(\beta) = \infty.$$

And, as we can check that $I(\beta)$ is continuous in β, this means that

$$I((-\infty, 0]) = I([0, \infty)) = [0, \infty)$$

by the mean-value theorem. Combining this fact with the monotonicity of $I(\beta)$ restricted to either the positive and negative half-line yields the fact that for any $\iota > 0$, there exists $\beta_1 < 0 < \beta_2$ such that $I(\beta_1) = I(\beta_2) = \iota$. $\square$.

Lemma A.4 *For any measure G on $[0, \infty]$, let G^k denote the measure defined by*

$$G^k(A) = G(A)^k,$$

and define

$$E[G] = \int x dG(x).$$

$$I[G] = \int x \log x dG(x)$$

and

$$\psi_k[G] = \int x d(G^k)(x).$$

Then, defining Q_c and c_ι as in Theorem 1, we have

$$\sup_{G: E[G]=1, I[G] \leq \iota} \psi_k[G] = \int_0^1 Q_{c_\iota}(t) t^{k-1} dt.$$

Furthermore, the supremum is attained by a measure G that has cdf equal to Q_c^{-1}, and thus has a density g with respect to Lesbegue measure.

Proof of Lemma A.4. (This will appear in the appendix of the paper.)

Consider the quantile function $Q(t) = \inf_{x\in[0,1]} : G((-\infty, x]) \geq t$. $Q(t)$ must be a monotonically increasing function from $[0,1]$ to $[0,\infty)$. Let $\mathcal{Q}$ denote the collection of all such quantile functions.

We have

$$E[G] = \int_0^1 Q(t)dt$$

$$\psi_k[G] = \int_0^1 Q(t)x^{k-1}dt.$$

and

$$I[G] = \int_0^1 Q(t)\log Q(t)dt.$$

For any given ι, let P_ι denote the class of probability distributions G on $[0,\infty]$ such that $E[G] = 1$ and $I[G] \leq \iota$. From Markov's inequality, for any $G \in P_\iota$ we have

$$G([x,\infty]) \leq x^{-1}$$

for any $x \geq 0$, hence P_ι is tight. From tightness, we conclude that P_ι is closed under limits with respect to weak convergence. Hence, since ψ_k is a continuous function, there exists a distribution $G^* \in P_\iota$ which attains the supremum

$$\sup_{G\in P_\iota} \psi_k[G].$$

Let $\mathcal{Q}_\iota$ denote the collection of quantile functions of distributions in P_ι. Then, $\mathcal{Q}_\iota$ consists of monotonic functions $Q : [0,1] \to [0,\infty]$ which satisfy

$$E[Q] = \int_0^1 Q(t)dt = 1,$$

and

$$I[Q] = \int_0^1 Q(t)\log Q(t)dt \leq \iota.$$

Let $\mathcal{Q}$ denote the collection of *all* quantile functions from measures on $[0,\infty]$. And letting Q^* be the quantile function for G^*, we have that Q^* attains the supremum

$$\sup_{Q\in\mathcal{Q}_\iota} \phi_k[Q] = \sup_{Q\in\mathcal{Q}_\iota} \int_0^1 Q(t)t^{k-1}dt. \tag{2}$$

Therefore, there exist Lagrange multipliers $\lambda \geq 0$ and $\nu \geq 0$ such that defining

$$\mathcal{L}[Q] = -\phi_k[Q] + \lambda E[Q] + \nu I[Q] = \int_0^1 Q(t)(-t^{k-1} + \lambda + \nu \log Q(t))dt,$$

Q^* attains the infimum of $\mathcal{L}[Q]$ over *all* quantile functions,

$$\mathcal{L}[Q^*] = \inf_{Q \in \mathcal{Q}} \mathcal{L}[Q].$$

The global minimizer Q^* is also necessarily a stationary point: that is, for any perturbation function $\xi : [0,1] \to \mathbb{R}$ such that $Q^* + \xi \in \mathcal{Q}$, we have $\mathcal{L}[Q^*] \leq \mathcal{L}[Q^* + \xi]$. For sufficiently small ξ, we have

$$\mathcal{L}[Q + \xi] \approx \mathcal{L}[Q] + \int_0^1 \xi(t)(-t^{k-1} + \lambda + \nu + \nu \log Q(t))dt. \tag{3}$$

Define

$$\nabla \mathcal{L}_{Q^*}(t) = -t^{k-1} + \lambda + \nu + \nu \log Q(t). \tag{4}$$

The function $\nabla \mathcal{L}_{Q^*}(t)$ is a *functional derivative* of the Lagrangian. Note that if we were able to show that $\nabla \mathcal{L}_{Q^*}(t) = 0$, this immediately yields

$$Q^*(t) = \exp[-1 - \lambda\nu^{-1} + \nu^{-1}t^{k-1}]. \tag{5}$$

At this point, we know that the right-hand side of (5) gives a stationary point of $\mathcal{L}$, but we cannot be sure that it gives the global minimzer. The reason is because the optimization occurs on a constrained space. We will show that (5) indeed gives the global minimizer Q^*, but we do so by showing that the set of points t where $\nabla \mathcal{L}_{Q^*}(t) \neq 0$ is of zero measure. Since sets of zero measure don't affect the integrals defining the optimization problem (2), we conclude there exists a global optimal solution with $\nabla \mathcal{L}_{Q^*}(t) = 0$ everywhere, which is therefore given explicitly by (5) for some $\lambda \in \mathbb{R}$, $\nu \geq 0$.

We will need the following result: that for $\iota > 0$, any solution to (2) satisfies $\phi_k[Q] < 1$. This follows from the fact that

$$E[Q] - \phi_k[Q] = \int_0^1 (1 - t^{k-1})Q(t)dt,$$

where the term $(1 - t^{k-1})$ is negative, except for the one point $t = 1$. Therefore, in order for $\phi_k[Q] = 1 = E[Q]$, we must have $Q(t) = 0$ for $t < 1$.

However, this yields a contradiction since $Q(t) = 0$ for $t < 1$ implies that $E[Q] = 0$, a violation of the hard constraint $E[Q] = 1$.

Let us establish that $\nu > 0$: in other words, the constraint $I[Q] = \iota$ is tight. Suppose to the contrary, that for some $\iota > 0$, the global optimum Q^* minimizes a Lagrangian with $\nu = 0$. Let $\phi^* = \phi_k[Q^*] < 1$. However, if we define $Q_\kappa(t) = I\{t \geq 1 - \frac{1}{\kappa}\}\kappa$, we have $E[Q_\kappa] = 1$, and also for some sufficiently large $\kappa > 0$, $\phi_k[Q_\kappa] > \phi^*$. But since the Lagrangian lacks a term corresponding to $I[Q]$, we conclude that $\mathcal{L}[Q_\kappa] < \mathcal{L}[Q^*]$, a contradiction.

The rest of the proof proceeds as follows. We will use Lemmas A.1 and A.2 to define a decomposition $A = D_0 \cup D_1 \cup D_2$, where D_2 is of measure zero. First, we show that assuming the existence of $t \in D_0$ yields a contradiction, and hence $D_0 = \emptyset$. Then, again using argument from contradiction we establish that $D_1 = \emptyset$. Finally, since D_2 is a set of zero measure, this allows us to conclude that the $Q^*(t) = 0$ on all but a set of zero measure.

We will now apply the Lemmas to obtain the necessary ingredients for constructing the sets D_i. Since $\nabla\mathcal{L}_{Q^*}(t)$ is a difference between an increasing function and a continuous stricly increasing function, we can apply Lemma A.1 to conclude that there exists a countable partition $\mathcal{A}$ of the set $A : \{t \in [0,1] : \nabla\mathcal{L}_{Q^*}(t) \neq 0\}$ into intervals such that for all $J \in \mathcal{A}$, $|\text{Sign}(\nabla Q^*(J))| = 1$ and $\inf J < \sup I$. Applying Lemma A.2 we get a countable partition $\mathcal{B} = \mathcal{B}_0 \cup \mathcal{B}_1$ of $[0,1]$ so that each element $J \in \mathcal{B}_0$ is an interval such that $\nabla\mathcal{L}_{Q^*}(t)$ is constant on J, and furthermore is not properly contained in any interval with the same property, and each element $J \in \mathcal{B}_1$ is an interval, such that J contains no positive-length subinterval where $\nabla\mathcal{L}_{Q^*}(t)$ is constant. Also define B_i as the union of the sets in $\mathcal{B}_i$ for $i = 0, 1$.

Note that B_0 is necessarily a subset of A. That is because if $\nabla\mathcal{L}_{Q^*}(t) = 0$ on any interval J, then that $Q^*(t)$ is necessarily not constant on the interval.

We will construct a new countable partition of A, called $\mathcal{D}$. The partition $\mathcal{D}$ is constructed by taking the union of three families of intervals,

$$\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1 \cup \mathcal{D}_2.$$

Define D_i to be the union of intervals in $\mathcal{D}_i$ for $i = 0, 1, 2$.

Define $\mathcal{D}_0 = \mathcal{B}_0$, Define a countable partition $\mathcal{D}_1$ by

$$\mathcal{D}_1 = \{J \cap L : J \in \mathcal{A}, L \in \mathcal{B}_1, \text{ and } |L| > 1\},$$

in order words, $\mathcal{D}_1$ consists of positive-length intervals where $\nabla Q^*(t)$ is entirely positive or negative and is not constant. Define

$$\mathcal{D}_2 = \{J \in \mathcal{B}_1 : J \subset A \text{ and } |J| = 1\},$$

i.e. $\mathcal{D}_2$ consists of isolated points in A.

One verifies that $\mathcal{D}$ is indeed a partition of A by checking that $D_0 = B_0$, $D_1 \cup D_2 = B_1 \cap A$, so that $D_0 \cup D_2 \cup D_2 = A$: it is also easy to check that elements of $\mathcal{D}$ are disjoint. Furthermore, as we mentioned earlier, the set D_2 is indeed of zero measure, since it consists of countably many isolated points.

Now we will show that the existence of $t \in D_0$ implies a contradiction. Take $t \in D$ for $D \in \mathcal{D}_0$, and let $a = \inf D$ and $b = \sup D$. Define

$$\xi^+ = I\{t \in D\}(Q^*(b) - Q^*(t))$$

and

$$\xi^- = I\{t \in D\}(Q^*(a) - Q^*(t)).$$

Observe that $Q + \epsilon\xi^+ \in \mathcal{Q}$ and $Q + \epsilon\xi^- \in \mathcal{Q}$ for any $\epsilon \in [0, 1]$. Now, if $\nabla\mathcal{L}_{Q^*}(t)$ is strictly positive on D, then for some $\epsilon > 0$ we would have $\mathcal{L}[Q^* + \epsilon\xi^-] < \mathcal{L}[Q^*]$, a contradiction. A similar argument with ξ^+ shows that $\nabla\mathcal{L}_{Q^*}(t)$ cannot be strictly negative on D either. From this pertubation argument, we conclude that $\nabla\mathcal{L}_{Q^*}(t) = 0$. Since this argument applies for all $t \in D_0$, we conclude that $D_0 = \emptyset$: therefore, on the set $[0, 1] \setminus (D_1 \cup D_2)$, we have $\nabla\mathcal{L}_{Q^*}(t) = 0$.

The following observation is needed for the next stage of the proof. If we look at the function $Q^*(t)$, then up so sets of neglible measure, it is given by the expression (5) on the set $[0, 1] \setminus D_1$, and it is piecewise constant in-between. But since (5) gives a strictly increasing function, and since Q^* is increasing, this implies that Q^* is discontinuous at the boundary of D_1.

Now we are prepared to show that $\nabla\mathcal{L}_{Q^*}(t) = 0$ for $t \in D_1$. Take $t \in D$ for $D \in \mathcal{D}_1$, and let $a = \inf D$ and $b = \sup D$. From the previous argument, there is a discontinuity at both a and b, so that $\lim_{u \to a^-} Q(u) < Q(t) < \lim_{u \to b^+} Q(u)$. Therefore, for any $\xi(t)$ which is increasing on D and zero elsewhere, there exists $\epsilon > 0$ such that $\nabla Q^* + \epsilon\xi \in \mathcal{Q}$. It remains to find such a perturbation ξ such that $\mathcal{L}[Q + \epsilon\xi] < \mathcal{L}[Q]$.

Also, since by definition $\nabla\mathcal{L}_{Q^*}(t)$ is constant on D, follows from(4) that ∇Q^* is strictly decreasing, and thus either

- Case 1: $\nabla\mathcal{L}_{Q^*}(t) \geq 0$ on D,
- Case 2: $\nabla\mathcal{L}_{Q^*}(t) \leq 0$ on D, or
- Case 3: $\nabla\mathcal{L}_{Q^*}(t) \geq 0$ for all $t \in D \cap [a, t_0]$ and $\nabla\mathcal{L}_{Q^*}(t) \leq 0$ for all $t \in D \cap [t_0, b]$.

Depending on the case, we construct a suitable perturbation ξ:

- Case 1: Construct $\xi(t) = -I\{t \in D\}$.
- Case 2: Construct $\xi(t) = I\{t \in D\}$
- Case 2: Construct

$$\xi(t) = \begin{cases} -1 & \text{for } t \in D \cap [a, t_0], \\ 0 & \text{otherwise.} \end{cases}$$

In all three cases, given the corresponding construction for $\xi(t)$ we get

$$\int_0^1 \xi(t) \nabla \mathcal{L}_{Q^*}(t) dt < 0.$$

Therefore, from (3), there exists some $\epsilon > 0$ such that $\mathcal{L}[Q + \epsilon\xi] < \mathcal{L}[Q]$, a contradiction. Again, since the contradiction applies for all $t \in D_1$, we conclude that $D_1 = \emptyset$.

By now we have established that a global optimum for (2) exists, and is given by (5) for some $\lambda \in \mathbb{R}$, $\nu > 0$. It remains to determine the values of λ and ν.

Reparameterize $\alpha = \exp[-1 - \lambda\nu^{-1}]$ and $\beta = \nu^{-1}$. Therefore,

$$Q^*(t) = \alpha \exp[\beta t^{k-1}]$$

for $\alpha > 0$, $\beta > 0$. There is a one-to-one mapping from $(\alpha, \beta) \in (0, \infty)^2$ to $(\lambda, \nu) \in \mathbb{R} \times (0, \infty)$.

Now, from the constraint

$$1 = E[Q^*] = \int_0^1 \alpha \exp[\beta t^{k-1}] dt.$$

we conclude that

$$\alpha = \frac{1}{\int_0^1 \exp[\beta t^{k-1}] dt.}$$

Therefore, we have reduced the set of possible solutions Q^* to a one-parameter family,

$$Q^*(t) = \frac{\exp[\beta t^{k-1}]}{Z(\beta)}.$$

where

$$Z(\beta) = \int_0^1 \exp[\beta t^{k-1}]dt.$$

Next, note that

$$I[Q^*] = \int_0^1 Q^*(t) \log Q^*(t) = \beta\mu_\beta - \log Z(\beta),$$

as a function of β, is completely characterized by Lemma A.3. Let us define c_ι as the unique positive solution to the equation

$$c_\iota \mu_{c_\iota} - \log Z(c_\iota) = \iota$$

given by Lemma A.3. We therefore have

$$Q^*(t) = \frac{\exp[c_\iota t^{k-1}]}{\int_0^1 \exp[c_\iota t^{k-1}]},$$

as needed. □

Lemma A.5 *The map*

$$\iota \to \int_0^1 Q_{c_\iota}(t) t^{k-1} dt$$

is concave in $\iota > 0$.

Proof of Lemma A.5. It is equivalent to show that the inverse function

$$C_k^{-1}(p) = \inf_{G: E[G]=1, \phi_k[G]=p} I[G]$$

is convex. Let $p_1, p_2 \in [0, 1]$. From lemma A.4, we can find measures G_1, G_2 on $[0, \infty)$ which minimize $I[G_i]$ subject to $E[G_i] = 1$ and $\phi_k[G_i] = p_i$. Define the measure

$$H = \frac{G_1 + G_2}{2}.$$

Since ϕ_k is a linear functional,

$$\phi_k[H] = \frac{\phi_k[G_1] + \phi_k[G_2]}{2} = \frac{p_1 + p_2}{2}.$$

But since I is a convex functional,

$$I[H] \leq \frac{I[G_1] + I[G_2]}{2}.$$

Therefore,

$$C_k^{-1}\left(\frac{p_1+p_2}{2}\right) \leq I[H] = \frac{I[G_1] + I[G_2]}{2} = \frac{C_k^{-1}(p_1) + C_k^{-1}(p_2)}{2}.$$

□.

B References

- Gastpar, M. Gill, P. Huth, A. Theunissen, F. "Anthropic Correction of Information Estimates and Its Application to Neural Coding." *IEEE Trans. Info. Theory*, Vol 56 No 2, 2010.
- A. Borst and F. E. Theunissen, "Information theory and neural coding" Nature Neurosci., vol. 2, pp. 947?957, Nov. 1999.
- L. Paninski, "Estimation of entropy and mutual information," Neural Comput., vol. 15, no. 6, pp. 1191?1253, 2003.
- I. Nelken, G. Chechik, T. D. Mrsic-Flogel, A. J. King, and J. W. H. Schnupp, "Encoding stimulus information by spike numbers and mean response time in primary auditory cortex," J. Comput. Neurosci., vol. 19, pp. 199?221, 2005.
- Cover and Thomas. Elements of information theory.
- Muirhead. Aspects of multivariate statistical theory.
- van der Vaart. Asymptotic statistics.