A saturating lower confidence bound for mutual information based on classification error

Anonymous Author(s)
Affiliation
Address
email

Abstract

Estimating the mutual information $I(X;Y)$ based on observations becomes statistically infeasible in high dimensions without some kind of modeling assumption. One approach is to assume a parametric joint distribution on (X, Y), but in many applications, such a strong modeling assumption cannot be justified. An alternative approach is to obtain a lower bound on the mutual information based on a classification task. Existing methods include lower confidence bounds based on the confusion matrix of the classifier, as well as Fano's inequality and its generalizations. One might hope that if the classifier is *consistent*, in the sense that the classification error approaches the Bayes error in the large-sample limit, that the information lower bound $\underline{I}(X, Y)$ should also approach the true information $I(X;Y)$. However, existing methods always produce a bound which is on the order $O(\log k)$, where K is the number of classes, so when $I(X;Y) \gg \log k$, the lower confidence bound is inconsistent even when the classifier is consistent. On the other hand, consistency is not possible with a fixed number of classes since the full distribution of X is not revealed. In this paper, we construct a novel lower bound based on high-dimensional asymptotics; our proposed bound satisfies a weaker property than consistency, called *saturation.* A saturating lower bound has the property that as $I(X;Y)$ and the number of observations grow to infinity (while the number of classes K stays fixed,) that $\underline{I}((X, Y)) = O(I(X, Y))$, assuming that the classifier used is consistent. While the theory is based on a large-sample, high-dimensional limit, we demonstrate through simulations that our proposed lower confidence bound has superior performance to the alternatives in problems of moderate dimensionality.

1 Introduction

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

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

In its original context of information theory, the mutual information describes the rate at which a noisy communications channel Y can communicate bits from a source stream X, but by now, the quantity $I(X, Y)$ has found many new uses in science and engineering. Mutual information is used to test for conditional independence, to quantifying the information between a random stimulus X and the signaling behavior of an ensembles of neurons, Y (Borst 1999); for use as an objective function for training neural networks (CITE), for feature selection in machine learning, and even as an all-purpose nonlinear measure of "correlation for the 21st century" (Speed.) What is common to all of these new applications, and what differs from the original setting of Shannon's theory of information, is that

Submitted to 29th Conference on Neural Information Processing Systems (NIPS 2016). Do not distribute.

the variables X and Y have unknown distributions which must be inferred from data. In the case when X and Y are both low-dimensional, for instance, when summarizing the properties of a single neuron in response to a single stimulus feature, $I(X;Y)$ can be estimated nonparametrically using a reasonable number of observations. There exists a huge literature on nonparametric estimation of entropy and mutual information exists, see (CITE) for a review.

However, for high-dimensional X and Y the sample complexity grows exponentially with the dimension, making nonparametric approaches intractable in applications with high-dimensional data. One such application includes multivariate pattern analysis (MVPA), an area of neuroscience research pioneered by Haxby (2001), which studies how entire regions of the human brain respond to stimuli, using function magnetic resonance imaging (fMRI) data; in MVPA studies, the input X could be a natural image parameterized by $p = 10000$ image features, while the output Y is a $q = 20000$-dimensional vector of brain activation features obtained from the fMRI scan. In problems of such dimensionality, 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. Rather than assuming a full parametric generative model, one can empirically select a good *discriminative* model by using machine learning. Treves (1997) first proposed using the empirical mutual information of the classification matrix in order to obtain a lower bound of the mutual information $I(X;Y)$; this confusion-matrix-based lower bound has subsequently enjoyed widespread use in the MVPA literature (Quiroga 2009.) But even earlier that this, the idea of linking classification performance to mutual information can be found in the beginnings of information theory: after all, 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.) How do these different classification-based methods for lower bounding mutual information compare, to each other, and to nonparametric and parametric estimators of mutual information? Before discussing such comparisons, we must first delineate a number of assumptions on the sampling regime, and the properties of the classifiers.

1.1 Sampling assumptions

Assume that the variables X, Y have a joint distribution F, and that one can define a conditional distribution of Y given X,

$$Y|X \sim F_X,$$

and let G denote the marginal distribution of X. We consider two different types of sampling procedures:

- *pair sampling*: For $i = 1, \ldots, n$, the data (X^i, Y^i) are sampled i.i.d. from the joint distribution of (X, Y).
- *stratified sampling*: For $j = 1, \ldots, k$, sample i.i.d. *exemplars* $X^{(1)}, \ldots, X^{(k)} \sim G$. For $i = 1, \ldots, n$, draw Z^i iid from the uniform distribution on $1, \ldots, k$, then draw Y^i from the conditional distribution $F_{X^{(Z_i)}}$.

Pair sampling occurs in observational studies, where one observes both X and Y externally. On the other hand, stratified sampling is more commonly seen in controlled experiments, where an experimenter chooses an input X to feed into a black box, which outputs Y. An example from fMRI studies is an experimental design where the subject is presented a stimulus X, and the experimenter measures the subject's response via the brain activation Y.

Mutual information can be defined for discrete or continuous random variables (X, Y), or a combination of discrete input X and continuous output Y and vice-versa. Shannon's original paper (CITE) begins with the case of discrete X and discrete Y, and he considers the problem of decoding X from Y; this is the same problem as labelling a feature vector Y with class labels taking the possible values of X. In the case that X is uniformly distributed on its support, Fano's inequality provides a link between mutual information and classification via

$$I(X;Y) \leq (1 - e_{class}) \log K + ...$$

where e_{class} is the Bayes error and K is the size of the support of X. Since the generalization error of any classifier is greater than the Bayes error, Fano's inequality also holds when e_{class} is taken to mean the generalization error of the classifier. However, the generalization error of any classifier is an unknown parameter: at best, we can obtain upper and lower confidence bounds. If $\bar{e}$ is an α-upper confidence bound, in the sense that

$$\Pr[\bar{e} < e_{gen}] \leq \alpha,$$

then substituting $\bar{e}$ into Fano's inequality yields the lower confidence bound for mutual information,

$$\underline{I}_{Fano} = \log k + ...$$

In the discrete case, there is little consequence to the distinction between pair sampling and stratified sampling as long as the number of sampled classes k is much larger than the support of X. However, in the case of continuous X, the classification tasks must be defined differently depending on the sampling scheme. Under pair sampling, one can no longer take distinct inputs X to define distinct classes, since the notion of generalization error depends on repeated sampling from the same class. Instead, one can define a fixed number classes by specifying a partition on the support of X. For instance, in fMRI imaging experiments, the experimenter may divide a set of stimuli into intuitive categories (car, dog, person, etc.) In contrast, under stratified sampling, one can take the distinct exemplars $X^{(1)}, \ldots, X^{(k)}$ to define distinct classes. While there is no need to specify an arbitrary partition on the input space, the k classes will now be *randomly* defined. One consequence is that the Bayes error e_{Bayes} is a random variable: when the sampling produces k similar exemplars, e_{Bayes} will be higher, and when the sampling produces well-separated exemplars e_{Bayes} may be lower. For this reason, Fano's inequality no longer produces a lower bound–it could produce an overestimate of $I(X;Y)$ for an exceptionally well-separated exemplar set.

Most nonparametric estimators of $I(X;Y)$ are derived under the pair sampling assumption, and may perform badly in the stratified sampling case. On the other hand, there exist nonparametric estimators which are specialized for stratified sampling. Using the fact that

$$I(X;Y) = H(Y) - H(Y|X),$$

one can estimate $I(X;Y)$ by first estimating $H(Y)$ from the empirical marginal distribution of Y, and then estimating $H(Y|X)$ from the distributions within each class:

$$\hat{H}(Y|X) = \frac{1}{k}\sum_{i=1}^{k} \hat{H}(Y|X^{(i)})$$

After Gastpar et al. (2009), we call the resulting estimator $\hat{I}_0$. In their paper, Gastpar et al. showed that $\hat{I}_0$ is biased downwards due to undersampling of the exemplars; to counteract this bias, they introduce the anthropic correction estimator $\hat{I}_\alpha$. If the parameter $\alpha \in [0, 1)$ is chosen correctly, the estimator is unbiased, but no method is given to tune the parameter.

Parametric estimators tend to work similarly in either type of sampling, as long as the sampling is correctly accounted for in the likelihood model. For instance, Gastpar et al. combined their anthropic correction estimator with a gaussian model to estimate information in a high-dimensional dataset.

The most straightforward type of comparison that can be made is between different estimators (or confidence bounds) which use the same type of sampling. But when designing an experiment, a researcher may have a choice between a pair sampling design and a stratified sampling design. The cost of the design may depend simply on the total number of observations n, or there might be an extra cost associated with the number of unique exemplars k; or the opposite could be true–it may cost extra to obtain repeats from the same class. We make an initial stab at the topic of experimental design in our simulation study, with the assumption that the total number of observations n is constrained.

Our primary tool for comparing different estimators (or lower bounds) of mutual information will be through simulation studies, though we will outline some general ideas about the strengths and weaknesses of the three big modeling approaches–nonparametric, parametric, and discriminative–in the discussion.

The main subject of the paper, however, is our proposal of a new lower confidence bound based on classification error. In the following subsection we outline the assumptions and criteria we use in comparing methods *within* the family of classification-based estimators.

1.2 Classification

Formally, a classification rule is any (possibly stochastic) mapping $f : \mathcal{Y} \to \{1, \ldots, k\}$. The *generalization error* of the classification rule for classes $x^{(1)}, \ldots, x^{(k)}$ is

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

A trivial classification rule which outputs the result of a k-sided die roll for all inputs y would achieve a generalization error of $e_{gen} = \frac{k-1}{k}$. Conversely, even a single counterexample with $e_{gen} < \frac{k-1}{k}$ 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

$$H_0 : e_{gen}(f) = \frac{k-1}{k}$$

versus the alternative

$$H_1 : e_{gen}(f) < \frac{k-1}{k}.$$

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.

1.3 Style

Papers to be submitted to NIPS 2016 must be prepared according to the instructions presented here. Papers may only be up to eight pages long, including figures. Since 2009 an additional ninth page *containing only acknowledgments and/or cited references* is allowed. Papers that exceed nine pages will not be reviewed, or in any other way considered for presentation at the conference.

The margins in 2016 are the same as since 2007, which allow for $\sim$15% more words in the paper compared to earlier years.

Authors are required to use the NIPS LaTeX style files obtainable at the NIPS website as indicated below. Please make sure you use the current files and not previous versions. Tweaking the style files may be grounds for rejection.

1.4 Retrieval of style files

The style files for NIPS and other conference information are available on the World Wide Web at

`http://www.nips.cc/`

The file `nips_2016.pdf` contains these instructions and illustrates the various formatting requirements your NIPS paper must satisfy.

The only supported style file for NIPS 2016 is `nips_2016.sty`, rewritten for LaTeX 2_ε. **Previous style files for LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

The new LaTeX style file contains two optional arguments: `final`, which creates a camera-ready copy, and `nonatbib`, which will not load the `natbib` package for you in case of package clash.

At submission time, please omit the `final` option. This will anonymize your submission and add line numbers to aid review. Please do *not* refer to these line numbers in your paper as they will be removed during generation of camera-ready copies.

The file `nips_2016.tex` may be used as a "shell" for writing your paper. All you have to do is replace the author, title, abstract, and text of the paper with your own.

The formatting instructions contained in these style files are summarized in Sections 2, 3, and 4 below.

2 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

For the final version, authors' names are set in boldface, and each name is centered above the corresponding address. The lead author's name is to be listed first (left-most), and the co-authors' names (if different address) are set to follow. If there is only one co-author, list both author and co-author side by side.

Please pay special attention to the instructions in Section 4 regarding figures, tables, acknowledgments, and references.

3 Headings: first level

All headings should be lower case (except for first word and proper nouns), flush left, and bold.

First-level headings should be in 12-point type.

3.1 Headings: second level

Second-level headings should be in 10-point type.

3.1.1 Headings: third level

Third-level headings should be in 10-point type.

Paragraphs There is also a `\paragraph` command available, which sets the heading in bold, flush left, and inline with the text, with the heading followed by 1 em of space.

4 Citations, figures, tables, references

These instructions apply to everyone.

4.1 Citations within the text

The `natbib` package will be loaded for you by default. Citations may be author/year or numeric, as long as you maintain internal consistency. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

The documentation for `natbib` may be found at

`http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

Of note is the command `\citet`, which produces citations appropriate for use in inline text. For example,

```
\citet{hasselmo} investigated\dots
```

produces

Hasselmo, et al. (1995) investigated...

If you wish to load the `natbib` package with options, you may add the following before loading the `nips_2016` package:

```
\PassOptionsToPackage{options}{natbib}
```

If `natbib` clashes with another package you load, you can add the optional argument `nonatbib` when loading the style file:

```
\usepackage[nonatbib]{nips_2016}
```

As submission is double blind, refer to your own published work in the third person. That is, use "In the previous work of Jones et al. [4]," not "In our previous work [4]." If you cite your other papers that are not widely available (e.g., a journal paper under review), use anonymous author names in the citation, e.g., an author of the form "A. Anonymous."

4.2 Footnotes

Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number[1] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).

Note that footnotes are properly typeset *after* punctuation marks.[2]

4.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction. The figure number and caption always appear after the figure. Place one line space before the figure caption and one line space after the figure. The figure caption should be lower case (except for first word and proper nouns); figures are numbered consecutively.

You may use color figures. However, it is best for the figure captions and the paper body to be legible if the paper is printed in either black/white or in color.

4.4 Tables

All tables must be centered, neat, clean and legible. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

[1] Sample of the first footnote.
[2] As in this example.

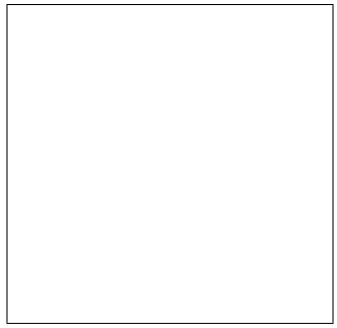

Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	$\sim$100
Axon	Output terminal	$\sim$10
Soma	Cell body	up to 10^6

Note that publication-quality tables *do not contain vertical rules.* We strongly suggest the use of the `booktabs` package, which allows for typesetting high-quality, professional tables:

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

5 Final instructions

Do not change any aspects of the formatting parameters in the style files. In particular, do not modify the width or length of the rectangle the text should fit into, and do not change font sizes (except perhaps in the **References** section; see below). Please note that pages should be numbered.

6 Preparing PDF files

Please prepare submission files with paper size "US Letter," and not, for example, "A4."

Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or Embedded TrueType fonts. Here are a few instructions to achieve this.

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

  followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for $\mathbb{R}$, $\mathbb{N}$ or $\mathbb{C}$. You can also use the following workaround for reals, natural and complex:

```
\newcommand{\RR}{I\!\!R} %real numbers
\newcommand{\Nat}{I\!\!N} %natural numbers
\newcommand{\CC}{I\!\!\!\!C} %complex numbers
```

Note that `amsfonts` is automatically loaded by the `amssymb` package.

If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

6.1 Margins in LaTeX

Most of the margin problems come from figures positioned by hand using `\special` or other commands. We suggest using the command `\includegraphics` from the `graphicx` package. Always specify the figure width as a multiple of the line width as in the example below:

```
\usepackage[pdftex]{graphicx} ...
\includegraphics[width=0.8\linewidth]{myfile.pdf}
```

See Section 4.4 in the graphics bundle documentation (`http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf`)

A number of width problems arise when LaTeX cannot properly hyphenate a line. Please give LaTeX hyphenation hints using the `\-` command when necessary.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments go at the end of the paper. Do not include acknowledgments in the anonymized submission, only in the final paper.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. **Remember that you can use a ninth page as long as it contains *only* cited references.**

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.