A Finite-Sample, Distribution-Free, Probabilistic Lower Bound on Mutual Information

Nathan D. VanderKraats and Arunava Banerjee

{*ndv,arunava*}*@cise.ufl.edu*

Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA 32611

Abstract

For any memoryless communication channel with a binary-valued input and a one-dimensional real-valued output, we introduce a probabilistic lower bound on the mutual information given empirical observations on the channel. The bound is built upon the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality and is distribution-free. A quadratic time algorithm is described for computing the bound and its corresponding class-conditional distribution functions. We compare our approach to existing techniques and show the superiority of our bound to a method inspired by Fano's inequality where the continuous random variable is discretized.

1 Introduction

Determining the mutual information over a communication channel is a ubiquitous problem in many fields. In neuroscience, feedforward communication channels are seen extensively in the analysis of sensory systems, where neural responses are analyzed in an effort to determine the class of the generating stimulus. Our particular case – a discrete input random variable and a continuous output random variable – is encountered directly whenever a response spike train is described as a continuous quantity. Perhaps the most common continuous response representation is the time-averaged spike rate, which has been used to differentiate stimuli in the cat lemniscal auditory thalamus and cortex (L. M. Miller et al., 2002), the macaque middle temporal cortex (Buračas et al., 1998), the macaque motor cortex (Georgopoulos et al., 1982), and the macaque ventrolateral prefrontal cortex (Lee et al., 2009), to name a few. As another example, the latency between the stimulus and a response neuron's first spike has been shown to carry a large amount of information about some stimuli (Gollisch & Meister, 2008; Gerstner & Kistler, 2002). The effectiveness of various scalar representations of single-neuron responses has often been investigated, for instance in locations throughout the cat auditory pathway (Middlebrooks et al., 1994; Chechik et al., 2006; Nelkin & Chechik, 2005).

Even when a neural response is high-dimensional, such as the general case of an ensemble of neurons with an unbounded number of real-valued spike times, its dimensionality is often reduced during the course of processing. For instance, continuous stimuli are often reconstructed from the response (Bialek et al., 1991; Warland et al., 1997). Viewed as a transformation, reconstruction amounts to mapping a high-dimensional vector to a scalar value. When the stimuli are discrete, reconstruction is essentially a classification problem (Duda et al., 2001). Several classifiers, commonly called decoders in this context, have been applied, including linear discriminant analysis (Nicolelis et al., 1998), support vector machines (Mesgarani et al., 2008), and maximum likelihood (Panzeri et al., 1999; Samengo, 2002). Additionally, unsupervised dimensionality-reduction techniques have been used to preprocess ensembles of spike trains, such as principle component analysis (Richmond et al., 1987) and the wavelet-based method of Laubach (2004).

Our type of channel is also prevalent outside of the stimulus-response domain. Many epilepsy studies, for example, observe time-varying EEG or SEEG data to decode the binary states of seizure detection (Quiroga et al., 2000) or prediction (Elger & Lehnertz, 1998). In the field of computer vision, the image segmentation problem has been re-posed as a maximization of the information between an image's pixel intensities and the discrete region labels (Kim et al., 2005).

1.1 Estimating Information from Finite Samples

Mutual information (MI) (Shannon, 1948; Cover & Thomas, 2006) is a natural quantification of how much the output of a channel depends on its input (Rieke et al., 1997; Borst & Theunissen, 1999). Over a physical channel, MI must be estimated based on a finite number of empirical observations. Therefore, the utility of an MI estimate must also be qualified in terms of a probabilistic error term, or confidence interval.

In the majority of neuroscience literature, the MI estimation problem begins by assuming a high-dimensional continuous response – an ensemble of spike trains – that is discretized through a form of binning (Rieke et al., 1997). For an excellent discussion of current entropy estimation methods in neuroscience, see the reviews by Panzeri et al. (2007) and Victor (2006). Many of these techniques are based on the naïve MI estimator obtained by using the empirical joint distributions directly as approximations of the true underlying distributions (Strong et al., 1998); this approach is also termed the maximum-likelihood or plug-in estimator. The naïve method is known to be inherently biased, even in the asymptotic regime where the amount of data is unlimited (G. Miller, 1955). A popular strategy for estimating information for a finite number of observations is to start with the naïve estimator and attempt to accurately correct for this bias. Of particular interest is the estimator of Paninski (2003), which utilizes a probabilistic lower error bound on the MI derived from McDiarmid's Inequality (McDiarmid, 1989). Another recent effort uses an anthropic correction to the naïve estimator to give a positively-biased MI estimate (Gastpar et al., 2010).

For the finite-sample regime, other information estimates that focus on reducing the bias of the naïve estimator have been obtained through assumptions about the underlying response distributions. These include model selection methods (Montemurro et al., 2007; Shlens et al., 2007; Kennel et al., 2005) and Bayesian inference methods (Nemenman et al., 2004; Wolpert & Wolf, 1995). One unique strategy by Victor (2002) provides an asymptotically-unbiased information estimator without binning.

In this article, we explore a fundamentally different approach to the MI estimation problem. For any finite sample, the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956) gives a probabilistic bound on the difference between the empirical cumulative distribution function and the true distribution function. The result is an extension to the finite-sample regime of the well-known work of Kolmogorov

(1933) and Smirnov (1944) who proved it for the asymptotic regime (Shorack & Wellner, 1986; Doob, 1949). The DKW inequality was further refined by Massart (1990), who showed that his version was tight. The DKW inequality, then, provides tight probabilistic bounds around any empirical distribution, such as the distributions involved in the direct calculation of MI. This result allows us to derive a probabilistic lower bound on MI that is distribution-free in the finite-sample regime. A similar application of the DKW inequality was recently used to address the problem of maximizing differential entropy on a single variable (Learned-Miller & DeStefano, 2008).

It is important to note that, by the Glivenko-Cantelli theorem (Glivenko, 1933; Cantelli, 1933), the empirical cumulative distribution functions in question converge uniformly to the true distribution functions almost surely (i.e., with probability 1) in the limiting case where the sample size grows to infinity. Our probabilistic lower bound on MI, therefore, approaches the true MI as the number of observations approaches infinity. For a fixed confidence level, the DKW inequality quantifies the rate of this convergence with respect to the number of samples, yielding optimal probabilistic bounds on the underlying distribution functions for any given sample size.

We construct a probabilistic lower bound on the MI over any memoryless communication channel with binary input and one-dimensional continuous output. Additionally, we develop a worst-case quadratic time algorithm to efficiently compute this bound for any data set.

2 A First Approach: Post-Discretization Analysis

Consider the binary-valued random variable X and real-valued random variable Y, denoting the input and output, respectively, of some feedforward communication channel. Since the distribution of X is controlled by the experimenter, we assume the marginal probabilities $P(X = 0)$ and $P(X = 1)$ to be fixed *a priori* such that $P(X = 0) =$

$P(X = 1) = 0.5$. We seek a lower bound on the MI:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= 1 - H(X|Y) \\ &= 1 + \int_{Y} \left[\sum_{X\in\{0,1\}} P(X|Y=y) \log_2 P(X|Y=y) \right] f(y)\,\mathrm{d}y \end{aligned} \tag{1}$$

where $H(\cdot)$ denotes the Shannon entropy and $f(y)$ is the density of the continuous random variable Y. In spite of our trivialized input, estimating the continuous function I based solely on the observation of a finite sample is no easy task. The standard approach to this problem is to discretize Y, leading to a lower bound on MI through the data processing inequality (Cover & Thomas, 2006):

$$I(X;Y) \geq I(X;T(Y))$$

for any function $T(Y)$. In what follows, we let T_m denote some function that discretizes Y into m bins.

2.1 The Fano Method

The simplest discretization, T_2, partitions the output into two bins. While seemingly rudimentary, this function is interesting because it produces a classification space on Y: in essence, each element of the output is mapped to a prediction of the input that produced it. For notational convenience we denote the new binary output random variable, which is the value of the function T_2, also as Y. Motivated by Fano's inequality (Cover & Thomas, 2006) we can derive a lower bound on MI using the binary input and the per-class error rates on the binary Y. For binary random variables, the conditional entropy of X is equivalent to the conditional entropy of the error, $H(E|Y)$. Therefore, our discretized estimate of the MI can be expressed as:

$$I(X;Y) = 1 - H(E|Y) \tag{2}$$

which can easily be rewritten as a function depending only on the true class-conditional probabilities of error, $P(E=1|X=0) \equiv P(Y=1|X=0)$ and $P(E=1|X=1) \equiv P(Y=0|X=1)$.

Now then, given a sample of $\frac{n}{2}$ independent, identically distributed (i.i.d.) random variables per class, Hoeffding's inequality (Hoeffding, 1963) states probabilistic bounds on each true class-conditional test error:

$$P\left(\left|\frac{2S_0}{n} - P(E=1|X=0)\right| \geq \epsilon_h\right) \leq 2e^{-n\epsilon_h^2} \quad \text{and independently}$$
$$P\left(\left|\frac{2S_1}{n} - P(E=1|X=1)\right| \geq \epsilon_h\right) \leq 2e^{-n\epsilon_h^2}$$

where for each class $x \in \{0,1\}$, S_x denotes the total number of errors and the true error $P(E=1|X=x)$ lies within $\pm\epsilon_h$ of the empirical error with confidence $\gamma = 1-2e^{-n\epsilon_h^2}$. For any desired confidence, therefore:

$$\epsilon_h = \sqrt{-\frac{\ln\frac{1-\gamma}{2}}{n}} \tag{3}$$

A probabilistic lower bound on MI can be calculated by considering the worst case: that the true class-conditional error rates are ϵ_h larger than observed.

2.2 Generalized Discretization

The previous technique relies on the results of Hoeffding and Fano to achieve a tight lower bound on MI given any binary discretization of the output. Unfortunately, generalizing this strategy to arbitrary discretizations T_m is not as simple.

An m-bin discretization can be approached in one of two ways. First, in direct analogy to the Fano-inspired method, we could apply a probabilistic bound on the m-dimensional multinomial distribution. For example, Takeuchi (1993) has derived probabilistic bounds for the Kullback-Leibler divergence between the true and the empirical multinomial distribution. Such a bound would identify a set on the m-simplex around the empirical distribution for which the probability is bounded to a given confidence, as before. (See also Globerson et al. (2009), where this set is defined as those multinomial distributions for which the expectation of fixed functions are given.) However, minimizing the MI function over this set – a straightforward exercise when I could be expressed in terms of only error rates – is now complicated due to the combinatorics of optimization over m variables.

Another avenue for the m-dimensional discretization was thoroughly explored by Paninski (2003). In this work, a distribution-free bound is created around the bias-corrected empirical estimate of MI using McDiarmid's inequality (McDiarmid, 1989).

Although ultimately his MI estimate is useful, the approach faces difficulties stemming from the weakness of McDiarmid's inequality, due to its generality. To avoid these discretization issues, we derive an analytical solution for the case when Y is real-valued. This view is equivalent to the multinomial approach above as $m \to \infty$.

3 A Novel Lower Bound on MI

Consider a communication channel with the input random variable, X, taking *discrete* (binary) values, i.e., $X \in \{0, 1\}$, and the output random variable, Y, taking *real* values in a *bounded* range, i.e., $Y \in [a, b] : a, b \in \mathbb{R}$. The channel is modeled by a pair of unknown (conditional) *continuous* distribution functions $P(Y \leq y|X = 0)$ and $P(Y \leq y|X = 1)$ with density functions f_0 and f_1 such that:

$$F_0(y) \triangleq \int_a^y f_0(t)\,\mathrm{d}t \triangleq P(Y \leq y|X = 0) \quad \text{and}$$
$$F_1(y) \triangleq \int_a^y f_1(t)\,\mathrm{d}t \triangleq P(Y \leq y|X = 1)$$

As before, we assume $P(X = 0) = P(X = 1) = 0.5$. However, our results are easily generalized for other values.

Let $y_1^0, y_2^0, \ldots, y_{(\frac{n}{2})}^0$ be a sample of $\frac{n}{2}$ i.i.d. random variables with distribution function $F_0(y)$, and $y_1^1, y_2^1, \ldots, y_{(\frac{n}{2})}^1$ be a sample of $\frac{n}{2}$ i.i.d. random variables with distribution function $F_1(y)$. Also, let $\widehat{F}_0(y)$, $\widehat{F}_1(y)$ be the empirical distribution functions, defined by:

$$\widehat{F}_0(y) = \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} \mathbb{1}_{(y_i^0 \leq y)} \quad \text{and} \quad \widehat{F}_1(y) = \frac{2}{n} \sum_{i=1}^{\frac{n}{2}} \mathbb{1}_{(y_i^1 \leq y)}$$

where $\mathbb{1}_E$ represents the indicator function for the event E, and the traditional subscript denoting the sample size for $\widehat{F}$ is understood.

In what follows, we utilize the order statistics[1] of the combined sample $\mathbf{y^0} \cup \mathbf{y^1}$, denoted by $\langle z_i | i = 1 \ldots n \rangle$. For notational convenience, we define the points $z_0 \triangleq a$ and $z_{n+1} \triangleq b$, so that $F_0(z_0) = F_1(z_0) = 0$ and $F_0(z_{n+1}) = F_1(z_{n+1}) = 1$.

[1] The order statistics $z_1, z_2, \ldots, z_n$ of a sample $y_1, y_2, \ldots y_n$ are the values of the sample arranged in non-decreasing order.

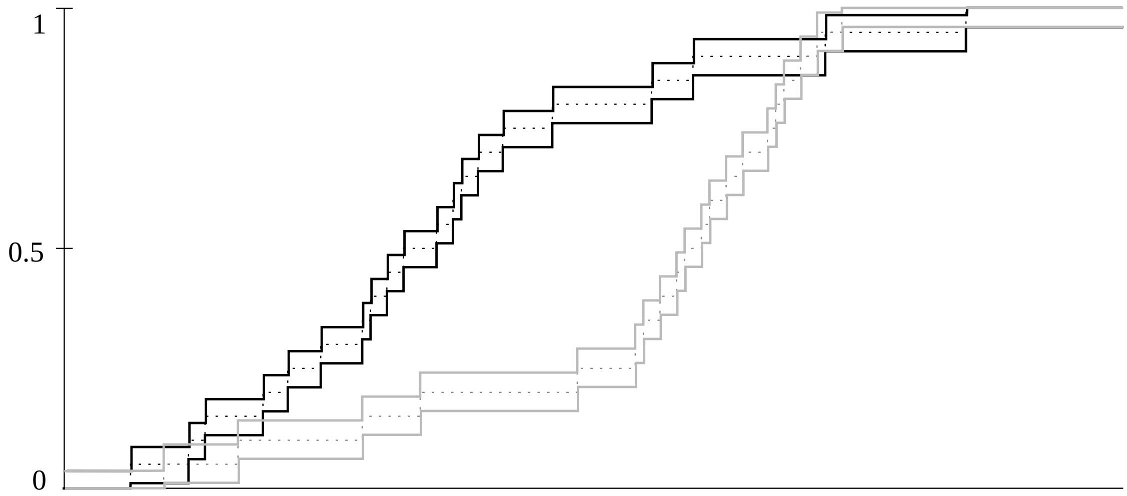


Figure 1: Two empirical class-conditional distributions and their DKW tubes. The dotted lines represent the empirical distributions and the solid lines depict the upper and lower tubes.

The DKW inequality, described previously, places a tight probabilistic bound on the difference between each empirical distribution function and its true distribution function (Dvoretzky et al., 1956; Massart, 1990). Using samples with $\frac{n}{2}$ i.i.d. random variables for each distribution, the bounds on F_0 and F_1 are given by:

$$P\left(\sqrt{\frac{n}{2}}\sup_t |\widehat{F}_0(t) - F_0(t)| > \delta\right) \leq 2e^{-2\delta^2} \quad \text{and independently}$$
$$P\left(\sqrt{\frac{n}{2}}\sup_t |\widehat{F}_1(t) - F_1(t)| > \delta\right) \leq 2e^{-2\delta^2}$$

Therefore, given any desired confidence γ, the DKW inequality guarantees that the true distributions will lie within the fixed tube drawn $\pm\epsilon_{dkw}$ around the empirical distributions, where:

$$\epsilon_{dkw} = \sqrt{-\frac{\ln\frac{1-\gamma}{2}}{n}} \tag{4}$$

Within this framework, we seek two distribution functions F_0^* and F_1^* on $[z_0, z_{n+1}]$ that minimize the mutual information $I(X;Y)$ subject to the DKW tube constraints. Since $P(X=0) = P(X=1) = 0.5$, the entropy $H(X) = 1$, and:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= 1 + \frac{1}{2}\int_{z_0}^{z_{n+1}} \left[f_0(t)\log\frac{f_0(t)}{f_0(t)+f_1(t)} + f_1(t)\log\frac{f_1(t)}{f_0(t)+f_1(t)}\right] \mathrm{d}t \\ &\triangleq 1 + \frac{1}{2}M_{\int} \end{aligned} \tag{5}$$

We focus hereafter on the variable component $M_{\int}$.

3.1 The Tube-Unconstrained Solution

Before undertaking the general problem, we address a simpler subproblem:

Theorem 1. *Consider any pair of values $c : c \geq a$ and $d : d \leq b$ such that the solutions for the distribution functions F_0 and F_1 are known at $F_0^*(c)$, $F_0^*(d)$, $F_1^*(c)$, and $F_1^*(d)$ (we call these four points "pins", and the corresponding interval $[c, d]$ a "pinned interval"). Assuming that F_0^* and F_1^* are monotonically non-decreasing on $[c, d]$, then, in the absence of further constraints, the solution that minimizes the interval's contribution to the total MI is given by any two curves with the property:*

$$f_0(t) = v \cdot f_1(t) \quad \forall t : c \leq t \leq d \tag{6}$$

for some $v \in \mathbb{R}$. In other words, the two solution densities are multiples of one another on the interval $[c, d]$. Furthermore, the minimum contribution to $I(X, Y)$ by this interval is:

$$m = \alpha \log \frac{\alpha}{\alpha + \beta} + \beta \log \frac{\beta}{\alpha + \beta} \tag{7}$$

where $\alpha = F_0^*(d) - F_0^*(c)$ and $\beta = F_1^*(d) - F_1^*(c)$.

Proof of Theorem 1. As stated, let $\alpha = F_0^*(d) - F_0^*(c)$, denoting the increase in F_0 between its two pins, and $\beta = F_1^*(d) - F_1^*(c)$, denoting F_1's increase between its pins. The minimal MI on $[c, d]$ is given by the curves that minimize the functional:

$$M_{\int_c^d} = \int_c^d \left[f_0(t) \log \frac{f_0(t)}{f_0(t) + f_1(t)} + f_1(t) \log \frac{f_1(t)}{f_0(t) + f_1(t)} \right] dt$$

subject only to the constraints:

$$\begin{aligned} \int_c^d f_0(t)\, dt &= \alpha \\ \int_c^d f_1(t)\, dt &= \beta \end{aligned} \tag{8}$$

$$f_0(t) \geq 0 \qquad\qquad f_1(t) \geq 0 \tag{9}$$

$$\forall t : c \leq t \leq d$$

Using the objective and these constraints, the Lagrangian integrand function for the resulting calculus of variations problem is:

$$\begin{gathered} f_0(t) \log \frac{f_0(t)}{f_0(t) + f_1(t)} + f_1(t) \log \frac{f_1(t)}{f_0(t) + f_1(t)} \\ +\nu_1 f_0(t) + \nu_2 f_1(t) - \lambda_1(t) f_0(t) - \lambda_2(t) f_1(t) \end{gathered} \tag{10}$$

where ν_1 and ν_2 are constants and $\lambda_1(t),\ \lambda_2(t) \geq 0.$ The Hessian of the integrand of $M_{\int_c^d}$ is:

$$H(M) = \begin{bmatrix} \frac{f_1(t)}{f_0(t)(f_0(t)+f_1(t))} & \frac{-1}{f_0(t)+f_1(t)} \\ \frac{-1}{f_0(t)+f_1(t)} & \frac{f_0(t)}{f_1(t)(f_0(t)+f_1(t))} \end{bmatrix}$$

which is positive semi-definite. Therefore the integrand is convex, and consequently the functional $M_{\int_c^d}$ is also convex. Since all the constraints are affine, any extremals of the Lagrangian must be minima. Two Euler-Lagrange equations are:

$$0 = \log \frac{f_0(t)}{f_0(t) + f_1(t)} + \nu_1 - \lambda_1(t) \tag{11}$$

$$0 = \log \frac{f_1(t)}{f_0(t) + f_1(t)} + \nu_2 - \lambda_2(t) \tag{12}$$

$$\forall t : c \leq t \leq d$$

Complementary slackness requires that:

$$\lambda_1(t) f_0(t) = 0 = \lambda_2(t) f_1(t)$$

$$\forall t : c \leq t \leq d$$

Now for any t, if $f_0(t) = 0$, then the right-hand side of Equation 11 must be non-finite, and therefore nonzero. Similarly, $f_1(t)$ must be nonzero in order to satisfy Equation 12. Therefore, $\lambda_1(t) = \lambda_2(t) = 0$ for all t. Rewriting Equations 11, 12 shows that:

$$v \triangleq \frac{f_0(t)}{f_1(t)} = \frac{1}{2^{\nu_1} - 1} = 2^{\nu_2} - 1 \quad \forall t : c \leq t \leq d$$

So then, $v \in \mathbb{R}$ is a constant that does not depend on t, and f_0 and f_1 differ by a constant multiple on the pinned interval. Since ν_1 and ν_2 have no other constraints, it is also clear that any two densities differing by a constant multiple will be equivalent minimizers.

Furthermore, if the multiplier v is such that $v \cdot f_1(t) = f_0(t)$, then it is easy to show that $\alpha = v\beta$, and therefore:

$$\frac{\alpha}{\alpha + \beta} = \frac{v \cdot \int_c^d f_1(t)\,dt}{v \cdot \int_c^d f_1(t)\,dt + \int_c^d f_1(t)\,dt} = \frac{v}{v+1} = \frac{f_0(t)}{f_0(t) + f_1(t)} \quad \forall t : c \leq t \leq d$$

$$\frac{\beta}{\alpha+\beta} = \frac{\int_c^d f_0(t)\,dt}{\int_c^d f_0(t)\,dt + v \cdot \int_c^d f_0(t)\,dt} = \frac{1}{v+1} = \frac{f_1(t)}{f_0(t)+f_1(t)} \quad \forall t : c \leq t \leq d$$

Therefore, the minimum value on the pinned interval is:

$$\begin{aligned} m &= min_{\{f_0,f_1\}} \int_c^d \left[f_0(t) \log \frac{f_0(t)}{f_0(t)+f_1(t)} + f_1(t) \log \frac{f_1(t)}{f_0(t)+f_1(t)} \right] dt \\ &= \alpha \log \frac{\alpha}{\alpha+\beta} + \beta \log \frac{\beta}{\alpha+\beta} \end{aligned}$$

□

For reasons that will soon be made clear, any solution meeting the requirements of Theorem 1 will be known as a *tube-unconstrained solution*.

3.2 A Discrete Formulation

At any point z_i, we denote the (unknown) conditional probability mass for each $X \in \{0, 1\}$ by:

$$\begin{aligned} f_0^i &\triangleq \int_{z_{i-1}}^{z_i} f_0(t)\,\mathrm{d}t = F_0(z_i) - F_0(z_{i-1}) \quad \text{and} \\ f_1^i &\triangleq \int_{z_{i-1}}^{z_i} f_1(t)\,\mathrm{d}t = F_1(z_i) - F_1(z_{i-1}) \end{aligned}$$

Also, we denote the lower DKW tube boundaries at z_i for the two distribution functions as $\overline{F_0^-}(z_i)$ and $\overline{F_1^-}(z_i)$, and the upper tube boundaries correspondingly as $\overline{F_0^+}(z_i)$ and $\overline{F_1^+}(z_i)$. In order for the distribution functions to be feasible throughout the entire interior of the tubes, the F^- boundaries are raised to 0 whenever they fall below 0, and the F^+ boundaries are lowered to 1 whenever they rise above 1. The tubes have nonzero width at the minimum and maximum order statistics, z_1 and z_n, and are defined as being collapsed at the interval edges such that:

$$\begin{aligned} \overline{F_0^-}(z_0) = \overline{F_0^+}(z_0) = \overline{F_1^-}(z_0) = \overline{F_1^+}(z_0) = 0 \qquad \text{and} \\ \overline{F_0^-}(z_{n+1}) = \overline{F_0^+}(z_{n+1}) = \overline{F_1^-}(z_{n+1}) = \overline{F_1^+}(z_{n+1}) = 1 \end{aligned}$$

Using Theorem 1, the problem of minimizing MI with respect to two functions can be simplified to a constrained optimization problem with a finite number of constraints. Since the empirical distributions are each defined using a sample of $\frac{n}{2}$ i.i.d. random variables, the DKW tubes will be step functions on $[z_0, z_{n+1}]$. By definition, the tubes

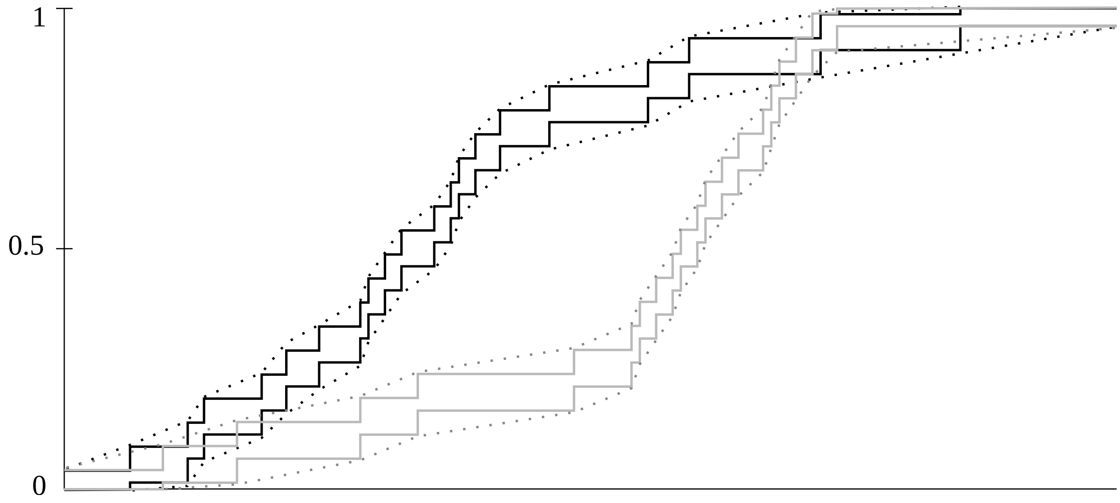


Figure 2: Relaxation of the DKW tubes. The solid lines denote DKW tubes computed from two empirical distributions. The dotted lines show the piecewise linear versions of the tubes.

are flat between any two successive order statistics, z_i and z_{i+1}, without loss of generality. However, we consider a weakened version of the DKW tubes such that successive tube boundaries are piecewise linear, as shown in Figure 2. Formally:

$$\overline{F_0^+}(t) = c_0(t - z_i) + \overline{F_0^+}(z_i) \qquad \text{and} \qquad \overline{F_0^-}(t) = c_0(t - z_i) + \overline{F_0^-}(z_{i-1})$$
$$\overline{F_1^+}(t) = c_1(t - z_i) + \overline{F_1^+}(z_i) \qquad \text{and} \qquad \overline{F_1^-}(t) = c_1(t - z_i) + \overline{F_1^-}(z_{i-1})$$
$$\forall t : z_i \leq t \leq z_{i+1}$$

where

$$c_0 = \frac{\overline{F_0^+}(z_{i+1}) - \overline{F_0^+}(z_i)}{z_{i+1} - z_i} \qquad \text{and} \qquad c_1 = \frac{\overline{F_1^+}(z_{i+1}) - \overline{F_1^+}(z_i)}{z_{i+1} - z_i}$$

Now then, consider any solution to the general problem. The distribution functions will take values within their tubes at $F_0^*(z_i)$, $F_0^*(z_{i+1})$, $F_1^*(z_i)$, and $F_1^*(z_{i+1})$. On the interval $[z_i, z_{i+1}]$, consider the linear solution obtained by drawing one straight line between $F_0^*(z_i)$ and $F_0^*(z_{i+1})$, and another between $F_1^*(z_i)$ and $F_1^*(z_{i+1})$. These lines clearly lie within the relaxed DKW tubes. Furthermore, since they are linear, this solution necessarily has the property that f_0 and f_1 are multiples on $[z_i, z_{i+1}]$.

Applying Theorem 1, the general problem may be simplified by placing pins at all order statistics, yielding a system of $2n$ variables in lieu of two functions. The functional $I(X;Y)$ can thus be written discretely:

$$\begin{aligned} I(X;Y) &= 1 + \frac{1}{2}\sum_{i=1}^{n+1} f_0^i \log \frac{f_0^i}{f_0^i + f_1^i} + \frac{1}{2}\sum_{i=1}^{n+1} f_1^i \log \frac{f_1^i}{f_0^i + f_1^i} \\ &= 1 + \frac{1}{2} M(f_0^1, f_0^2, \ldots, f_0^n, f_0^{n+1}, f_1^1, f_1^2, \ldots, f_1^n, f_1^{n+1}) \end{aligned} \tag{13}$$

where I is equivalently minimized by the function M, subject to a host of constraints. Finding the minimum of M can now be cleanly posed as a constrained optimization problem (Boyd & Vandenberghe, 2004).

3.3 Constraints on the Distribution Functions

For any statistic $z_i : 1 \leq i \leq n$, four constraints are imposed by the DKW tubes of F_0 and F_1, and two more ensure the non-negativity of f_0 and f_1:

$$\begin{aligned} g_i^1 &= \sum_{j=1}^{i} f_0^j - \overline{F_0^+}(z_i) \leq 0 & g_i^2 &= \overline{F_0^-}(z_i) - \sum_{j=1}^{i} f_0^j \leq 0 \\ g_i^3 &= \sum_{j=1}^{i} f_1^j - \overline{F_1^+}(z_i) \leq 0 & g_i^4 &= \overline{F_1^-}(z_i) - \sum_{j=1}^{i} f_1^j \leq 0 \\ g_i^5 &= -f_0^i \leq 0 & g_i^6 &= -f_1^i \leq 0 \end{aligned} \tag{14}$$

Two more constraints necessitate that the total probability under each curve must sum to 1:

$$h_0 = \sum_{j=1}^{n+1} f_0^j - 1 = 0 \qquad h_1 = \sum_{j=1}^{n+1} f_1^j - 1 = 0 \tag{15}$$

Note that the subscripts for the inequality constraints do not include the point z_{n+1}, since these conditions would be redundant.

The Lagrangian for this optimization problem is therefore:

$$L = M + \sum_{i=1}^{n} \left[\lambda_i^1 g_i^1 + \lambda_i^2 g_i^2 + \lambda_i^3 g_i^3 + \lambda_i^4 g_i^4 + \lambda_i^5 g_i^5 + \lambda_i^6 g_i^6\right] + \nu_\alpha h_0 + \nu_\beta h_1 \tag{16}$$

With respect to the arguments of the objective function M, the constraints g_i^1, g_i^2, g_i^3, g_i^4, g_i^5, g_i^6, h_0, and h_1 are both linear and continuously-differentiable. Without loss of generality, the Hessian of M at any z_i is:

$$H(M) = \begin{bmatrix} \frac{f_1^i}{f_0^i(f_0^i+f_1^i)} & \frac{-1}{f_0^i+f_1^i} \\ \frac{-1}{f_0^i+f_1^i} & \frac{f_0^i}{f_1^i(f_0^i+f_1^i)} \end{bmatrix}$$

which is positive semi-definite. Since all constraints are affine and the problem is strictly feasible, the Karush-Kuhn-Tucker (KKT) conditions are both necessary and sufficient (Boyd & Vandenberghe, 2004).

3.4 KKT Conditions for the General Problem

The first KKT condition specifies that the gradient of the Lagrangian must be zero:

$$\nabla L = 0 \tag{17}$$

where

$$\begin{aligned}
\frac{\partial L}{\partial f_0^1} &= \log \frac{f_0^1}{f_0^1 + f_1^1} + \nu_\alpha - \lambda_1^5 + \lambda_1^1 + \lambda_2^1 + \cdots + \lambda_n^1 - \lambda_1^2 - \lambda_2^2 - \cdots - \lambda_n^2 &&= 0 \\
\frac{\partial L}{\partial f_1^1} &= \log \frac{f_1^1}{f_0^1 + f_1^1} + \nu_\beta - \lambda_1^6 + \lambda_1^3 + \lambda_2^3 + \cdots + \lambda_n^3 - \lambda_1^4 - \lambda_2^4 - \cdots - \lambda_n^4 &&= 0 \\
\frac{\partial L}{\partial f_0^2} &= \log \frac{f_0^2}{f_0^2 + f_1^2} + \nu_\alpha - \lambda_2^5 + \lambda_2^1 + \lambda_3^1 + \cdots + \lambda_n^1 - \lambda_2^2 - \lambda_3^2 - \cdots - \lambda_n^2 &&= 0 \\
\frac{\partial L}{\partial f_1^2} &= \log \frac{f_1^2}{f_0^2 + f_1^2} + \nu_\beta - \lambda_2^6 + \lambda_2^3 + \lambda_3^3 + \cdots + \lambda_n^3 - \lambda_2^4 - \lambda_3^4 - \cdots - \lambda_n^4 &&= 0 \\
&\vdots \\
\frac{\partial L}{\partial f_0^n} &= \log \frac{f_0^n}{f_0^n + f_1^n} + \nu_\alpha - \lambda_n^5 + \lambda_n^1 - \lambda_n^2 &&= 0 \\
\frac{\partial L}{\partial f_1^n} &= \log \frac{f_1^n}{f_0^n + f_1^n} + \nu_\beta - \lambda_n^6 + \lambda_n^3 - \lambda_n^4 &&= 0 \\
\frac{\partial L}{\partial f_0^{n+1}} &= \log \frac{f_0^{n+1}}{f_0^{n+1} + f_1^{n+1}} + \nu_\alpha &&= 0 \\
\frac{\partial L}{\partial f_1^{n+1}} &= \log \frac{f_1^{n+1}}{f_0^{n+1} + f_1^{n+1}} + \nu_\beta &&= 0
\end{aligned}$$

Notably, with regard to the tube-related constraints λ_i^1, λ_i^2, λ_i^3, and λ_i^4, the terms of successive partials are upper triangular when viewing the order statistics in increasing order.

The remaining KKT conditions are primal feasibility, dual feasibility, and complementary slackness. Primal feasibility requires the satisfaction of all constraints:

$$\begin{aligned}
&h_0 = h_1 = 0 \\
&g_i^k \leq 0 \qquad \forall k = 1 \ldots 6, \quad \forall i = 1 \ldots n
\end{aligned}$$

Dual feasibility enforces positivity on all of the λ multipliers:

$$\lambda_i^k \geq 0 \qquad \forall k = 1 \ldots 6, \quad \forall i = 1 \ldots n$$

Complementary slackness dictates that:

$$\lambda_i^k g_i^k = 0 \qquad \forall k = 1 \ldots 6, \quad \forall i = 1 \ldots n \tag{18}$$

As in the proof of Theorem 1, the complementary slackness criteria can be used to determine the monotonicity constraints λ_i^5 and λ_i^6. For any i without loss of generality, if λ_i^5 is nonzero, then $f_0^i = 0$ by Equation 18. This implies that $\frac{\partial L}{\partial f_0^i}$ is non-finite, contradicting Equation 17. A similar argument can be made for λ_i^6 and f_1^i. Therefore:

$$\lambda_i^5 = \lambda_i^6 = 0 \quad \forall i = 1 \ldots n$$

It is also important to note the symmetries between λ_i^1 and λ_i^2, and λ_i^3 and λ_i^4. When λ_i^1 is nonzero, the curve $F_0^*(z_i)$ is said to be *tight* against the top of its tube. Since the tube must have nonzero width at any non-boundary point, $F_0^*(z_i)$ cannot also be tight against the bottom of the tube. Consequently, Equation 18 implies that λ_i^2 is 0. Similarly, when λ_i^2 is nonzero, λ_i^1 must be 0, corresponding to $F_0^*(z_i)$ being tight against the bottom of its tube. If $F_0^*(z_i)$ lies in the middle of the tube, then $\lambda_i^1 = \lambda_i^2 = 0$. An analogous relationship exists between λ_i^3, λ_i^4, and F_1. For convenience, we take advantage of these properties to define two new variable sets:

$$\lambda_i^{12} \triangleq \lambda_i^1 - \lambda_i^2 \quad \text{and}$$
$$\lambda_i^{34} \triangleq \lambda_i^3 - \lambda_i^4$$

Conceptually, λ_i^{12} is positive if and only if $F_0^*(z_i)$ is tight against the top of its tube, negative if and only if $F_0^*(z_i)$ is tight against the bottom of its tube, and 0 if the curve lies within the tube without the need for slack. λ_i^{34} is defined similarly with respect to $F_1^*(z_i)$.

As a consequence of all the above, we observe that ν_α and ν_β, as well as λ_i^{12} and λ_i^{34} for $i = 1 \ldots n$, are pairs of dependent variables whose values are governed by the equations:

$$2^{-\nu_\alpha - \sum_{i=1}^{j} \lambda_{n-i+1}^{12}} + 2^{-\nu_\beta - \sum_{i=1}^{j} \lambda_{n-i+1}^{34}} = 1 \quad \forall j = 0 \ldots n \tag{19}$$

These equations imply that ν_α and ν_β are strictly positive. For any pair λ_i^{12} and λ_i^{34}, it is easy to show that either variable takes a value of 0 if and only if its counterpart also has a value of 0.

3.5 A Constructive Algorithm for the General Solution

An optimal solution to the general problem can be obtained by constructing a set of values that satisfy the KKT conditions from Section 3.4. Informally, we take advantage of the upper triangular structure of Equation 17 to arrive at a feasible solution for the KKT constraints. We propose an algorithm that starts at the bottom of the system and rises to the top, incrementally generating a solution. Figure 3 gives a schematic diagram of the algorithm's operation. Beginning from the rightmost point z_{n+1} and working left, the algorithm locates the pinned points at which the distribution function is tight against the DKW tubes. At termination, the subset of the pinned points for which λ_i^{12} and λ_i^{34} are nonzero have been identified, which in turn enables a solution to be determined through repeated application of Theorem 1. An overview of the procedure is as follows:

1) Create a variable, z_p to denote the leftmost determined pin (not including z_0). Set $z_p = z_{n+1}$. Create two variables, λ_p^{12} and λ_p^{34}, to represent the leftmost undetermined slack variables. Assign the variable ν_α to λ_p^{12}, and the variable ν_β to λ_p^{34}.

2) Check whether there is a *tube-inactive solution* on the interval $[z_0, z_p]$ using the method of Sections 3.6 and 3.7. A tube-inactive solution on an interval $[z_a, z_b]$ is one for which F_0^* and F_1^* are not tight against their tubes throughout the interior of the interval, so that $\lambda_i^{12} = \lambda_i^{34} = 0 \quad \forall i : a < i < b$. In other words, the tube constraints are inactive on the interior of the interval $[z_a, z_b]$. Such a solution clearly reduces to the tube-unconstrained problem addressed in Theorem 1.

   If a tube-inactive solution exists on $[z_0, z_p]$, find the solution on this segment using Theorem 1, and stop. If not, proceed to Step 3.

3) Using the method of Section 3.8, find the rightmost statistic, z_k, at which a tube-inactive solution on $[z_k, z_p]$ is not possible, implying that some λ_i^{12} and λ_i^{34} pair on $[z_{k+1}, z_{p-1}]$ is nonzero. Through the method described in Section 3.9, determine the rightmost statistic, z_m, for which λ_m^{12} and λ_m^{34} must be nonzero, and determine the signs of both λ_m^{12} and λ_m^{34}. This knowledge in turn defines whether each of the cdfs touch the top or the bottom of its respective tube at z_m, thus pinning the solution at $F_0^*(z_m)$ and $F_1^*(z_m)$.

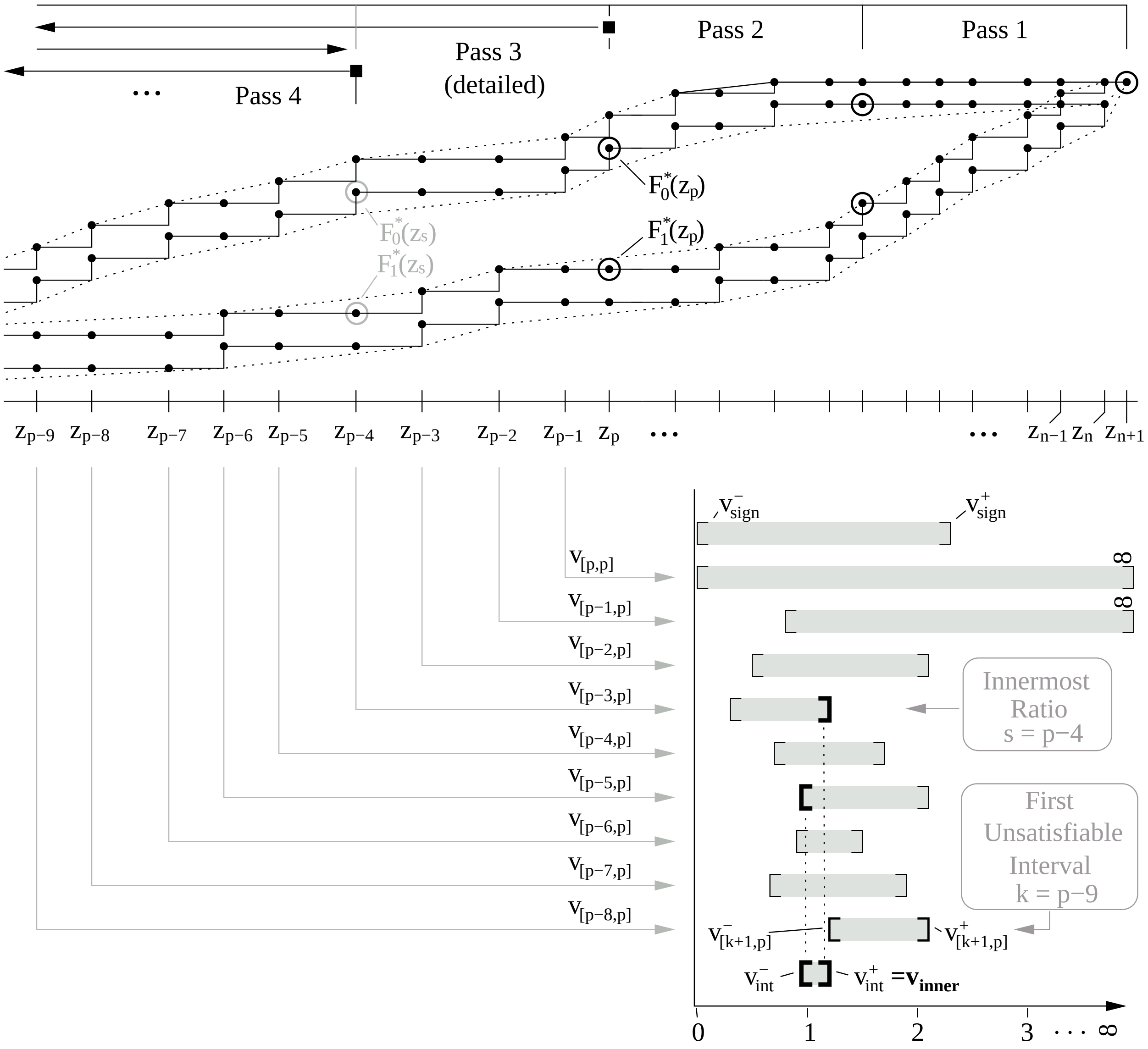


Figure 3: Visual depiction of the algorithm, highlighting Pass 3. Starting at statistic z_p, ratios are constructed from right to left until the first unsatisfiable ratio is encountered, $v_{[p-8,p]}$ in this example. Note that the pictured $[v^-_{sign}, v^+_{sign}]$ results from the direction of the pins from Pass 2, $F_0^*(z_p)$ and $F_1^*(z_p)$. The interval $[v^-_{int}, v^+_{int}]$ marks the intersection of the satisfiable intervals. Next, the innermost ratio on the same side as the unsatisfiable interval is found, which subsequently determines the next point to be pinned. Here, the innermost ratio is $v^+_{[p-3,p]} = v^+_{int}$, and the point to be pinned is therefore $z_s = z_{p-4}$. The algorithm then proceeds to Pass 4, setting $z_p = z_s$.

4) Find a tube-inactive solution on $[z_m, z_p]$, thereby solving for λ_p^{12} and λ_p^{34}.

5) Set $z_p = z_m$. Set $\lambda_p^{12} = \lambda_m^{12}$ and $\lambda_p^{34} = \lambda_m^{34}$. Record the signs of λ_p^{12} and λ_p^{34} for use in Step 3. Go to Step 2.

3.6 The Existence of a Tube-Inactive Solution

By definition, a pinned interval $[z_i, z_p]$ has a tube-inactive solution if the solution curves F_0^* and F_1^* are monotonically non-decreasing and are not affected by the tube constraints of Equation 14. Equivalently, the KKT conditions on the interval are satisfied with:

$$\lambda_{i+1}^{12} = \lambda_{i+1}^{34} = \lambda_{i+2}^{12} = \lambda_{i+2}^{34} = \ldots = \lambda_{p-1}^{12} = \lambda_{p-1}^{34} = 0$$

The primal feasibility, dual feasibility, and complementary slackness conditions are therefore trivially satisfied. Consequently, a tube-inactive solution exists on the interval if and only if it is possible to satisfy the zero-gradient conditions:

$$\begin{array}{ll}
\log \dfrac{f_0^{i+1}}{f_0^{i+1} + f_1^{i+1}} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \qquad \log \dfrac{f_1^{i+1}}{f_0^{i+1} + f_1^{i+1}} = -\lambda_p^{34} - C_{(p,n+1]}^{34} \\
\log \dfrac{f_0^{i+2}}{f_0^{i+2} + f_1^{i+2}} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \qquad \log \dfrac{f_1^{i+2}}{f_0^{i+2} + f_1^{i+2}} = -\lambda_p^{34} - C_{(p,n+1]}^{34} \\
\vdots & \qquad \vdots \\
\log \dfrac{f_0^{p}}{f_0^{p} + f_1^{p}} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \qquad \log \dfrac{f_1^{p}}{f_0^{p} + f_1^{p}} = -\lambda_p^{34} - C_{(p,n+1]}^{34}
\end{array}$$

where $C_{(p,n+1]}^{12}$ and $C_{(p,n+1]}^{34}$ are constants determined by previous iterations of the algorithm:

$$C_{(p,n+1]}^{12} = \begin{cases} \sum_{j=p+1}^{n} \lambda_j^{12} + \nu_\alpha & \text{if } p \leq n \\ 0 & \text{else} \end{cases}$$

$$C_{(p,n+1]}^{34} = \begin{cases} \sum_{j=p+1}^{n} \lambda_j^{34} + \nu_\beta & \text{if } p \leq n \\ 0 & \text{else} \end{cases}$$

To simplify the problem, the zero-gradient conditions can be rewritten into an equivalent system involving the ratios between f_0 and f_1 at each point z_j. This substitution is made possible by noting that at any z_j, setting $v_j = \frac{f_0^j}{f_1^j}$ means that:

$$\frac{f_0^j}{f_0^j + f_1^j} = \frac{v_j}{v_j + 1} \qquad \text{and} \qquad \frac{f_1^j}{f_0^j + f_1^j} = \frac{1}{v_j + 1}$$

Also,

$$\left[v_j^- \triangleq \frac{\min(f_0^j)}{\max(f_1^j)} \right] \leq v_j \leq \left[v_j^+ \triangleq \frac{\max(f_0^j)}{\min(f_1^j)} \right] \tag{20}$$

where monotonicity ensures that $v_j^- \geq 0$ and v_j^+ is finite. The zero-gradient conditions then become:

$$\begin{array}{cc}
\log \frac{v_{i+1}}{v_{i+1} + 1} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \log \frac{1}{v_{i+1} + 1} = -\lambda_p^{34} - C_{(p,n+1]}^{34} \\
\log \frac{v_{i+2}}{v_{i+2} + 1} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \log \frac{1}{v_{i+2} + 1} = -\lambda_p^{34} - C_{(p,n+1]}^{34} \\
\vdots & \vdots \\
\log \frac{v_p}{v_p + 1} = -\lambda_p^{12} - C_{(p,n+1]}^{12} & \log \frac{1}{v_p + 1} = -\lambda_p^{34} - C_{(p,n+1]}^{34}
\end{array}$$

Recall that by Theorem 1, any solution on a pinned interval that satisfies all feasibility and complementary slackness conditions and has the property that the densities f_0 and f_1 are multiples must be an optimal solution on that interval. The existence of a ratio v_j means that the two probability masses at z_j differ by the constant multiplier v_j. Consequently, the issue of finding whether a tube-inactive solution exists on $[z_i, z_p]$ is equivalent to finding whether there exists some ratio, v, that satisfies all of the v_j constraints simultaneously, meaning:

$$v = v_j \quad \forall j : (i+1) \leq j \leq p$$

We call the problem of finding such a v the *ratio satisfiability problem*.

Substituting the satisfying ratio v, the zero-gradient system simplifies to:

$$\log \frac{v}{v+1} = -\lambda_p^{12} - C_{(p,n+1]}^{12} \qquad \log \frac{1}{v+1} = -\lambda_p^{34} - C_{(p,n+1]}^{34} \tag{21}$$

One final transformation will facilitate an iterative solution to the ratio satisfiability problem. We define each $v_{[j,p]}$ to be the ratio of the probability masses of the two curves on the interval $[z_{j-1}, z_p]$:

$$\left[v^-_{[j,p]} \triangleq \frac{\min(\sum_{l=j}^p f_0^l)}{\max(\sum_{l=j}^p f_1^l)}\right] \leq v_{[j,p]} \leq \left[v^+_{[j,p]} \triangleq \frac{\max(\sum_{l=j}^p f_0^l)}{\min(\sum_{l=j}^p f_1^l)}\right]$$
$$\equiv \left[v^-_{[j,p]} = \frac{\max(0, F_0^*(z_p) - \overline{F_0^+}(z_{j-1}))}{F_1^*(z_p) - \overline{F_1^-}(z_{j-1})}\right] \leq v_{[j,p]} \leq \left[v^+_{[j,p]} = \frac{F_0^*(z_p) - \overline{F_0^-}(z_{j-1})}{\max(0, F_1^*(z_p) - \overline{F_1^+}(z_{j-1}))}\right] \tag{22}$$

It is straightforward to show that either set of ratios has a satisfying v if and only if the other ratio set is satisfied by the same v:

$$v_{i+1} = v_{i+2} = \ldots = v_p = v \Leftrightarrow v_{[i+1,p]} = v_{[i+2,p]} = \ldots = v_{[p,p]} = v$$

Henceforth, we refer to the ratio satisfiability problem for the $v_{[j,p]}$ ratio set, meaning:

$$v = v_{[j,p]} \quad \forall j : (i+1) \leq j \leq p$$

The ratio satisfiability problem is pictured in Figure 3. Algorithmically, it can be solved by computing the intersection of all the $v_{[j,p]}$ intervals between $v_{[i+1,p]}$ and $v_{[p,p]}$, which is all such ratios on the interval $[z_i, z_p]$. If the intersection is non-empty:

$$\bigcap_{j=i+1}^{p} \left[v^-_{[j,p]}, v^+_{[j,p]}\right] \neq \emptyset \tag{23}$$

then the conditions are satisfiable, and the interval $[z_i, z_p]$ has a tube-inactive solution. Otherwise, the solution on the given interval must be tight against the tubes at some intermediate point.

3.7 The Extended Ratio Satisfiability Problem

It is clear from the description of the algorithm that $\lambda_p^{12} \neq 0$ and $\lambda_p^{34} \neq 0$ for all algorithmic passes. In the case of the first pass, $\lambda_p^{12} \triangleq \nu_\alpha > 0$ and $\lambda_p^{34} \triangleq \nu_\beta > 0$. For all subsequent passes, the signs of λ_p^{12} and λ_p^{34} will have already been determined by the previous pass. Therefore, the ratio satisfiability problem must be amended to account for this additional constraint. Let the variable t represent the index of the current pass of the algorithm: $t = 1$ denotes the first pass, $t = 2$ the second pass, and so on.

An equivalent condition for the signs in terms of a satisfying ratio $v = v(t)$ becomes clear when examining the zero-gradient conditions from Equation 21. The constants

$C^{12}_{(p,n+1]}$ and $C^{34}_{(p,n+1]}$ can be unrolled as:

$$\begin{aligned} -\log\frac{v(t-1)}{v(t-1)+1} &= C^{12}_{(p,n+1]} = C^{12}_{(p+1,n+1]} + \lambda^{12}_{p+1} \\ -\log\frac{v(t-1)}{v(t-1)+1} &= C^{34}_{(p,n+1]} = C^{34}_{(p+1,n+1]} + \lambda^{34}_{p+1} \end{aligned}$$

The t^{th} pass zero-gradient conditions from Equation 21 are then equivalent to:

$$\frac{v(t)}{v(t)+1} = \frac{v(t-1)}{v(t-1)+1}2^{-\lambda^{12}_p} \qquad \frac{1}{v(t)+1} = \frac{1}{v(t-1)+1}2^{-\lambda^{34}_p} \tag{24}$$

Now if $F^*_0(z_p)$ is pinned at the top of its tube, then $\lambda^{12}_p > 0$, and subsequently:

$$\begin{aligned} & \frac{v(t)}{v(t)+1} < \frac{v(t-1)}{v(t-1)+1} \\ \equiv \qquad & v(t) < v(t-1) \end{aligned} \tag{25}$$

On the other hand, if $F^*_0(z_p)$ is pinned at the bottom of its tube, then $\lambda^{12}_p < 0$, and:

$$v(t) > v(t-1) \tag{26}$$

Also, it follows from Equation 19 that if the solution for one cdf is tight against its tube at any statistic z_p (for $z_p \neq z_{n+1}$), then the other cdf must also be tight against its tube.

Corollary 1. *At any statistic z_p where $z_p < z_{n+1}$ and the solution curves F^*_0 and F^*_1 are tight against their tubes, the two curves must be be positioned either towards each other or away from each other:*

$$\begin{aligned} F^*_0(z_p) = \overline{F^+_0}(z_p) \Leftrightarrow F^*_1(z_p) = \overline{F^-_1}(z_p) \\ F^*_0(z_p) = \overline{F^-_0}(z_p) \Leftrightarrow F^*_1(z_p) = \overline{F^+_1}(z_p) \end{aligned} \tag{27}$$

Proof of Corollary 1. As shown above (Equation 25):

$$F^*_0(z_p) = \overline{F^+_0}(z_p) \Rightarrow v(t) < v(t-1)$$

A similar result for F_1 can be derived from Equation 24, yielding:

$$F^*_1(z_p) = \overline{F^+_1}(z_p) \Rightarrow v(t) > v(t-1) \tag{28}$$

Hence, if both curves are tight against the tops of their tubes, then no consistent v exists for algorithmic pass t, which contradicts the assumption that this pinned point is part of a global solution. Analogous logic shows that the tubes cannot be both tight against the bottom of their tubes. Therefore the two curves must be tight either towards each other or away from each other. □

So then $v(t-1)$ places a bound on the new ratio $v(t)$, and this bound is equivalent to enforcing the signs of λ_p^{12} and λ_p^{34} in the zero-gradient system. Let $v^-(t-1)$ denote the minimum ratio $v(t-1)$ that would have satisfied the interval handled by the previous pass. Since z_p must be pinned as in Corollary 1, the definitions of Equation 22 imply that $v(t-1) = v^-(t-1)$ if and only if $F_0^*(z_p)$ is pinned at the top of its tube and $F_1^*(z_p)$ is pinned at the bottom of its tube. Similarly, let $v^+(t-1)$ denote the maximum ratio $v(t-1)$ that would have satisfied the previous interval. Then $v(t-1) = v^+(t-1)$ when $F_0^*(z_p)$ is pinned down and $F_1^*(z_p)$ is pinned up. Incorporating Equations 25 and 26, the extended ratio satisfiability problem can be completed by including a constraint imposed by the range:

$$\left[v_{sign}^-, v_{sign}^+\right] \triangleq \begin{cases} [0,\infty) & \text{if } t=1 \\ [0, v(t-1)\,) & \text{if } v(t-1) = v^+(t-1) \\ (\,v(t-1),\infty) & \text{if } v(t-1) = v^-(t-1) \end{cases} \tag{29}$$

Then a satisfying ratio can be found if and only if:

$$\left[v_{sign}^-, v_{sign}^+\right] \cap \bigcap_{j=i+1}^{p} \left[v_{[j,p]}^-, v_{[j,p]}^+\right] \neq \emptyset \tag{30}$$

3.8 The Nonexistence of a Tube-Inactive Solution

Because a satisfying v can be found for some interval if and only if the interval has a tube-inactive solution, the lack of a satisfying v on an interval $[z_k, z_p]$ indicates that no tube-inactive solution exists:

$$\left[v_{sign}^-, v_{sign}^+\right] \cap \bigcap_{j=k+1}^{p} \left[v_{[j,p]}^-, v_{[j,p]}^+\right] = \emptyset \tag{31}$$

During the execution of the algorithm, the statistic z_k must be determined for each pass, relative to the current z_p. The intervals $\{[v_{[j,p]}^-, v_{[j,p]}^+] \mid j = (k+2)\dots p\}$, are collectively referred to as the *satisfiable intervals*. The intersection of the satisfiable intervals and the current range constraint is denoted as:

$$[v_{int}^-, v_{int}^+] \triangleq \left[v_{sign}^-, v_{sign}^+\right] \cap \bigcap_{j=k+2}^{p} \left[v_{[j,p]}^-, v_{[j,p]}^+\right] \tag{32}$$

The interval $[v_{[k+1,p]}^-, v_{[k+1,p]}^+]$ is referred to as the *first unsatisfiable interval* since, as the algorithm works left from z_p, z_k is the first point at which the above intersection is empty. In this case, there must be some statistic on $[z_{k+1}, z_{p-1}]$ at which the curves are tight against their tubes. Since the algorithm seeks to place pins at every tight point, we must find the rightmost such point, which we denote z_m. Once z_m has been found for the current pass, the algorithm proceeds to the next pass using z_m as the new z_p.

3.9 Finding the Rightmost Tight Statistic, z_m

Identifying z_m, the rightmost statistic on an interval $[z_{k+1}, z_{p-1}]$ whose solution is tight against the tubes, follows from a simple property of the set of all minimum and maximum ratios on the interval. We define the *innermost ratio* $v_{[s+1,p]}$ and a corresponding *innermost statistic* z_s as follows:

$$v_{[s+1,p]} \triangleq \begin{cases} v^-_{int} & \text{if } v^+_{[k+1,p]} < v^-_{int}; \\ v^+_{int} & \text{if } v^-_{[k+1,p]} > v^+_{int} \end{cases} \tag{33}$$

and

$$z_s \triangleq \begin{cases} z_{l-1} & \text{if } v^+_{[k+1,p]} < v^-_{int}; \\ z_{r-1} & \text{if } v^-_{[k+1,p]} > v^+_{int} \end{cases} \tag{34}$$

where

$$l = \operatorname*{arg\,max}_{j \,|\, (k+2) \leq j \leq p} \left(v^-_{[j,p]} \right) \qquad \text{and} \qquad r = \operatorname*{arg\,min}_{j \,|\, (k+2) \leq j \leq p} \left(v^+_{[j,p]} \right)$$

So by definition, the innermost ratio always lies on the same side of the satisfiable intervals as the first unsatisfiable interval $[v^-_{[k+1,p]}, v^+_{[k+1,p]}]$.

Theorem 2 will prove that the innermost statistic and the rightmost tight statistic are equivalent. The theorem relies on the following lemma:

Lemma 1. *Given a statistic z_m that is chosen to pin the interval $[z_m, z_p]$ on the t^{th} algorithmic pass, if z_m is pinned so that:*

$$F^*_0(z_m) = \overline{F^+_0}(z_m) \qquad \textit{and} \qquad F^*_1(z_m) = \overline{F^-_1}(z_m)$$

(which is true when $v_{[m+1,p]}(t) = v^-_{[m+1,p]}(t)$), and if

$$F^*_0(z_p) \geq \overline{F^+_0}(z_{m-1}) \qquad \textit{and} \qquad v^-_{[m,p]}(t) < v^-_{[m+1,p]}(t)$$

then:

$$v^-_{[m,m]}(t+1) < v^-_{[m,p]}(t)$$

Similarly, if z_m is pinned so that:

$$F_0^*(z_m) = \overline{F_0^-}(z_m) \qquad \textit{and} \qquad F_1^*(z_m) = \overline{F_1^+}(z_m)$$

(which is true when $v_{[m+1,p]}(t) = v^+_{[m+1,p]}(t)$), and if

$$F_1^*(z_p) \geq \overline{F_1^+}(z_{m-1}) \qquad \textit{and} \qquad v^+_{[m,p]}(t) > v^+_{[m+1,p]}(t)$$

then:

$$v^+_{[m,m]}(t+1) > v^+_{[m,p]}(t)$$

In other words, when placing a pin at z_m, if $F_0^*(z_m)$ is tight against the top of its tube and $F_1^*(z_m)$ is tight against the bottom, then the minimum ratio $v^-_{[m,m]}(t+1)$, the first new minimum ratio to the left of the pin on the algorithm's next pass, will be less than the old ratio $v^-_{[m,p]}(t)$ as long as $v^-_{[m,p]}(t) < v^-_{[m+1,p]}(t)$. Similarly, if $F_0^*(z_m)$ is tight against the bottom of its tube and $F_1^*(z_m)$ is tight against the top, then the maximum ratio $v^+_{[m,m]}(t+1)$ will be greater than the old ratio $v^+_{[m,p]}(t)$ as long as $v^+_{[m,p]}(t) > v^+_{[m+1,p]}(t)$.

Proof of Lemma 1.

$$\begin{aligned}
& v^-_{[m,p]}(t) < v^-_{[m+1,p]}(t) \\
\equiv \quad & \frac{F_0^*(z_p) - \overline{F_0^+}(z_{m-1})}{F_1^*(z_p) - \overline{F_1^-}(z_{m-1})} < \frac{F_0^*(z_p) - \overline{F_0^+}(z_m)}{F_1^*(z_p) - \overline{F_1^-}(z_m)} \\
\equiv \quad & \frac{\overline{F_0^+}(z_m) - \overline{F_0^+}(z_{m-1})}{\overline{F_1^-}(z_m) - \overline{F_1^-}(z_{m-1})} < \frac{F_0^*(z_p) - \overline{F_0^+}(z_{m-1})}{F_1^*(z_p) - \overline{F_1^-}(z_{m-1})} \\
\equiv \quad & v^-_{[m,m]}(t+1) < v^-_{[m,p]}(t)
\end{aligned}$$

For the second case,

$$\begin{aligned}
& v^+_{[m,p]}(t) > v^+_{[m+1,p]}(t) \\
\equiv \quad & \frac{F_0^*(z_p) - \overline{F_0^-}(z_{m-1})}{F_1^*(z_p) - \overline{F_1^+}(z_{m-1})} > \frac{F_0^*(z_p) - \overline{F_0^-}(z_m)}{F_1^*(z_p) - \overline{F_1^+}(z_m)} \\
\equiv \quad & \frac{\overline{F_0^-}(z_m) - \overline{F_0^-}(z_{m-1})}{\overline{F_1^+}(z_m) - \overline{F_1^+}(z_{m-1})} > \frac{F_0^*(z_p) - \overline{F_0^-}(z_{m-1})}{F_1^*(z_p) - \overline{F_1^+}(z_{m-1})} \\
\equiv \quad & v^+_{[m,m]}(t+1) > v^+_{[m,p]}(t)
\end{aligned}$$

□

Theorem 2. *Given some z_k and z_p such that a tube-inactive solution exists on $[z_{k+1}, z_p]$, but not on $[z_k, z_p]$, the rightmost statistic z_m at which the global solution must be tight against the tubes is exactly the innermost statistic z_s.*

Proof of Theorem 2. Clearly, the ratio $v_{[m+1,p]}$ corresponding to the rightmost statistic z_m must possess three attributes:

1) Feasibility. The ratio must represent two cdfs that touch their tubes, so that either:

$$\begin{aligned} F_0^*(z_m) = \overline{F_0^+}(z_m) \quad &\text{and} \quad F_1^*(z_m) = \overline{F_1^-}(z_m) \quad \text{or} \\ F_0^*(z_m) = \overline{F_0^-}(z_m) \quad &\text{and} \quad F_1^*(z_m) = \overline{F_1^+}(z_m) \end{aligned} \tag{35}$$

implying that either:

$$\begin{aligned} v_{[m+1,p]} &= v^-_{[m+1,p]} \quad \text{or} \\ v_{[m+1,p]} &= v^+_{[m+1,p]} \end{aligned} \tag{36}$$

2) Backward Consistency. The chosen ratio must satisfy the interval $[z_m, z_p]$, meaning:

$$v^-_{[j,p]} \leq v_{[m+1,p]} \leq v^+_{[j,p]} \quad \forall j : (m+1) \leq j \leq p \tag{37}$$

3) Forward Consistency. The ratio must be consistent with a global solution. In other words, it must not contradict the existence of a solution on the remaining interval $[z_0, z_m]$. The algorithm of Section 3.5 can proceed as long as the statistic z_m can be included in the next pass, thereby inductively guaranteeing a solution on $[z_0, z_m]$:

$$\left[v^-_{sign}(t+1), v^+_{sign}(t+1)\right] \cap \left[v^-_{[m,m]}(t+1), v^+_{[m,m]}(t+1)\right] \neq \emptyset \tag{38}$$

Now, consider the innermost statistic z_s and the innermost ratio $v_{[s+1,p]}$. This ratio is feasible by definition, and backwards consistent since there is a tube-inactive solution for the entire interval $[z_s, z_p]$.

Furthermore, the ratio can be shown to be forward consistent. First, consider the case when the first unsatisfiable interval lies to the left of the satisfiable intervals, so that $v_{[s+1,p]}(t) = v^-_{int}(t) = v^-_{[s+1,p]}(t)$. We show that Lemma 1 applies for $z_m = z_s$. By the definition of an innermost ratio, $v^-_{[s,p]} < v^-_{[s+1,p]}$. Furthermore, we see that $\overline{F_0^+}(z_{s-1}) < F_0^*(z_p)$, since if $\overline{F_0^+}(z_{s-1}) \geq F_0^*(z_p)$ then $\overline{F_0^+}(z_s) \geq F_0^*(z_p)$ and so $v_{[s+1,p]} = 0$, which contradicts the assumption that the first unsatisfiable interval lies to the left of v^-_{int}. By Lemma 1, the new minimum ratio $v^-_{[s,s]}(t+1)$ cannot be greater than the old minimum ratio, $v^-_{[s,p]}(t)$, which is also the bound derived from the signs of the λ_p's.

On the other hand, if the first unsatisfiable interval lies to the right of the satisfiable intervals, then $v_{[s+1,p]}(t) = v^+_{int}(t) = v^+_{[s+1,p]}(t)$, $v^+_{[s,p]} > v^+_{[s+1,p]}$, and $\overline{F^+_1}(z_{s-1}) < F^*_1(z_p)$. Again Lemma 1 applies, stating that the new maximum ratio $v^+_{[s,s]}(t+1)$ cannot be less than the old maximum ratio $v^+_{[s,p]}(t)$.

Therefore, the interval $\left[v^-_{[s,s]}(t+1), v^+_{[s,s]}(t+1)\right]$ must be subsumed by the interval $\left[v^-_{sign}(t+1), v^+_{sign}(t+1)\right]$. Since the interval itself must be nonempty, the ratio $v_{[s+1,p]}$ is forward consistent by Equation 38.

The innermost ratio z_s satisfies three properties of the rightmost statistic, showing that it is consistent with a global solution as a candidate for z_m. By Theorem 1, the unconstrained solution on $[z_s, z_p]$ obtained by pinning only at z_s and z_p must be optimal. Therefore, there can be no other statistic $z_j : z_s < z_j < z_p$ that is tight against its tubes without contradicting Theorem 1. So, the statistic z_s corresponding to the innermost ratio is exactly the rightmost statistic z_m. □

The algorithm of Section 3.5 is thereby proven to construct a solution on $[z_0, z_{n+1}]$. Since the procedure equivalently satisfies the KKT conditions for the general problem, the solution is guaranteed to be optimal. The running time of the algorithm is $O(n^2)$, since it moves linearly through the order statistics to determine the pin locations, with exactly one linear backtrack on each interval to find each rightmost z_m. Supplemental information about the algorithm, including pseudocode, is provided in Appendix A.

4 Evaluation of the MI Bound and the Naïve Method

4.1 Model Distributions

We demonstrate the performance of our lower bound on two well-known distribution families for which the true MI between class labels and outputs is computed numerically. The confidence used for each class-conditional distribution is 0.99. Letting n again denote the sample size, the DKW tube widths are fixed to $2\epsilon_{dkw}$ according to Equation 4. It is important to note that overall confidence in the lower bound on MI

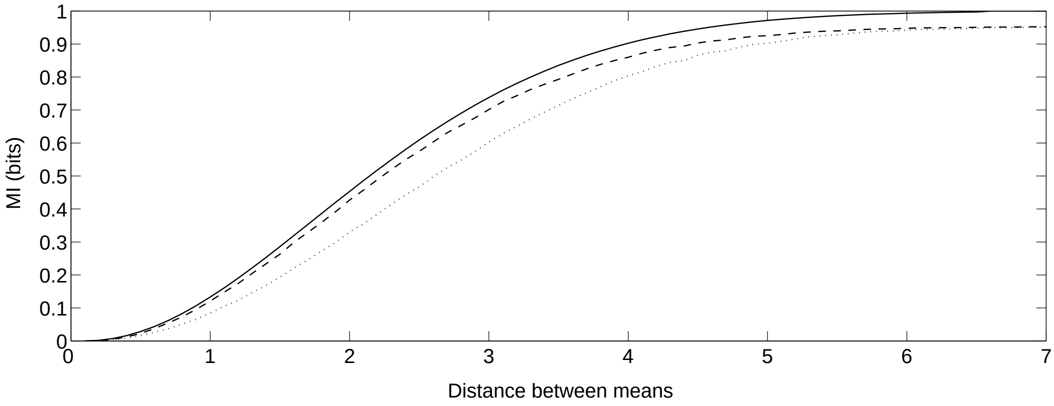


Figure 4: Lower bounds on MI for normally-distributed points from two classes. The x-axis denotes distance between the means of the two distributions. The solid line represents the true MI, the dashed line is the lower bound on MI given by the algorithm, and the dotted line shows the lower bound given by the Fano method.

is the product of the two confidence values from the bounds on $P(Y|X = 0)$ and $P(Y|X = 1)$. This relationship is an obvious consequence of the independence of the two distributions. For clarity we will write the overall confidence explicitly as a square: for the following results the overall confidence is 0.99^2.

We first generate data for two normally-distributed classes with unit variances. Letting d denote the distance between the means of the two distributions, the true MI is:

$$1+\frac{1}{2\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[e^{-\frac{(y-d)^2}{2}}\log\frac{e^{yd-\frac{d^2}{2}}}{1+e^{yd-\frac{d^2}{2}}}-e^{-\frac{y^2}{2}}\log\left(1+e^{yd-\frac{d^2}{2}}\right)\right]\mathrm{d}y$$

For $n =$200,000, Figure 4 compares the lower bound on MI obtained by our result to the true MI, as well as to the lower bound given by the Fano method, for a number of distances.

While our algorithm produces an MI bound without assuming any particular classification of the data, the Fano method relies on the per-class empirical probabilities of error, and thus inherently requires a classifier. The Fano results shown here are computed using the theoretically-optimal discriminant from the true generating distributions, in order to avoid any error stemming from the use of a sub-optimal discriminant. Using the correction from Equation 3 and a confidence of 0.99^2, the Fano method gives a lower bound on MI as in Equation 2.

In addition to the normally-distributed inputs, we sample two gamma distributions

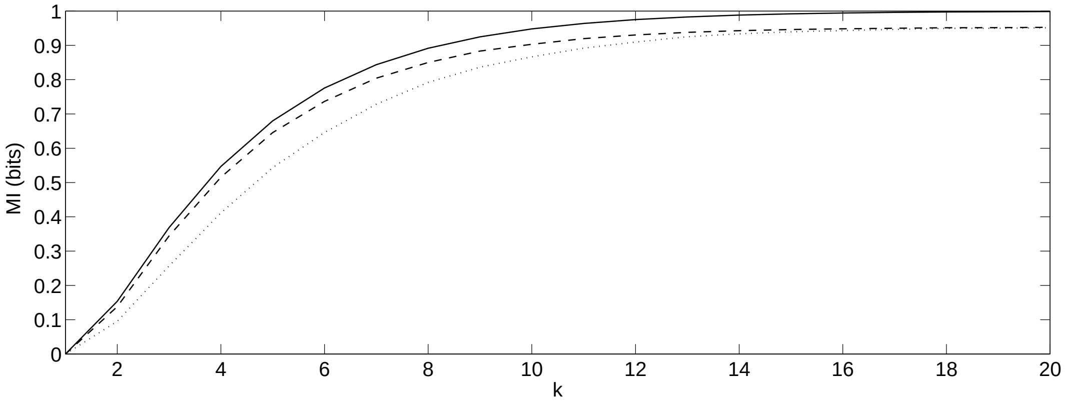


Figure 5: Lower bounds on MI for gamma-distributed points from two classes. The x-axis denotes the shape parameter for the second class distribution (the first distribution has a fixed shape parameter of 1). The solid line shows the true MI, the dashed line is the lower bound on MI given by the algorithm, and the dotted line depicts the lower bound given by the Fano method.

with unit scale parameters. The first class is generated with shape parameter $k = 1$, and the second class uses an integer shape parameter denoted by the variable k. Thus, both classes are drawn from Erlang distributions, and the true MI is:

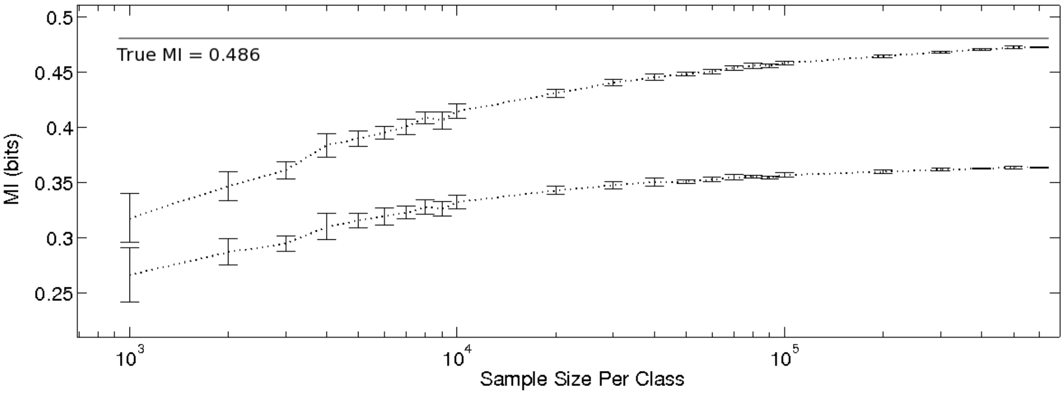
$$1 + \frac{1}{2}\int_{-\infty}^{\infty} e^{-y}\left[\log\frac{(k-1)!}{y^{k-1}+(k-1)!} + \frac{y^{k-1}}{(k-1)!}\log\frac{y^{k-1}}{y^{k-1}+(k-1)!}\right]\mathrm{d}y$$

A comparison of the lower bounds on MI and the actual MI using $n =$200,000 is shown in Figure 5.

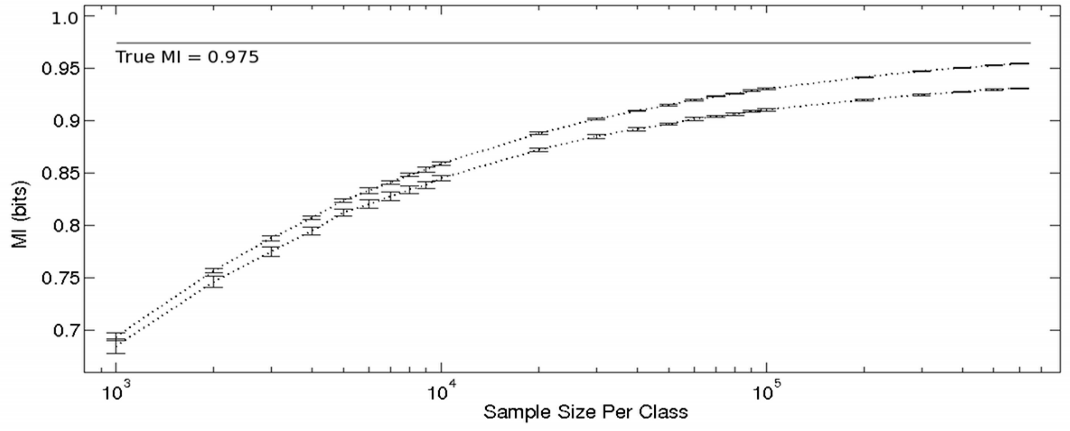
As for any MI estimation technique, the lower bound given by these two procedures is clearly dependent on the sample size. For model Gaussian distributions, the behaviors of the two methods as n increases are shown in Figure 6, with sample sizes ranging from 2,000 to 1,200,000.

4.2 Spike Train Data

To demonstrate the application of our method to neural data, we employ the Meddis Inner-Hair Cell Model, which generates realistic auditory nerve spike trains from digital sound stimuli (Meddis, 1986, 1988; Sumner et al., 2002). Our simulations are based on a standard human parameter set derived from psychophysical data (Lopez-Poveda et

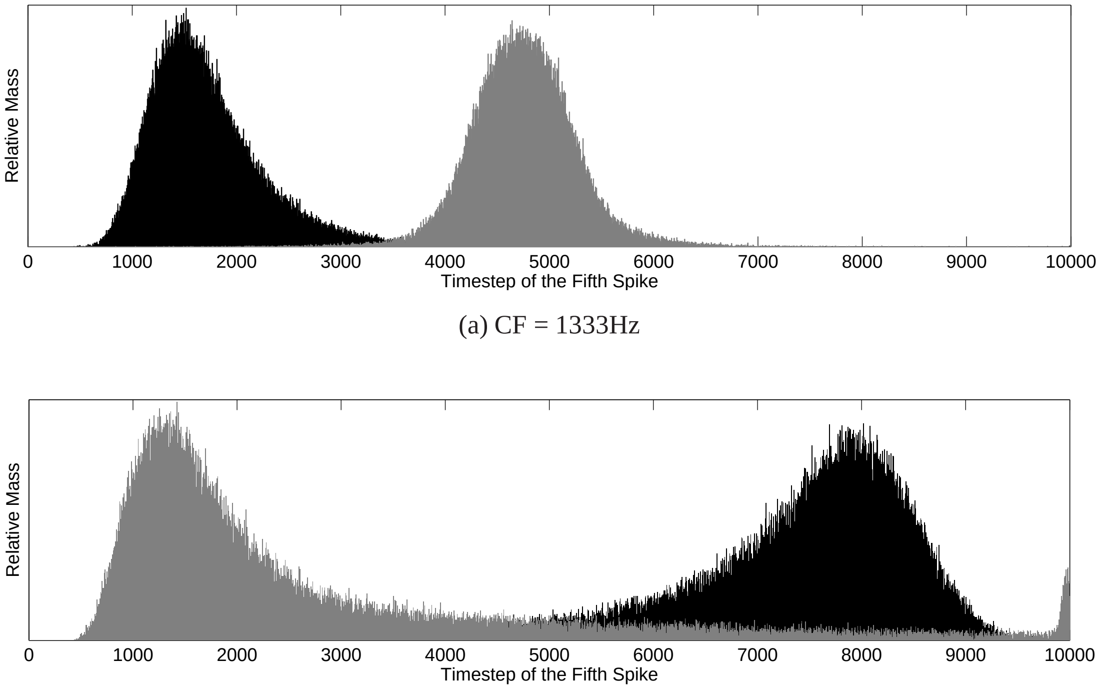


(a) Distance between means = 2



(b) Distance between means = 5

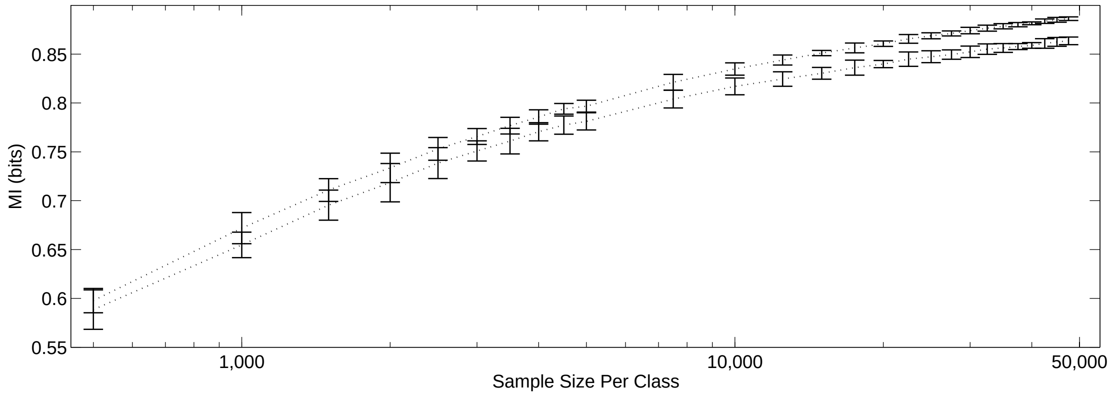
Figure 6: MI lower bounds as a function of sample size for two pairs of Gaussian classes. The upper panel shows results for two Gaussians whose means differ by 2, and the lower panel shows two Gaussians whose means differ by 5. The x-axes are scaled logarithmically. Each point represents the mean lower bound computed over 10 randomly-generated samples and the bars denote two standard deviations from the mean. For both plots, the upper curves depict the bounds given by our algorithm and the lower curves depict the bounds given by the Fano method.



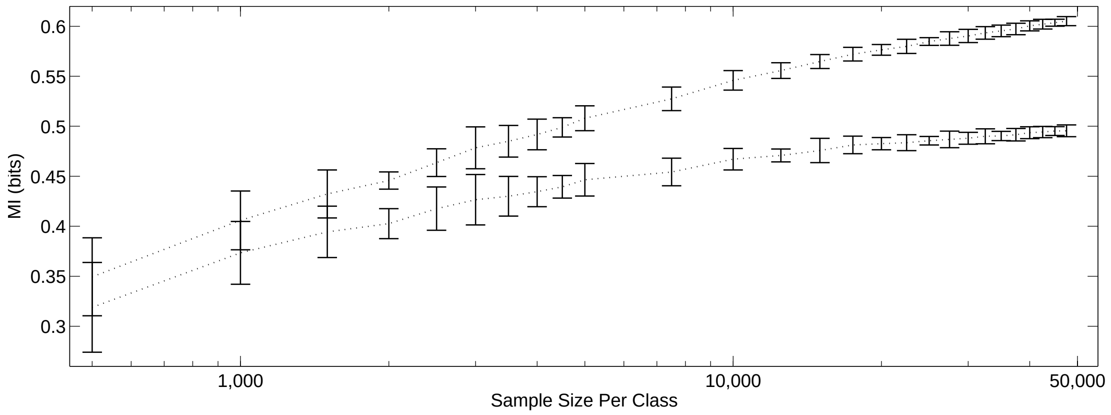
(a) CF = 1333Hz

(b) CF = 2200Hz

Figure 7: Response distributions of the time to fifth spike by two model auditory neurons for up/down frequency sweeps, n = 100,000. The upper panel shows responses for a neuron with a CF of 1333Hz; the lower panel shows responses for a neuron with a CF of 2200Hz. For both plots, the black distribution is the class of upward sweeps and the gray denotes downward sweeps.



(a) CF = 1333Hz



(b) CF = 2200Hz

Figure 8: MI lower bounds as a function of sample size for two CF responses. The upper panel shows results for a neuron with a CF of 1333Hz; the lower panel shows a CF of 2200Hz. The x-axes are scaled logarithmically. Each point represents the mean lower bound computed over 10 randomly-generated samples and the bars denote two standard deviations from the mean. For both plots, the upper curves show the bounds given by our algorithm and the lower curves show the bounds given by the Fano method.

al., 2001). We use only high spontaneous rate fibers and an optimized gain level that remains fixed for all trials. Two center frequencies (CF) are modeled: 1333Hz and 2200Hz, using bandwidths of 243.5Hz and 359.0Hz, respectively.

Our stimuli are $100ms$ sounds representing a simple binary task: upward and downward frequency sweeps, or chirps, of 1000Hz. To generate each 100ms stimulus, the class (up or down) is selected and a starting frequency is chosen at random. For ‘up’ stimuli, the starting frequencies range between 950Hz and 1050Hz, and for ‘down’ stimuli, the range is 1950Hz to 2050Hz. Each interval is generated by:

$$y(t) = \sin(2\pi t(\omega + zt * 1000Hz/0.1ms/2) + \phi)$$

where t ranges from 0ms to 100ms, ω denotes the starting frequency, $z = \pm 1$ is positive for ‘up’ stimuli and negative for ‘down’ stimuli, and ϕ is the initial phase of the sound. Every stimulus is phase-matched to the preceding interval so that discontinuities in the physical waveform do not mark the interval boundaries.

There are many ways in which one could map the spike train response of each neuron to a one-dimensional scalar value. Since differentiating between upward and downward sweeps requires precise spike timing information, a natural mapping candidate is the time between the stimulus onset and the first response spike. Due to the spontaneous firing of the spiral ganglion cells, however, this representation is not effective for neurons with CFs in the middle of the frequency range.

Instead, we record the time between the onset of the stimulus and the fifth spike of the response. Our implementation uses input sound files sampled at 100KHz, necessitating response bins, or timesteps, of 0.01ms. Distributions of the time to fifth spike for both CFs are portrayed in Figure 7 for a sample size of 100,000. For a confidence of 0.99^2, lower bounds on MI computed using our method and the Fano method for both CFs are plotted in Figure 8 as functions of the sample size.

5 Discussion

The tightness of our lower bound is made possible by the result of Massart (1990). Unfortunately, to our knowledge, there is no equivalently strong theory addressing the case of multi-dimensional output. In this case, our method may still be used to obtain a lower bound on MI through the following procedure: Using a hold-out set, find a good mapping from the high-dimensional space to $\mathbb{R}$. Then, apply our algorithm on the remainder of the data.

Determining a suitable mapping is a problem of wide interest. One intriguing idea is to obtain a mapping to $\mathbb{R}$ as the byproduct of an existing statistical classification technique. While the final goal of a classifier is to map its input to a discrete set, the inner workings of many classifiers can be viewed as first projecting data onto the real line, then partitioning the line to make class assignments. Discriminant classifiers, for instance, project each point onto the real line denoting distance from the discriminant, then predict a class label based on the projected sign. As another example, consider a clustering algorithm assigning points to two classes. For each point, a real-valued confidence can be derived based on a function of the Mahalanobis distance between the point and the two cluster centers. Clearly, such intermediate representations contain more information than the overall error rate of the classifier. This extra information can be extracted through our technique.

The probabilistic lower bound on MI developed here is a promising new approach to the MI estimation problem. It is distribution-free and operates in the finite-sample regime. Furthermore, a probabilistic bound has a clear interpretation, as the result is qualified simply by a confidence level chosen by the user. This method is ideal for appraising the MI over any feedforward channel with binary input and real-valued output.

Acknowledgments

This research was in part supported by NSF IIS-0902230 to AB. In addition, the authors acknowledge the University of Florida High-Performance Computing Center for providing computational resources and support that have contributed to the research results reported within this article.

A Implementing the Algorithm

A.1 Finite-precision sampling

The algorithm presented in the text assumes that the values of a sample are drawn with infinite precision from $\mathbb{R}$, which is obviously precluded in reality by the limits of instrumentation and digital computing. Consequently, the uniqueness of the algorithm's answer depends on the assumption that any two order statistics have the exact same value with probability 0. If in practice these values are not unique, the resulting ambiguity may be resolved through a simple, linear preprocessing of the sample before execution of the main algorithm.

Consider any set of observed order statistics $z_i, \ldots, z_j$ such that $z_i = \ldots = z_j$ to the maximum discernible precision, and let y denote this common value. We assume that the error due to the finite sampling precision is small, so that the true points represented by $z_i, \ldots, z_j$ are uniformly-distributed around the approximation y. For each class, we assign new, higher-precision values to the duplicate order statistics from that class to redistribute the points evenly around y. This procedure guarantees that no two points from the same class will share the same value, ensuring a consistent answer from the algorithm.

A.2 Pseudocode

Annotated pseudocode for our algorithm follows. In addition, source code for an implementation in C is available online at:

http://www.cise.ufl.edu/research/compbio1/LowerBoundMI/

```
procedure find_minMI
{
  MI := 0;
  $y_p$ := n+1;
  $F_0^*(z_0)$ := 0;        $F_0^*(z_{n+1})$ := 1;
  $F_1^*(z_0)$ := 0;        $F_1^*(z_{n+1})$ := 1;
  vmin := 0;        vmax := $\infty$;
  inner_min := $y_p$;  inner_max := $y_p$;

  // Work backwards to find the pin locations
```

```
while( $y_p$ != 0 )
{
    // Find the longest interval for which we have
    //    a tube-unconstrained solution
    for( $i$ := $(p-1)$ to 0, step -1 )
    {
        // Determine minimum ratio for $v = \frac{f_0}{f_1}$ at $y_i$
        if( $F_0^*(y_p)$ <= $\overline{F_0^+}(y_i)$ )
            vi_min := 0;
        else if( $F_1^*(y_p)$ == $\overline{F_1^-}(y_i)$ )
            vi_min := $\infty$;
        else
            vi_min := $\frac{F_0^*(y_p)-\overline{F_0^+}(y_i)}{F_1^*(y_p)-\overline{F_1^-}(y_i)}$;

        // Determine maximum ratio for $v = \frac{f_0}{f_1}$ at $y_i$
        if( $F_1^*(y_p)$ <= $\overline{F_1^+}(y_i)$ )
            vi_max := $\infty$;
        else
            vi_max := $\frac{F_0^*(y_p)-\overline{F_0^-}(y_i)}{F_1^*(y_p)-\overline{F_1^+}(y_i)}$;

        if( vi_max < vmin )       // First Unsat. Interval
        {
            // Fix the pins at inner_min to the minimum
            $y_k$ = inner_min;
            $F_0^*(y_k)$ := $\overline{F_0^+}(y_k)$;
            $F_1^*(y_k)$ := $\overline{F_1^-}(y_k)$;

            // Init vmin/vmax values for next while() pass
            //   (Extended Ratio Satisfiability)
            vmax := vmin;
            vmin := 0;
            break;
        }
        else if( vi_min > vmax )  // First Unsat. Interval
        {
            // Fix the pins at inner_max to the maximum
            $y_k$ := inner_max;
            $F_0^*(y_k)$ := $\overline{F_0^-}(y_k)$;
            $F_1^*(y_k)$ := $\overline{F_1^+}(y_k)$;

            // Init vmin/vmax values for next while() pass
            //   (Extended Ratio Satisfiability)
            vmin := vmax;
            vmax := $\infty$;
            break;
```

```
          }
          else
          {
              // Update running intersection with vi
              if( vi_min > vmin )
              {
                  vmin := vi_min;
                  inner_min := $y_i$;
              }
              if( vi_max < vmax )
              {
                  vmax := vi_max;
                  inner_max := $y_i$;
              }
              $y_k$ := $y_i$;
          }
      } // end for($i$)

      // Determine solution value on $[y_k, y_p]$ using Th.1
      $\alpha$ := $F_0^*(y_p) - F_0^*(y_k)$;
      $\beta$ := $F_1^*(y_p) - F_1^*(y_k)$;
      m := 0;
      if( $\alpha$ != 0 )
          m := m + $\frac{1}{2}\alpha \log_2 \frac{\alpha}{\alpha+\beta}$;
      if( $\beta$ != 0 )
          m := m + $\frac{1}{2}\beta \log_2 \frac{\beta}{\alpha+\beta}$;

      MI := MI + m;

      // Reset vars for next loop
      $y_p$ := $y_k$;
      inner_min := $y_p$;
      inner_max := $y_p$;
  } // end while()

  MI := 1 + MI/2;

} // end procedure

Output: MI holds the final, minimal MI value
        $F_0^*(z_i)$ and $F_1^*(z_i)$ hold solution $\forall i : 0 \leq i \leq n+1$
```

References

Bialek, W., Rieke, F., Steveninck, R. de Ruyter van, & Warland, D. (1991). Reading a neural code. *Science*, *252*, 1854–7.

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nat. Neurosci.*, *2*, 947–57.

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge Univ. Press.

Buračas, G. T., Zador, A. M., DeWeese, M. R., & Albright, T. D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron*, *20*, 959–69.

Cantelli, F. P. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.*, *4*, 221–424.

Chechik, G., Anderson, M. J., Bar–Yosef, O., Young, E. D., Tishby, N., & Nelkin, I. (2006). Reduction of information redundancy in the ascending auditory pathway. *Neuron*, *51*, 359–68.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). New Jersey: Wiley.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov-Smirnov theorems. *Ann. Math. Stat.*, *20*, 393–403.

Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern recognition*. New York: John Wiley & Sons.

Dvoretzky, A., Kiefer, J., & Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Stat.*, *27*, 642–69.

Elger, C. E., & Lehnertz, K. (1998). Seizure prediction by nonlinear time series analysis of brain electrical activity. *Eur. J. Neurosci.*, *10*, 786–9.

Gastpar, M., Gill, P., Huth, A., & Theunissen, F. (2010). Anthropic correction of information estimates and its application to neural coding. *IEEE Trans. Inf. Theory*, *56*(2), 890–900.

Georgopoulos, A. P., Kalaska, J. F., Caminiti, R., & Massey, J. T. (1982). On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *J. Neurosci.*, *2*(11), 1527–37.

Gerstner, W., & Kistler, W. (2002). *Spiking neuron models: single neurons, populations, plasticity*. Cambridge: Cambridge Univ. Press.

Glivenko, V. I. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.*, *4*, 92–99.

Globerson, A., Stark, E., Vaadia, E., & Tishby, N. (2009). The minimum information principle and its application to neural code analysis. *PNAS*, *106*, 3490–5.

Gollisch, T., & Meister, M. (2008). Rapid neural coding in the retina with relative spike latencies. *Science*, *391*, 1108–11.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *J. Am. Stat. Assoc.*, *58*, 13–30.

Kennel, M. B., Shlens, J., Abarbanel, H. D. I., & Chichilnisky, E. J. (2005). Estimating entropy rates with bayesian confidence intervals. *Neural Computation*, *17*, 1531–76.

Kim, J., Fisher, J. W., Yezzi, A., Çetin, M., & Willsky, A. S. (2005). A nonparametric statistical method for image segmentation using information theory and curve evolution. *IEEE Trans. Image Processing*, *14*, 1486–1502.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giorn. Ist. Ital. Attuari.*, *4*, 83–91.

Laubach, M. (2004). Wavelet-based processing of neuronal spike trains prior to discriminant analysis. *J. Neurosci. Meth.*, *134*, 159–68.

Learned-Miller, E., & DeStefano, J. (2008). A probabilistic upper bound on differential entropy. *IEEE T. Inform. Theory*, *54*, 5223–30.

Lee, J. H., Russ, B. E., Orr, L. E., & Cohen, Y. (2009). Prefrontal activity predicts monkeys' decisions during an auditory category task. *Front. Integr. Neurosci*, *3*(16).

Lopez-Poveda, E. A., O'Mard, L. P., & Meddis, R. (2001). A human nonlinear cochlear filterbank. *J. Acoust. Soc. Am.*, *110*, 3107–18.

Massart, P. (1990). The tight constant in the Dvoretsky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, *18*, 1269–83.

McDiarmid, C. (1989). On the method of bounded differences. In J. Siemons (Ed.), *Surveys in combinatorics* (pp. 148–88). Cambridge: Cambridge Univ. Press.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *J. Acoust. Soc. Am.*, *79*, 702–11.

Meddis, R. (1988). Simulation of auditory-neural transduction: Further studies. *J. Acoust. Soc. Am.*, *83*, 1056–63.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *J. Acoust. Soc. Am.*, *123*, 899–909.

Middlebrooks, J. C., Clock, A. E., Xu, L., & Green, D. M. (1994). A panoramic code for sound location by cortical neurons. *Science*, *264*, 842–4.

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

Miller, L. M., Escabi, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J. Neurophysiol.*, *87*, 516–27.

Montemurro, M. A., Senatore, R., & Panzeri, S. (2007). Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Computation*, *19*, 2913–57.

Nelkin, I., & Chechik, G. (2005). Estimating stimulus information by spike numbers and mean response time in primary auditory cortex. *J. Comput. Neurosci.*, *19*, 199–221.

Nemenman, I., Bialek, W., & Steveninck, R. de Ruyter van. (2004, May). Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E*, *69*(5), 056111.

Nicolelis, M. A. L., Ghazanfar, A. A., Stambaugh, C. R., Oliveira, L. M. O., Laubach, M., Chapin, J. K., et al. (1998). Simultaneous encoding of tactile information by three primate cortical areas. *Nat. Neurosci.*, *1*, 621–30.

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

Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol*, *98*(3), 1064–72.

Panzeri, S., Treves, A., Schultz, S., & Rolls, E. T. (1999). On decoding the responses of a population of neurons from short time windows. *Neural Computation*, *11*, 1553–77.

Quiroga, R. Q., Arnhold, J., Lehnertz, K., & Grassberger, P. (2000, Dec). Kullback-leibler and renormalized entropies: Applications to electroencephalograms of epilepsy patients. *Phys. Rev. E*, *62*(6), 8380–8386.

Richmond, B. J., Optican, L. M., Podell, M., & Spitzer, H. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. *J. Neurophysiol.*, *57*, 132–46.

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

Samengo, I. (2002). Information loss in an optimal maximum likelihood decoding. *Neural Computation*, *14*, 771–9.

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

Shlens, J., Kennel, M. B., Abarbanel, H. D. I., & Chichilnisky, E. J. (2007). Estimating information rates with confidence intervals in neural spike trains. *Neural Computation*, *19*, 1683–1719.

Shorack, G. R., & Wellner, J. A. (1986). *Empirical processes with applications to statistics*. New York: Wiley.

Smirnov, N. V. (1944). Approximate laws of distribution of random variables from empirical data. *Uspekhi Mat. Nauk.*, *10*, 179–206. (In Russian)

Strong, S., Koberle, R., de Ruyter van Steveninck, R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Phys. Rev. Lett.*, *80*, 197–200.

Sumner, C. J., Lopez-Poveda, E. A., O'Mard, L. P., & Meddis, R. (2002). A revised model of the inner-hair cell and auditory nerve complex. *J. Acoust. Soc. Am.*, *111*, 2178–88.

Takeuchi, J. (1993). Some improved sample complexity bounds in the probabilistic PAC learning model. In S. Doshita, K. Furukawa, K. Jantke, & T. Nishida (Eds.), *Algorithmic learning theory* (Vol. 743, pp. 208–219). Berlin Heidelberg: Springer.

Victor, J. D. (2002). Binless strategies for the estimation of information from neural data. *Phys. Rev. E*, *66*, 51903–51918.

Victor, J. D. (2006). Approaches to information-theoretic analysis of neural activity. *Biol. Theory*, *1*, 302–16.

Warland, D. K., Reinagel, P., & Meister, M. (1997). Decoding visual information from a population of retinal ganglion cells. *J. Neurophysiol.*, 2336–50.

Wolpert, D., & Wolf, D. (1995). Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E*, *52*, 6841–54.