Estimating Mutual Information from Average Misclassification Error

Charles Zheng and Yuval Benjamini

March 16, 2016

Abstract

Mutual information is a useful measure of dependence between the input of a neural subsystem, X, and its output, or measured output, Y, due to its flexibility for capturing nonlinear associations, and its rich information-theoretic context. In high-dimensional settings, non-parametric estimators of mutual information perform poorly, and the only remaining options for estimating mutual information are to either make a parametric assumption (such as multivariate Gaussianity) or obtain a lower bound on the mutual information $I(X;Y)$ from the estimated mutual information between X and an estimator or classification rule, $I(X;\hat{X}(Y))$. But, assuming multivariate Gaussianity reduces mutual information to a linear correlation-based statistic; meanwhile, lower bounds based on $I(X;\hat{X}(Y))$ tend to be overconservative. We propose a new estimator of mutual information based on the concept of "averaged Bayes error," which we prove to be asymptotically a function of the mutual information. The average Bayes error can be estimated using the average misclassification rate from random classification tasks; our estimation procedure consists of plugging in such an estimate of average Bayes error into the inverse of the asymptotic formula. We demonstrate the utility of our method in obtaining accurate estimates of mutual information in simulated data, as well as an fMRI dataset, without having to make parametric assumptions.

1 Introduction

Understanding the amount and type of 'information' that flows from sensory inputs to the various components and subcomponents of the human brain is a key motivation for neuroimaging studies. However, it is not straightforward to decide how exactly to quantify 'information' in the human brain, nor whether quantification should even be attempted, depending on the goals of the experiementer, the nature of the data collected, the complexity of the stimuli and the subsystems being studied. In practice, neuroscientists have a large repetoire of methods for quantifying information: these methods can be subdivided into methods based on linear dependency, such as correlation, multivariate R-squared, or signal-to-noise ratio; and methods based on nonlinear dependency such as mutual information, Fisher information, and prediction error.

Different methods tend to be used in different settings due to tradeoffs between interpretability, flexibility, and high-dimensional scalability. Linear measures of dependence such as correlation is highly interpretable and scalable, but are inflexible: such measures fail to capture many forms of nonlinear dependence. It is a well-established fact that the relationship between the level of chemical and electrical input to a neuron, and the level of its corresponding output is nonlinear; however, it is reasonable to approximate the input-output relationship with a linear relationship, and such approximations may be especially appropriate if signals are averaged at a population level.

Mutual information is interpretable and flexible, but it is generally difficult to estimate in high-dimensional settings. Fisher information is easier to estimate when the stimulus X is low-dimensional, and the response Y is high-dimensional; however, for p-dimensional X the Fisher information takes the form of a $p \times p$ matrix, rendering interpretation difficult when p is large. In cases when both the stimulus space and the response space are high-dimensional a powerful approach is to apply supervised learning methods to either predict X based on Y (decoding), or Y based on X (encoding). Of course, predictive models are extremely interesting beyond the measure of prediction error: one commonly examines the fitted model to find clues to the underlying dynamics of the system. However, one is still often interested in a one-dimensional summary of the dependence structure: in that regard, while prediction error and mutual information are both interpretable, mutual information has the added advantage of its rich context in information

theory, while prediction error has the disadvantage of the arbitrariness of the loss function. Even when considering misclassification error, one often faces the problem of how to partition a high-dimensional space into discrete classes. Furthermore, the ideal definition of prediction error is the *Bayes error*, the prediction error of the optimal rule, but obtaining the Bayes error depends on having the correct model, as well as having infinite data to fit the model. That said, in many problems it may be more feasible to estimate the Bayes error than to obtain a fully nonparametric estimate of the mutual information, since we can easily exploit prior knowledge about the dependence structure between x and Y (for instance, a generalized linear model) to train the predictive model, while nonparametric estimators of mutual information fail to exploit this prior knowledge.

In fact, one could exploit the strong relationship between mutual information and the prediction error to obtain an estimate of the mutual information from the observed classification rate. For example, using generalizations of Fano's inequality, one can obtain a lower bound on mutual information in relation to the optimal prediction error, or Bayes error. Such a technique for obtaining estimates of mutual information from classification rates can be understood as a way to leverage the prior information about X and Y implied by the prediction model in order to obtain a *model-based* estimate of mutual information, $\hat{I}(X;Y)$. However, a prominent challenge to such an approach is the *finite sample bias* resulting from having a limited number of observations N for training and testing the prediction rule.

However, in low-SNR settings, which are commonly encountered in applications, we find that the connection between mutual information and prediction error in the form of misclassification rate, can be made even stronger than the lower bound implied by Fano's inequality. In a particular low-SNR regime, we find an exact asymptotic relationship between the Bayes misclassification probability and the mutual information. Furthermore, our framework allows us to characterize the discrepancy between the observed misclassification rate, and the Bayes error, which allows us to derive that the sample complexity of estimating the mutual information.

1.1 Motivation

The specific setup we consider was motivated by a number of studies

- Face recognition in monkeys

- Identification of natural images

1.2 Setup

Assume X and Y are real random vectors with the same dimensionality, d. Our results are derived under a model where X has a continuous density $p(x)$, but in which the experimenter observes multiple repeats of Y conditional on a common X. The data therefore consists of tuples $(x^{(i)}, y^{(i,1)}, \ldots, y_i^{(i,r)})$, where $x^{(i)}$ is the ith unique stimulus, and $y^{(i,1)}, \ldots, y^{(i,r)}$ are the repeats of Y given $X = x^{(i)}$, which are assumed to be conditionally independent given X.

Let K denote the number of unique stimuli. The data therefore consists of $n = Kr$ observations. When r is large, the data can be nearly considered as i.i.d. observations from the joint distribution of $(\tilde{X}, \tilde{Y})$, where $\tilde{X}$ has a distribution $\tilde{p}$ consisting of a mixture of point masses at $x^{(i)}$:

$$\tilde{p} = \frac{1}{K} \sum_{i=1}^{K} \delta_{x^{(i)}},$$

and $\tilde{Y}|\tilde{X} = x^{(i)}$ has the same distribution as $Y|X = x^{(i)}$.

Yet, although our data was collected from the distribution $(\tilde{X}, \tilde{Y})$, our goal is to estimate $I(X;Y)$ rather than $I(\tilde{X};\tilde{Y})$. In order for the two quantities to have any connection, the selected stimuli $x^{(i)}$ must be 'representative' of the continuous distribution X. When the stimulus X is very high-dimensional, it becomes quite reasonable to draw $x^{(i)}$ i.i.d. from the marginal distribution $p(x)$. This ensures that $I(\tilde{X};\tilde{Y})$ converges to $I(X;Y)$ as $K \to \infty$. Though, as noted by Gastpar et. al., for finite K, $I(\tilde{X};\tilde{Y})$ tends to result in an underestimate of $I(X;Y)$. This motivates their antropic correction method for estimating $I(X;Y)$, which can be applied directly in this setting supposing that one has a method for estimating the conditional entropies $H(Y|X = x^{(i)})$.

In contrast, we will consider the misclassification errror as a means to estimate the mutual information. Letting $p(x, y)$ denote the density of (X, Y), the Bayes rule for predicting $\tilde{X}$ from $\tilde{Y} = y^*$ is given by

$$\hat{X}_{Bayes} = \text{argmax}_{x=x^{(1)},\ldots,x^{(k)}} \log p(y^*|x)$$

where $p(y|x) = p(x, y)/p(x)$. The Bayes error is

$$\Pr[\tilde{X} \neq \hat{X}_{Bayes}],$$

where the probability is taken over the joint distribution of $(\tilde{X}, \tilde{Y})$. Since the Bayes error depends on the sample of representative stimuli $\{x^{(i)}\}$, we find it more useful to consider the average Bayes error:

$$\mathrm{MC} = \mathbf{E}[\Pr[\tilde{X} \neq \hat{X}_{Bayes}]],$$

where the outer expectation is over the distribution of $x^{(i)} \sim p(x)$. The following sections explore the relationship between MC and $I(X;Y)$.

As a means to estimate the average Bayes error MC, we fit a predictive model for $\tilde{X}$ given $\tilde{Y}$. This results in a K-class classification problem. While in practice, a variety of multi-class classification methods can be employed, our theory depends on having a known, semiparametric generative model for the conditional distribution of Y: we study the misclassification rate obtained by using the maximum-likelihood plugin estimate of the Bayes rule.

Hence, when deriving sample complexity results, we make the further assumptions that

$$p(x, y) = p(x)q(y|\mu(x))$$

where μ is an unknown bijection from $\mathbb{R}^p \to \mathbb{R}^p$, and $q(y|\mu)$ is a known parametric family of density functions which are jointly differentiable in y and μ. The model is semiparametric since we do not make any constraints on the function μ, other than invertibility. In fact, X can be removed from the picture since $I(X;Y) = I(\mu;Y)$, where $\mu = \mu(X)$. This reflects practice in many neuroimaging studies where the actual pixel values of the stimuli are not incorporated in the model at all; rather, one simply models the joint distribution of the class of the stimulus and the response. On the other hand, it is worth noting that the model-based approach demonstrated in Kay et al., and others, do model the mapping μ.

In order to get an estimate of the misclassification rate, one has to *hold out* a number r_{test} of the repeats from each class. The classification rule is based on estimates of $\mu^{(i)} = \mu(x^{(i)})$, given by the MLE estimator on the training set,

$$\hat{\mu}^{(i)} = \mathrm{argmax}_{\mu} \sum_{j=1}^{r_{train}} \log q(y^{(i,j)}|\mu).$$

The MLE classification rule is therefore defined as

$$\hat{X}_{MLE} = x^{(i)} \text{ where } i = \mathrm{argmax}_i \log q(y^*|\hat{\mu}^{(i)}).$$

The sample test error is therefore

$$\frac{1}{Kr_{test}} \sum_{i=1}^{K} \sum_{j=r_{train}+1}^{r} I(\hat{x}_{MLE}^{(i,j)} \neq x^{(i)}).$$

As an estimate of MC, the sample test error has variability both from the randomness in $\tilde{Y}$ conditional on the sampled stimuli $x^{(i)}$, and from the randomness in the sampled stimuli drawn from $p(x)$. Therefore, it makes sense to repeat the procedure for m independent *samples* of $(x^{(1)}, \ldots, x^{(K)})$, and then averaging the resulting test errors. Let the resulting average misclassification rate be denoted $\hat{\text{MC}}$. In later sections we will study the discrepancy between $\hat{\text{MC}}$ and MC, and how to optimally choose the experimental parameters K and r given a total budget of $N = Kmr$ observations.

2 Theory

2.1 Application of classical results

- Using Fano's inequality
- Limitations
- Define $\tilde{X}$ to be the discretization of X
- Define $I(F)$ to be the mutual information $I(X;Y)$ when $(X,Y) \sim F$.

2.2 Low-SNR model

We have seen in the previous section that the lower bound implied by Fano's inequality is quite inaccurate when (...). Certainly, an exact relationship between $I(X;Y)$ and the Bayes error cannot hold since given two different joint distibutions F, G with $I(F) = I(G)$, the K-class misclassification rate MC may be quite different between F and G. Yet, we observe that under the two conditions that (i) the dimensionality of (X,Y) is high, and (ii) the signal-to-noise ratio is low, in the sense that $H(X,Y) >> I(X;Y)$, the relationship between information and misclassification rate begins to cohere.

- Give a counterexample (gaussian)

- Plot of MC depending on K.
- Examples of low SNR regime. Varying $I(X;Y)$ and also dimensionality
- In all plots, compare with Fano inequality
- As we can see, low SNR regime gets more accurate than Fano

Assume that X and Y have joint density $p(x,y)$ with respect to Lesbegue measure on $\mathbb{R}^{2d}$. Draw i.i.d. $(X^{(i)}, Y^{(i)})$ from the joint distribution, for $i = 0, \ldots, K-1$, and let (X^*, Y^*) denote $(X^{(0)}, Y^{(0)})$. Define

$$Z_i = \log p(Y^*|X_i) = \log p(Y^*, X_i) - \log p(X_i).$$

The Bayes rule is therefore

$$\hat{X} = x^{(i)} \text{ where } i = \text{argmax}_i Z_i$$

It turns out the reason why the dimensionality and signal-to-noise ratio play a role is because those conditions ensure that the vector $Z = (Z_*, Z_1, \ldots, Z_{K-1})$ has an approximately normal distribution. However, to formally prove this fact, we require an asymptotic framework.

We consider a limiting sequence of problems of increasing dimensionality d. Let $(X^{[d]}, Y^{[d]})$ denote the joint distributions in the sequence, for $d \in \{1, 2, \ldots\}$. As d increases, the ratio of the information $I(X^{[d]}; Y^{[d]})$ and the joint entropy $H(X^{[d]}, Y^{[d]})$ decreases.

2.2.1 Gaussian Example

Before giving a general result, we illustrate this aymptotic regime by the following gaussian example. Let

$$\begin{bmatrix} X^{[d]} \\ Y^{[d]} \end{bmatrix} \sim N\left(0, \begin{bmatrix} I & \frac{1}{\sqrt{1+d\sigma^2}} I \\ \frac{1}{\sqrt{1+d\sigma^2}} I & I \end{bmatrix}.\right)$$

For fixed d, we have $(X_i^{[d]}, Y_i^{[d]})$ drawn i.i.d. from a bivariate normal $N(0, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$ where $\rho = (1 + d\sigma^2)^{-\frac{1}{2}}$. Recalling that the mutual information

of the components of such a bivariate normal is $-\log(1-\rho^2)/2$, we easily calculate:

$$I(X^{[d]}, Y[d]) = \sum_{i=1}^{d} I(X_i^{[d]}, Y_i^{[d]}) = -\frac{d}{2}\log(1 - \frac{1}{1+d\sigma^2}).$$

Let ι denote the limit of the mutual information as $d \to \infty$: we have

$$\begin{aligned}\iota &= \lim_{d\to\infty} I(X^{[d]}, Y^{[d]}) = \lim_{d\to\infty} -\frac{d}{2}\log(1 - \frac{1}{1+d\sigma^2}) \\ &= \lim_{d\to\infty} \frac{d}{2}\frac{1}{1+d\sigma^2} = \frac{1}{2\sigma^2}.\end{aligned}$$

Meanwhile, $H(X^{[d]}) = H(Y^{[d]}) = \frac{d}{2}\log(2\pi)$, so it is clear that $H(X^{[d]}, Y^{[d]}) >> I(X^{[d]}; Y^{[d]})$.

A simple calculation shows that

$$Z_i = \log p(Y^*|X^{(i)}) = -\frac{1}{2(1-\rho^2)}||Y^* - \rho X^{(i)}||^2 + C_\rho$$

where the first term is a scaled chi-squared distribution with d degrees of freedom: the scale is $-\frac{1}{2}\frac{1+\rho^2}{1-\rho^2}$ for $i = 1, \ldots, K-1$ and $-1/2$ for $i = 0$. The omitted constant is

$$C_\rho = -\frac{1}{2}\log(2\pi(1-\rho)^2)$$

Since we can separate Z_i into independent, componentwise sums,

$$Z_i = C_\rho - \frac{1}{2(1-\rho^2)}\sum_{j=1}^{d}(Y_j^* - \rho X_j^{(i)})^2,$$

it follows from the multivariate central limit theorem that Z_i are asymptotically jointly normal.

A straightforward computation using multivariate normal moments (c.f.

Muirhead) yields the limiting moments:

$$\begin{aligned}
\mathbf{E}[Z_*] &= -\frac{d}{2} + C_\rho \\
\mathbf{E}[Z_i] &= -\frac{d}{2}\frac{1+\rho^2}{1-\rho^2} + C_\rho \\
\mathrm{Var}[Z_*] &= \mathrm{Cov}(Z_*, Z_i) = \frac{d}{2} \\
\mathrm{Var}[Z_i] &= \frac{d}{2}\frac{(1+\rho^2)^2}{(1-\rho^2)^2} \\
\mathrm{Cov}[Z_i, Z_j] &= \frac{d}{2}\frac{1}{(1-\rho^2)^2}
\end{aligned}$$

for $i \neq j \neq 0$. Taking limits, the moments simplify to yield

$$\begin{bmatrix} Z_* \\ Z_1 \\ \vdots \\ Z_{K-1} \end{bmatrix} \xrightarrow{d} N\left(\begin{bmatrix} C_0 - \frac{d}{2} \\ C_0 - \frac{d}{2} - \frac{1}{\sigma^2} \\ \vdots \\ C_0 - \frac{d}{2} - \frac{1}{\sigma^2} \end{bmatrix}, \begin{bmatrix} \frac{d}{2} & \frac{d}{2} & \cdots & \frac{d}{2} \\ \frac{d}{2} & \frac{d}{2} + \frac{2}{\sigma^2} & \cdots & \frac{d}{2} + \frac{1}{\sigma^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d}{2} & \frac{d}{2} + \frac{1}{\sigma^2} & \cdots & \frac{d}{2} + \frac{2}{\sigma^2} \end{bmatrix} \right),$$

where $C_0 = -\log(2\pi)/2$. By the central limit theorem, the misclassification probability is

$$\mathrm{MC} = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

for a random multivariate normal vector $(Z_*, Z_1, \ldots, Z_{K-1})$ with the given mean and covariance matrix. It is worth noting that the probability $\Pr[Z_* < \max_{i=1}^{K-1} Z_i]$ directly gives the *averaged* Bayes error: indeed, in high dimensions it is not trivial to compute the Bayes error for fixed configuration. To obtain a simplified expression of this multivariate normal probability, we employ the following lemma.

Lemma. *Suppose $(Z_*, Z_1, \ldots, Z_{K-1})$ are jointly multivariate normal, with $\boldsymbol{E}[Z_* - Z_1] = \alpha$, $Var(Z_*) = \beta$, $Cov(Z_*, Z_i) = \gamma$, $Var(Z_i) = \delta$, and $Cov(Z_i, Z_j) = \epsilon$ for all $i, j = 1, \ldots, K-1$. Then, letting*

$$\mu = \frac{\boldsymbol{E}[Z_* - Z_i]]}{\sqrt{\frac{1}{2}\,Var(Z_i - Z_j)}} = \frac{\alpha}{\sqrt{\delta - \epsilon}},$$

$$\nu^2 = \frac{Cov(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\,Var(Z_i - Z_j)} = \frac{\beta + \epsilon - 2\gamma}{\delta - \epsilon},$$

we have

$$\begin{aligned}\Pr[Z_* < \max_{i=1}^{K-1}] &= \Pr[W < M_{K-1}] \\ &= 1 - \int -\frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(w-\mu)^2}{2\nu^2}} (1-\Phi(w))^{K-1} dw,\end{aligned}$$

where $W \sim N(\mu, \nu^2)$ *and* M_{K-1} *is the maximum of* $K-1$ *independent standard normal variates, which are independent of* W*.*

(See appendix for proof.)

Applying the lemma, we compute

$$\mu = \frac{\sigma^{-2}}{\sqrt{\sigma^{-2}}} = \frac{1}{\sigma},$$

$$\nu^2 = \frac{\sigma^{-2}}{\sigma^{-2}} = 1,$$

hence

$$\mathrm{MC} = \Pr[N(\frac{1}{\sigma}, 1) < M_{K-1}].$$

Defining the function $f_K(\mu) = \Pr[N(\mu, 1) < M_{K-1}]$, we therefore get

$$\mathrm{MC} = f_K\left(\frac{1}{\sigma}\right) = f_K(\sqrt{2\iota}),$$

recalling that ι is the limiting value of the mutual information.

Hence we obtain a fairly explicit relationship between average Bayes error MC and limiting mutual information in the gaussian case. In the following section, we see that the formula

$$\mathrm{MC} = f_K(\sqrt{2\iota})$$

applies more generally!

2.2.2 Generalization

How far can we generalize the previous example? For starters, we can allow $(X_i^{[d]}, Y_i^{[d]})$ to have a non-Gaussian bivariate distribution, with a density with respect to Lesbegue measure. However, we still require that $(X_i^{[d]}, Y_i^{[d]})$ are i.i.d. for $i = 1, \ldots, d$. We let $b_d(x, y)$ denote the bivariate joint density of

$(X_i^{[d]}, Y_i^{[d]})$, and $b(x)$ and $b(y)$ to denote the marginal distributions of $b_d(x, y)$, which are also assumed to be fixed. The independence allows us to decompose the mutual information

$$I(X^{[d]}, Y^{[d]}) = \sum_{i=1}^{d} I(X_i^{[d]}, Y_i^{[d]}) = dI(X_i^{[d]}.Y_i^{[d]}).$$

Given some additional conditions on the marginal bivariate distributions $b_d(x, y)$, we can conclude joint asymptotic normality of all of the quantities $\log p(X^{(i)}, Y^{(j)}$ for $i, j = 1, \ldots, K$.

Now consider what happens if the total mutual information stays fixed, $I(X^{[d]}.Y^{[d]}) = c$, while the dimensionality increases. Since $I(X_i^{[d]}.Y_i^{[d]}) = c/d$, and since the marginals are fixed, we conclude that the density functions $b_d(x, y)$ are converging to the product $b(x)b(y)$. Now if we define

$$u_d(x, y) = \frac{b_d(x, y)}{b(x)b(y)} - 1,$$

we can say that $u_d(x, y) \to 0$ as $d \to \infty$.

It turns out that in such a limit, the moments of $u_d(X_i^{(j)}, Y_i^{(k)})$ determines many of the information theoretic-quantities of interest. From the definition, we have

$$0 = \mathbf{E}[u(X_i, Y_i)|X_i] = \mathbf{E}[u(X_i, Y_i)|Y_i] = \mathbf{E}[u(X_i^{(j)}, Y_i^{(k)})]$$

for $j, k \in \{1, \ldots, K\}$. Then observe that

$$\begin{aligned}
-H(X_1, Y_1) &= \int \log(b(x,y))b(x,y)dxdy \\
&= \int \log(b(x)b(y)(1+u(x,y))b(x)b(y)(1+u(x,y))dxdy \\
&= \int \log(b(x))b(x) \left[\int b(y)(1+u(x,y))dy\right] dx \\
&\quad + \int \log(b(y))b(y) \left[\int b(x)(1+u(x,y))dx\right] dx \\
&\quad + \int \log(1+u(x,y))(1+u(x,y))b(x)b(y)dxdy \\
&= \int \log(b(x))b(x)\mathbf{E}[1+u(X,Y)|X=x]dx \\
&\quad + \int \log(b(y))b(y)\mathbf{E}[1+u(X,Y)|X=y]dx \\
&\quad + \mathbf{E}[\log(1+u(X_1,Y_*))(1+u(X_1,Y_1^*))] \\
&= -H(X_1) - H(Y_1) + \mathbf{E}[\log(1+u(X_1,Y_1^*))(1+u(X_1,Y_1^*))]
\end{aligned}$$

where here X_1, Y_1^* are drawn from the product marginal $b(x)b(y)$. Hence

$$I(X_1; Y_1) = \mathbf{E}[\log(1+u(X_1,Y_1^*))(1+u(X_1,Y_1^*))].$$

Since $I(X_1; Y_1)$ is 'small'–order $O(1/d)$, the function $u(x,y)$ must become 'small' in some sense as well, as d grows. Assume for the moment that we can justifiably replace $\log(1+u(x,y))$ with its second-order Taylor expansion,

$$\log(1+u_d(x,y)) \approx u_d(x,y) - \frac{1}{2}u_d(x,y)^2.$$

Then we get

$$I(X_1; Y_1) \approx \mathbf{E}\left[u_d(X_1,Y_1^*) + \frac{1}{2}u_d(X_1,Y_1)^2 - \frac{1}{2}u_d(X_1,Y_1^*)^3\right],$$

which, since $\mathbf{E}[u_d(X_1,Y_1^*)] = 0$, and since we have been neglecting third-order terms, gives

$$I(X_1; Y_1) \approx \frac{1}{2}\mathbf{E}[u_d(X_1,Y_1^*)^2] = \frac{1}{2}\mathrm{Var}[u_d(X_1,Y_1^*)].$$

Many other similar identities occur in the following proof–all of which depend on neglecting higher-order terms of u_d. But when can we justifiably ignore terms of the form $u_d(X_i, Y_i^*)^k$? Ideally, we need $\mathbf{E}[|u_d(X_i, Y_i^*)|^k]$ uniformly bounded by $O(d^{-1+\epsilon})$ for $k \geq 3$. However, in order to conclude such a result, it is necessary to assume a few regularity conditions. For instance, it suffices to assume that

????

We mention it to give a concrete example of a sufficient condition for being able to neglect higher moments of $u_d(X_i, Y_i^*)$.

Collected together, we will assume the following:

A1. For all d, $I(X^{[d]}, Y^{[d]}) = \iota$.

A2. For all d, d' and $i \leq d$, $j \leq d'$, we have $X_i^{[d]}$ is equal to $X_j^{[d']}$ in distribution, $Y_i^{[d]}$ is equal to $Y_j^{[d']}$ in distribution. Let $b(x)$ denote the marginal distribution of X_j and $b(y)$ denote the marginal distribution of Y_j.

A3. Assume X_j and Y_j have finite third moments.

A4. For each $d = 1, 2, \ldots$, the components $(X_i^{[d]}, Y_i^{[d]})$ are drawn i.i.d. from a bivariate density $b_d(x, y)$ for $i = 1, \ldots, d$.

A6. Defining $u_d(x, y) = \frac{b_d(x,y)}{b(x)b(y)}$, we have $\mathbf{E}[|u_d(X_i, Y_i^*)^3|] = O(d^{-1-\epsilon})$, where $\epsilon > 0$.

Theorem. *Let $X^{[d]}, Y^{[d]}$ be a sequence of distributions satisfying assumptions A1-A6. Then, as $d \to \infty$, the misclassification probability*

$$MC = \Pr[Z_* < \max_{i=1}^{K-1} Z_i]$$

satisfies

$$\lim_{d\to\infty} MC = f_K(\sqrt{2\iota}),$$

where f_K is defined in Lemma.

Proof. As mentioned in the preceding discussion, assumption A6 allows us to find some $\epsilon > 0$, allowing us to write

$$I(X_1; Y_1) = \frac{1}{2}\mathrm{Var}[u(X_1, Y_1^*)] + O(d^{-1-\epsilon}).$$

Furthermore, we conclude that

$$\mathrm{Cov}(u_d(X_i^{(j)}, Y_i^*)^2, u_d(X_i^{(k)}, Y_i^*)) = O(d^{-1-\epsilon})$$

$$\mathrm{Cov}(u_d(X_i^{(j)}, Y_i^*), u_d(X_i^{(k)}, Y_i^*)) = O(d^{-1-\epsilon})$$

for all $j, k = 0, \ldots, K-1$. And for $i \neq j \neq 0$, we have

$$\begin{aligned}\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^{(j)}, Y_1^*)) =& \mathrm{Cov}(\mathbf{E}[u_d(X_1^{(i)}, Y_1^*)|Y_1^*], \mathbf{E}[u_d(X_1^{(j)}, Y_1^*)|Y_1^*]) \\ &+ \mathbf{E}[\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^{(j)}, Y_1^*))|Y_1^*] \\ &= \mathrm{Cov}(0,0) + \mathbf{E}[0] = 0\end{aligned}$$

due to the identity $\mathbf{E}[u_d(X_1, y)|y] = 0$, and conditional independence, respectively. By a similar argument,

$$\mathrm{Cov}(u_d(X_1^{(i)}, Y_1^*), u_d(X_1^*, Y_1^*)) = 0.$$

Noting that $\mathbf{E}[u_d(X_1, Y_1^*)^2] = 2\iota/d$, we can compute

$$\begin{aligned}\mathbf{E}[u_d(X_1^*, Y_1^*)] &= \int u_d(x,y)b(x,y)dxdy \\ &= \int u_d(x,y)(1+u_d(x,y))b(x)b(y)dxdy \\ &= \mathbf{E}[u_d(X_1, Y_1^*)] + \mathbf{E}[u_d(X_1, Y_1^*)^2] \\ &= 0 + 2\iota/d = 2\iota/d\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}[u_d(X_1^*, Y_1^*)^2] &= \int u_d(x,y)^2(1+u_d(x,y))b(x)b(y)dxdy \\ &= \mathbf{E}[u_d(X_1, Y_1^*)^2] + \mathbf{E}[u_d(X_1, Y_1^*)^3] \\ &= 2\iota/d + O(d^{-1-\epsilon}),\end{aligned}$$

hence

$$\mathrm{Var}(u_d(X_1^*, Y_1^*)) = \frac{2\iota}{d} - \frac{4\iota^2}{d^2} + O(d^{-1-\epsilon}) = \frac{2\iota}{d} + O(d^{-1-\epsilon})$$

Due to componentwise independence, the scores $Z_i = \log p(Y^*|X^{(i)}$ converge in distribution to a multivariate normal. Now let us compute the

moments of Z_i:

$$\begin{aligned}\mathbf{E}[Z_1] &= d\mathbf{E}[\log b(X_1, Y_1^*) - \log b(X_1)] \\ &= d\mathbf{E}[\log b(Y_1^*) + u(X_1, Y_1^*) - u(X_1, Y_1^*)^2/2] + O(d^{-\epsilon}) \\ &= -H(Y) - I(X;Y) + O(d^{-\epsilon}).\end{aligned}$$

Meanwhile, we know that

$$\mathbf{E}[Z_*] = \mathbf{E}[\log p(Y^*|X^*)] = H(X) - H(X,Y),$$

hence

$$\mathbf{E}[Z_* - Z_i] = 2I(X;Y) + O(d^{-\epsilon}) = 2\iota + O(d^{-\epsilon}).$$

We get

$$\begin{aligned}\mathrm{Var}(Z_i - Z_j) &= d\mathrm{Var}(\log(b_d(Y_1^*|X_1^{(i)})) - \log(b_d(Y_1^*|X_1^{(j)}))) \\ &= d\mathrm{Var}(\log(b_d(X_1^{(i)}, Y_1^*)) - \log(b_d(X_1^{(j)}, Y_1^*) - \log b(X_1^{(i)}) + \log b(X_1^{(j)}))) \\ &= d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*) - u_d(X_1^{(j)}, Y_1^*) - u_d(X_1^{(i)}, Y_1^*)^2/2 + u_d(X_1^{(j)}, Y_1^*)^2/2) \\ &= d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*) - u_d(X_1^{(j)}, Y_1^*)) + O(d^{-\epsilon}) \\ &= 2d\mathrm{Var}(u_d(X_1^{(i)}, Y_1^*)) + O(d^{-\epsilon}) = 4\iota + O(d^{-\epsilon}).\end{aligned}$$

Looking ahead to the application of the Lemma, we can already compute

$$\mu = \lim_{d\to\infty} \frac{\mathbf{E}[Z_* - Z_i]}{\sqrt{\frac{1}{2}\mathrm{Var}(Z_i - Z_j)}} = \sqrt{2\iota}.$$

It remains to compute

$$\begin{aligned}\mathrm{Cov}(Z_* - Z_i, Z_* - Z_j) &= d\mathrm{Cov}(u_d(X_1^*, Y_1^*) - u_d(X_1^{(i)}, Y_1^*), u_d(X_1^*, Y_1^*) - u_d(X_1^{(j)}, Y_1^*)) + O(d^{-\epsilon}) \\ &= d\mathrm{Var}(u_d(X_1^*, Y_1^*)) = 2\iota + O(d^{-1-\epsilon})\end{aligned}$$

We conclude that

$$\nu^2 = \lim_{d\to\infty} \frac{\mathrm{Cov}(Z_* - Z_i, Z_* - Z_j)}{\frac{1}{2}\mathrm{Var}(Z - i - Z_j)} = \frac{2\iota}{2\iota} = 1.$$

Hence, the desired result follows from the asymptotic normality of $(Z_*, Z_1, \ldots, Z_{K-1})$ and Lemma. □