Discriminative Estimation of $f$-Divergence

Le Song  
School of Computer Science  
CMU, Pittsburgh, PA  
lesong@cs.cmu.edu

Mark D. Reid, Robert C. Williamson  
Australian National University  
Canberra, ACT, Australia  
mark.reid,bob.williamson@nicta.com.au

Alex J. Smola  
Yahoo! Research  
Santa Clara, CA  
alex@smola.org

Abstract

We propose an approach for estimating $f$-divergences that exploits a new representation of an $f$-divergence as a weighted integral of cost-weighted Bayes risks. We are therefore able to reduce $f$-divergence estimation to a problem of posterior conditional probability estimation. We provide both batch and online implementation of our approach and analyze their convergence. Empirically, we show that our implementation compares favorably to other $f$-divergence estimators and demonstrate its application to an EEG dataset.

1 Introduction

Being able to compare the difference between distributions is a fundamental operation in statistics and machine learning. A rich class of “distance” relations between two distributions $Q(X)$ and $R(X)$ defined on the same space $X$ is the $f$-divergence, $I_f(Q,R) := \int_X f(dQ/dR)dR$ [1]. Members of this class include the well-known divergences such as the variational, Kullback-Leibler (KL), and Hellinger divergences. The estimation of these divergence measures from observations plays an important role in a diverse range of applications such as clustering [2] and covariate shift correction [3].

One way to estimate $f$-divergences is to directly estimate the densities $dQ$ and $dR$ and then simply apply the definitions. For this purpose, either Parzen windows or space partitioning techniques are needed [4] which makes these direct approaches encounter increasing difficulty as the dimensionality of the problem grows. To address this Nguyen et al. provide a technique based on estimating the density ratio $dQ/dR$ instead [5]. Their method employs a variational representation of $f$-divergences and estimates this ratio via a regularized convex risk minimization. However, this approach still model the space $X$ directly. When the dimensionality of $X$ is high, estimating $dQ/dR$ is still challenging.

We take a different approach and view $f$-divergence estimation in a binary classification framework where observations of the positive class are drawn from $Q$ and those of the negative class from $R$. Intuitively, the better we can classify observations from $Q$ and $R$, the more $Q$ and $R$ are different from each other. More specifically, observations are drawn from a joint distribution $P(X,Y)$ defined on an augmented space $X \times Y (Y = \{-1,1\})$; and now $Q$ and $R$ simply become the class-conditional distribution $P(X|Y = 1)$ and $P(X|Y = -1)$ respectively. Given a prior $P(Y)$ on the class probability and a prior $P(X)$ on the observations from $X$, Bayes’ theorem yields

$$P(X|Y)P(Y) = P(Y|X)P(X).$$  \hspace{1cm} (1)

This relation suggests that estimating $f$-divergence using $P(X—Y)$ terms in the lhs. of (1) can also be achieved using the $P(Y—X)$ terms in the rhs. of (1). We will refer to these complementary views as generative and discriminative, respectively.

Indeed, we show that $f$-divergences can be estimated via the observation-conditional class probability $\eta(X) := P(Y = 1|X)$. Compared to estimation approaches via $dQ$ and $dR$ or the density ratio $dQ/dR$, this new approach is appealing since we only need to estimate the density for a binary variable $Y$, and $\eta$ is well-behaved even when the density $dR$ is small making the density ratio difficult to estimate. Furthermore, our approach also has the additional advantage that it is able to estimate a whole range of different $f$-divergences by estimating $\eta(X)$ only once.

In the remainder of the paper, we will first explain the relation between $f$-divergences and binary classifications in Section 2. Then, we will present our algorithm for estimating $f$-divergences based on this relation in Section 3. Next we will analyze the stability of our estimator (Section 4), and extend it to
online setting (Section 5) and dependence estimation (Section 6). Last we will present experimental investigation of our method (Section 7).

2 Relation between f-Divergences and Binary Classifications

f-Divergences For two distributions \( Q(X) \) and \( R(X) \) defined on the same space \( X \), the \( f \)-divergence of \( Q \) from \( R \) is defined as [1]

\[
I_f(Q, R) := \int_X f \left( \frac{dQ}{dR} \right) dR,
\]

where \( f \in \mathcal{F} : \mathbb{R}^+ \mapsto \mathbb{R} \) is a convex function such that \( f(1) = 0 \). Different choices of \( f \) result in different \( f \)-divergences. For instance, choosing \( f(t) = -\ln(t) \) gives the KL divergence, and \( f(t) = |t - 1| \) yields the variational divergence.

Binary Classifications A binary classification task is characterized by a joint distribution \( P(X,Y) \) where we draw pairs \((X, Y)\) from domain \( X \times Y \). The variable \( Y \in Y = \{-1, 1\} \) is the class label for an observation. Besides the joint distribution \( P(X,Y) \), we may also have two prior distributions \( P(X) \) and \( P(Y) \) over the observation \( X \) and the class label \( Y \) respectively.

Rather than using the joint distribution as the foundation of our analysis we will consider two related representations:

- The generative view that decomposes \( P(X,Y) \) into two class-conditional distributions \( Q := P(X|Y = 1) \) and \( R := P(X|Y = -1) \), and the class label prior probability \( \pi := P(Y = 1) \).
- The discriminative view that decomposes \( P(X,Y) \) into the observation-conditional (or posterior) probability \( \eta(X) := P(Y = 1|X) \), and the observation prior \( P(X) \).

Note that we deliberately use \( Q \) and \( R \) to denote the two class-conditional distributions so that they correspond to the two distributions whose divergence is to be estimated. Furthermore, we note that the \( f \)-divergence of \( Q \) and \( R \) is actually independent of \( \pi \) and so we set it arbitrarily to \( \pi = \frac{1}{2} \).

The generative and discriminative views are related by the Bayes’ theorem \( P(X|Y)P(Y) = P(Y|X)P(X) \). From a discriminative point of view, the binary classification task boils down to estimating the posterior probability \( \eta(X) \). Then a simple classifier can be achieved by thresholding \( \eta(X) \). An attractive property of using a discriminative model is that such models enjoy better convergence properties than generative models whenever the model is not well specified [6].

Cost-weighted Loss The quality of a posterior probability estimator \( \hat{\eta}(X) \) can be assessed using a family of cost-weighted losses \( l_c(\hat{\eta}, Y) \) parameterized by a cost \( c \in [0, 1] \):

\[
l_c(\hat{\eta}(X), Y) := \begin{cases} c[I(Y = -1)], & \text{if } \hat{\eta}(X) \geq c, \\ (1 - c)[Y = 1], & \text{otherwise}, \end{cases}
\]

where \([\cdot]\) is 1 if its argument is true, and is 0 otherwise. Intuitively, a cost-weighted loss thresholds \( \hat{\eta} \) at \( c \) and assigns a cost if the resulting classification disagrees with the label \( Y \). Substituting \( c = \frac{1}{2} \) will verify that \( 2l_{\frac{1}{2}} \) is equivalent to the 0-1 classification loss.

Bayes Risk The conditional risk of the estimator \( \hat{\eta} \) is the expected value of the loss \( l_c(\hat{\eta}, Y) \) conditioned on a fixed value of \( X \), :

\[
r_c(\hat{\eta}, Y|X) := \mathbb{E}_{Y \sim P(Y|X)} [l_c(\hat{\eta}(Y), Y)|X]
\]

\[
= \begin{cases} c(1 - \eta(X)), & \text{if } \hat{\eta}(X) \geq c, \\ (1 - c)\eta(X), & \text{otherwise}. \end{cases}
\]

The infimum of the conditional risk, \( r_c(\eta, X) := \inf_{\hat{\eta}} r_c(\hat{\eta}, Y|X) = \min \{ (1 - \eta)c, (1 - c)\eta \} \), is called the conditional Bayes risk. Note that \( r_c(\eta, X) \) is a concave function of \( c \) since it is the minimum of two linear functions of \( c \).

The (full) Bayes risk is then the expected value of the conditional Bayes risk with respect to the prior \( P(X) \):

\[
r_c^\pi := \mathbb{E}_{X \sim P(X)} [r_c(\eta, X)].
\]

If the posterior \( \eta \) is equal to the prior \( \pi \) for the class labels, i.e. \( X \) is independent of \( Y \) or \( X \) is not informative about \( Y \), the Bayes risk in this case becomes:

\[
r_c^\pi := \mathbb{E}_X [r_c(\pi, X)] = \min \{ (1 - \pi)c, (1 - c)\pi \}.
\]

f-Divergence and Bayes Risk Statistical information is a measure of the reduction of Bayes risk when the observation \( X \) is informative about the class label \( Y \) [7], i.e. when \( \eta \neq \pi \). It is simply defined as:

\[
\Delta r_c^\pi := r_c^\pi - r_c^\eta.
\]

The following theorem is central to our estimation procedure. In essence, it states that \( f \)-divergences can be represented as weighted integrals of the statistical informations \( \Delta r_c^\pi := \min \{ \frac{1}{2}c, \frac{1}{2}(1 - c) \} = \mathbb{E}_X [\min \{ (1 - \eta)c, (1 - c)\eta \}]. \) (For convenience, we will abbreviate \( \Delta r_c^{\frac{1}{2}} \) as \( \Delta r_c \).

**Theorem 1** Let \( f \in \mathcal{F} \) be twice differentiable and \( \gamma(c) := \frac{2}{(1 - c)^2}f'' \left( \frac{c}{1 - c} \right) \). Then for all \( Q \) and \( R \)

\[
I_f(Q, R) = \int_0^1 \Delta r_c \gamma(c) dc
\]
Algorithm 1 Batch Estimation of f-Divergence

1: Estimate the posterior probability η using an exponential model, i.e., \( \hat{\eta} := (1 + \exp(-\theta, \phi(X)))^{-1} \);
2: For each \( c \), estimate \( \Delta r_c \) using sample average, i.e., \( \Delta r_c := \frac{1}{2} \hat{r}_c - \frac{1}{m} \sum_{i=1}^{m} r_c(\hat{\eta}, x_i) \);
3: Estimate the integral in (8) using quasi-Monte Carlo samples, i.e., \( \hat{I}_f := \frac{\Gamma(1 - \eta_0)}{n} \sum_{j=1}^{n} \Delta r_c \).

An important implication of this theorem is that f-divergences can be estimated discriminatively by estimating the posterior probability η.

3 Discriminative Estimation of f-Divergence

Based on Theorem 1, we derive a procedure for estimating f-divergences which contains 3 simple steps as shown in Algorithm 1.

Step 1 In the exponential model, \( \hat{\theta} \) is the natural parameter and \( \phi(X) \) is the sufficient statistic which corresponds to the feature map of a kernel function \( k(X, X') := \phi(X), \phi(X') \). If a universal kernel such as the Gaussian RBF kernel \( \exp(-s||X - X'||^2) \) is used, the exponential model will be rich enough to approximate any smooth \( \eta \) which has a full support [8].

Another reason to choose an exponential model is that it has good convergence guarantees [9]. The natural parameter \( \theta \) can be readily estimated via a regularized convex risk minimization, and the estimate \( \hat{\theta} \) converges to the true parameter \( \theta \) with rate \( O(m^{-\frac{1}{2}}) \) where \( m \) is the number of observations. That is \( \epsilon_{\theta} := \| \theta - \hat{\theta} \| = O(m^{-\frac{1}{2}}) \).

In our analysis, we also assume that the true posterior \( \eta \) is bounded away from 0 and 1 by a small number \( \eta_0 \in (0, \frac{1}{2}) \), i.e. \( \eta \in [\eta_0, 1 - \eta_0] \). Note that this assumption is equivalent to the assumption that the ratio \( dP/dQ \) is bounded away from zero and finite in Nguyen et al. [5]. Specifically, if \( \eta \in [\eta_0, 1 - \eta_0] \) then using \( \eta = \frac{dP}{dQ} \) for \( \pi = \frac{1}{2} \) gives us \( \frac{\eta_0}{1 - \eta_0} \leq \frac{dP}{dQ} \leq \frac{1 - \eta_0}{\eta_0} \).

Step 2 Since \( \Delta r_c = \hat{r}_c - \frac{1}{2} - E_X[\hat{r}_c(\hat{\eta}, X)] \), it can be simply estimated by its sample average, and \( \eta \) is also substituted by its estimator \( \hat{\eta} \) from step 1. Furthermore, we only need to focus on estimating those \( \Delta r_c \) where \( c \in [\eta_0, 1 - \eta_0] \). This is due to the following lemma:

Lemma 2 For \( c < \eta_0 \) or \( c > 1 - \eta_0 \), \( \Delta r_c = 0 \).

Step 3 We define \( \Gamma(c) := \int_{\eta_0}^{c} \gamma(c) \, dc \), which is analogous to a cumulative distribution for \( \gamma(c) \). More specifically, \( \Gamma(c) \) has the following form:

\[
\Gamma(c) = \left( \frac{2}{1-c} f' \left( \frac{c}{1-c} \right) - 2f \left( \frac{c}{1-c} \right) \right) \bigg|^{c}_{\eta_0} (9)
\]

We generate \( n \) uniformly spaced points \( u_j \ (j = 1, \ldots, n) \) in the interval \( [\Gamma(\eta_0), \Gamma(1 - \eta_0)] \), and then the corresponding \( c_j := \Gamma^{-1}(u_j) \) where \( \Gamma^{-1} \) denotes the inverse function. Note that we do not actually need to construct the inverse function. \( c_j \) can be computed by zero-finding algorithms [10].

This quasi-Monte Carlo sampling scheme allocates more points to the region where \( \gamma(c) \) changes faster [11]. By adaptively spacing the samples, it is expected better convergence can be achieved in estimating the integral (In particular, its convergence rate \( O(n^{-1}) \) is faster than the importance sampling which has a typical rate of \( O(n^{-\frac{1}{2}}) \)). Alternatively one can also use Gaussian quadratures [10]. However Gaussian quadratures involve higher order derivatives of \( \gamma(c) \) which can be inconvenient.

**Estimating Several f-Divergences in One Shot**

According to the representation in equation (8), various divergences differ only in their choice of \( \gamma(c) \). This suggests that we can estimate them all in one shot by performing a single posterior estimation \( \hat{\eta} \). The additional work is to pre-compute and store the \( c_j \)s for different \( \gamma(c) \), which can be done offline and re-used again and again. In a sense, the computational cost for estimating additional f-divergences is very low using our approach. This is in contrast to the approach by Nguyen et al. [5] where a new optimization needs to be constructed for each choice of f-divergence.

In upper half of Table 1, we list several common choice of f-divergences and their corresponding weights \( \gamma(c) \). Note that a much larger variety of weights not listed here will also correspond to valid f-divergences. Basically, any non-negative weight \( \gamma(c) \) will qualify. In the second half of Table 1, we also gives two such examples for unnamed f-divergences. Note that representing f-divergences as weighted integrals also provides insight.

Table 1: Various f-divergences and their corresponding weights \( \gamma(c) \) in the discriminative representation.

<table>
<thead>
<tr>
<th>Name</th>
<th>( f(t) )</th>
<th>( \gamma(c) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variational</td>
<td>(</td>
<td>t - 1</td>
</tr>
<tr>
<td>KL</td>
<td>( \log(t) )</td>
<td>( c(1-c)^2 )</td>
</tr>
<tr>
<td>Hellinger</td>
<td>( (\sqrt{t} - 1)^2 )</td>
<td>( \frac{(c(1-c))^2}{1-c} )</td>
</tr>
<tr>
<td>Pearson</td>
<td>( (t - 1)^2 )</td>
<td>( \frac{2}{c(1-c)} )</td>
</tr>
<tr>
<td>Div I</td>
<td>( \frac{1}{\sqrt{t(1+t)}} - \frac{1}{4} )</td>
<td>( \frac{2}{c(1-c)} )</td>
</tr>
<tr>
<td>Div II</td>
<td>( t \log\left(\frac{1}{1+t}\right) + \log\left(\frac{4}{1+t}\right) )</td>
<td>( \frac{2}{c(1-c)} )</td>
</tr>
</tbody>
</table>
4 Stability of the Algorithm

In this section, we will analyze the stability of the 3-step algorithm presented in Section 3. That is we would like to study the behavior of the estimation error \( \epsilon := |\hat{I}_f - I_f| \) as we increase the number \( m \) of observations and the number \( n \) of the \( c_j \)s.

Three sources of errors are introduced in our algorithm each of which is associated to one step of the algorithm. In the following, we will define two intermediate estimates of \( I_f \) each of which is performed as if only one or two sources of error were not present:

1. \( I_f^1 \): we assume that both \( \gamma \) and \( \Delta r_c \) are exact, and the error is introduced only by the samples drawn from \( \gamma(c) \) for the integral.

2. \( I_f^2 \): we assume that only \( \Delta r_c \) is exact, and the error is introduced by the estimate \( \hat{\gamma} \) and the samples from \( \gamma(c) \).

Thus, the estimation error \( \epsilon \) can be decomposed into three terms using the triangular inequality:

\[
\epsilon = |\hat{I}_f - I_f^1| + |I_f^1 - I_f^2| + |I_f^2 - I_f|
\]

(10)

Our strategy is to provide bounds for each of the three terms in the rhs. of (10). Combining these results will then yield a stability bound for the overall error \( \epsilon \):

Lemma 3 \( |\hat{I}_f - I_f^1| = O(m^{-\frac{1}{2}}) \).

Lemma 4 \( |I_f^1 - I_f^2| = O(m^{-\frac{3}{2}}) \).

Lemma 5 \( |I_f^2 - I_f| = O(n^{-1}) \).

Theorem 6 \( \epsilon = |\hat{I}_f - I_f| = O(m^{-\frac{1}{2}}) + O(n^{-1}) \).

5 Online Extension

Algorithm 2 Online Estimation of \( f \)-Divergence.

1. Initialize \( \Delta \hat{r}_c \) to 0.
2. Incrementally update \( \hat{\gamma} \) using stochastic gradient descent, ie. \( \hat{\gamma} \leftarrow \hat{\gamma} - \rho \frac{\partial f}{\partial \hat{\gamma}}(x_i) \) where \( \frac{\partial f}{\partial \hat{\gamma}}(x_i) \) denotes the stochastic gradient.
3. Incrementally update \( \Delta \hat{r}_c \), ie. \( \Delta \hat{r}_c \leftarrow \frac{1}{\tau} \frac{i}{\tau} \Delta \hat{r}_c + \frac{1}{\tau} \delta \hat{r}_c \) where \( \delta \hat{r}_c := \frac{1}{\tau} \frac{i}{\tau} (\hat{\gamma}, x_i) \).
4. Estimate \( f \)-divergences using \( \frac{\Gamma(1-\eta \sigma)}{\eta \sigma} \sum_{j=1}^{n} \Delta \hat{r}_{c_j} \).

2. Estimating the statistical information \( \Delta r_c \) is carried out via an incremental update scheme in the online setting.
3. An online algorithm has an additional parameter \( \rho \) for the learning rate. We typically use a discounting scheme \( \rho = \rho_0 \frac{1}{\rho_0 + \tau} \) where \( \rho_0 = 100 \).

This online algorithm allows us to scale up to large problems where there can be millions of observations.

Stability Our online algorithm also has good convergence guarantees. Basically, we need to show that the online estimation of the posterior \( \gamma \) converges to its batch counterpart with a good rate. For the exponential model, this issue has been extensively studied [12].

Typically, an online estimate converges to its batch counterpart also with a rate of \( O(m^{-\frac{1}{2}}) \). Taking into account this additional source of error in Theorem 6, we can show that the online version also converges with a rate of \( O(m^{-\frac{1}{2}}) + O(n^{-1}) \).

6 Estimation of Dependence for Paired Data Streams

Having a measure of “distance” between distributions allows us to assess the dependence between two random variables \( W \) and \( Z \). Let the joint distribution of \( W \) and \( Z \) be \( P(W,Z) \) and their marginal be \( P(W) \) and \( P(Z) \) respectively. The dependence between \( W \) and \( Z \) can be quantified by \( I_f(P(W,Z), P(W)P(Z)) \). For instance, mutual information can be recovered when \( f = t \log(t) \).

In a paired data stream where one stream corresponds to variable \( W \) and another corresponds to \( Z \), we draw observations \((w_i, z_i)\) from the joint distribution \( P(W,Z) \). To draw observations from the product of the marginals \( P(W)P(Z) \), we pair \( w_i \) with \( z_i-\tau \) and generate virtual observations \((w_i, z_{i-\tau})\) where \( \tau \) is a delay parameter. Note that we may need to choose a sufficiently large \( \tau \) such that \( z_{i-\tau} \) appear to be independent.
7 Experiments

We conduct three sets of experiments to investigate the empirical performance of our algorithm. The first and the second experiments concern the batch and the online algorithm respectively; the third experiment uses the online algorithm to study synchronization between EEG signals.

**Batch Experiments** In this experiment, we compare our estimator (OUR) to the estimator of Nguyen et al. [5] (NJW) and Wang et al. [4] (WKV)\(^1\). We emulate the experiments in Nguyen et al. [5] and focus on the comparison of KL divergence estimation using two combinations of \(Q\) and \(R\): Gaussian versus Uniform, and Gaussian versus Gaussian. Furthermore, we instantiate \(Q\) and \(R\) with different dimensions to investigate the scalability of the algorithms as the dimension of the problems increases.

For each of the estimation problems shown in Figure 1, we study the convergence of different algorithms as the sample size increases. For our method and NJW, we use a Gaussian RBF kernel \(\exp(-s\|x-x'\|^2)\) in the estimator where the scale parameter is set to be the reciprocal of the median of distances between observations. For the regularization parameter, we used a decay schedule similar to Nguyen et al. [5] which scales as \(O(m^{-1})\). We take 2000 \(c_i\) for our algorithm, and use random features of 200 dimensions as described in [13] for computing the kernel.

Results show that our method has better convergence than the NJW and WKV method. Especially when the dimension of the data increases, the advantage of our method over the NJW method becomes more evident, as can be seen from Figure 1(a,c,e,g). Furthermore, the performance of our method and the NJW method is more consistent over the WKV method which does not converge in some cases. These result are also consistent with our theory that both our method and NJW bypass the hard problem of directly modeling the distribution \(Q\) and \(R\). Additionally, our method employs a discriminative estimation of the class probability while the NJW method is a generative approach which estimates the ratio \(dP/dQ\). This explains why our method has better convergence especially when the dimension of the problems increases.

**Online Experiments** In this experiment, we generate 1,000,000 data points for each dataset studied in Figure 1. We use our online algorithm to estimate

\(^1\)We use the M2 estimator of Nguyen et al. [5], since it is showed M2 has better convergence than M1. We use an adaptive partitioning and bias correction version of Wang et al. [4].
iments and we apply no regularization in this online setting.

The results are shown in Table 2. For all datasets, the stochastic estimates are very close to the true value. Furthermore, we observed that KL divergence is relatively harder to estimate. This is because the $\gamma(c)$ of KL-divergence has a sharper change near 0 and 1 which makes the integral harder to estimate.

Measuring EEG Synchronization In this experiment, we use our online divergence estimator to study the synchronization or statistical dependence between electroencephalogram (EEG) signals. Many neuroscientists believe that synchronization between EEG signals from different parts of a brain is an important indicator of the collaboration between these brain regions [14]. Therefore, various dependence measures, such as mutual information, have been applied to study EEG synchronization and different measures provide qualitatively different results. Our method is able to estimate in one shot a family of dependence measures related by a weight function $\gamma(c)$. In a sense, our method can provide a dependence “spectrum” which we know how different “frequencies” are related. Furthermore, the choice of the weight function can also be used to tailor $f$-divergences for specific purposes, e.g., emphasizing the data where $\eta(x) = c$ for certain $c$.

We demonstrate our idea using the EEG signals of subject ‘aa’ from dataset IVa of Berlin BCI group.2 The signals are sampled at 100Hz which provide us an estimation problem with more than 280,000 observations. During the EEG recording, the subject was performing motor imagination. We therefore focus on channel 12 (from frontal cortex) and 60 (from motor cortex) to investigate how higher level cognitive structures are interacting with lower level motor controls.

We estimated the dependence using three different $\gamma(c)$: $\frac{1}{1-\frac{c}{\sqrt{2}}}$. We fixed the learning rate of the online algorithm $\rho$ to 1, and set the delay $\tau = 100$ and we only display the estimated dependence after the algorithm has learned from 50,000 observations and started tracking the changes in EEG. Furthermore, we standardize the estimated dependence such that they can be displayed in the same scale.

In Figure 2(a), all three dependence measures indicate increased interaction between brain regions as the imagination experiment proceeds; but as the experiment approaches the end the brain interaction starts to drop. This seems to be consistent with the arousal of the brain as the mental experiment proceeds and the disengagement of different brain regions as the task approaches the end.

Another interesting thing is that the time courses of different dependence measures can be quite different. For instance, the dependence measure with $\gamma(c) = \frac{1}{c}$ peaks earlier than the measure with $\gamma(c) = \frac{1}{\sqrt{1-c^2}}$ (mutual information). This is because two dependence measures focus on different part of the distribution: $\frac{1}{c}$ places more weights on the part where $\eta$ is close to 0 while mutual information places more weights to $\eta$ close to 1. We also showed a scatter plot of these two dependence measures in Figure 2(b). This scatter plot shows that there is no simple linear relation between the two dependence measures. This suggests that different dependence measures may be needed to capture different aspects of mental processes.

![Figure 2](http://ida.first.fraunhofer.de/projects/bci/competition-iii/)

Figure 2: (a) Synchronization between EEG signals measured with 3 different dependence measures. (b) There is no simple linear relationship between different dependence measures.

References


Table 2: Stochastic estimation of various divergence for 1,000,000 data points. The standard deviation is obtained by repeating the experiments 10 times. For each data set and each divergence, the number on the left is the true value, and the number on the right is the estimated value.

<table>
<thead>
<tr>
<th></th>
<th>True Estimate</th>
<th>True Estimate</th>
<th>True Estimate</th>
<th>True Estimate</th>
<th>True Estimate</th>
<th>True Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(a)</td>
<td>0.389 ± 0.001</td>
<td>0.231 ± 0.001</td>
<td>0.751 ± 0.002</td>
<td>0.048 ± 0.000</td>
<td>0.215 ± 0.001</td>
<td></td>
</tr>
<tr>
<td>(b)</td>
<td>0.500 ± 0.002</td>
<td>0.235 ± 0.001</td>
<td>0.766 ± 0.001</td>
<td>0.051 ± 0.000</td>
<td>0.223 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(c)</td>
<td>0.778 ± 0.002</td>
<td>0.436 ± 0.001</td>
<td>1.051 ± 0.002</td>
<td>0.086 ± 0.000</td>
<td>0.394 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(d)</td>
<td>1.000 ± 0.006</td>
<td>0.442 ± 0.001</td>
<td>1.041 ± 0.001</td>
<td>0.088 ± 0.000</td>
<td>0.403 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(e)</td>
<td>1.167 ± 0.003</td>
<td>0.617 ± 0.001</td>
<td>1.238 ± 0.002</td>
<td>0.115 ± 0.000</td>
<td>0.543 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(f)</td>
<td>1.500 ± 0.008</td>
<td>0.625 ± 0.001</td>
<td>1.227 ± 0.001</td>
<td>0.115 ± 0.000</td>
<td>0.551 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(g)</td>
<td>1.555 ± 0.009</td>
<td>0.777 ± 0.001</td>
<td>1.374 ± 0.001</td>
<td>0.137 ± 0.000</td>
<td>0.666 ± 0.000</td>
<td></td>
</tr>
<tr>
<td>(h)</td>
<td>2.000 ± 0.019</td>
<td>0.787 ± 0.001</td>
<td>1.364 ± 0.001</td>
<td>0.138 ± 0.000</td>
<td>0.676 ± 0.000</td>
<td></td>
</tr>
</tbody>
</table>

Appendix

Proof of Theorem 1 This theorem is a modification of the Theorem 1 appeared in [15] which states that, given a classification task with prior $\pi = P(Y = 1)$, class conditional distribution $Q = P(X|Y = 1)$ and $R = P(X|Y = -1)$, the $f$-divergence of $Q$ from $R$ can be expressed as:

$$I_f(Q, R) = \int_0^1 \Delta r_x \pi(w) \, dp$$  \hspace{1cm} (11)

where $w(\pi) = \frac{2}{\pi} f''(\frac{1-\pi}{\pi})$, and $\Delta r_x$ is the statistical information when 0-1 loss 2 is used.

However, this integral representation is not in a desired form: it ranges over all possible priors $\pi$ and for each $\pi$ a different posterior $\eta$ needs to be estimated. For divergence estimation, it is desirable that the prior $\pi$ is fixed and $\eta$ is estimated only once. We will show that by introducing the cost $c$, the role of $\pi$ and $\eta$ can be switched and thus equation (11) can be transformed into a desired form.

Note the definition of $\Delta r_x$ can be written as:

$$\Delta r_x = \min \left\{ (1-c)\pi, c(1-\pi) \right\}$$  \hspace{1cm} (12)

where we use the fact $\eta dP(X) = \pi dQ$ and $(1-\eta) dP(X) = (1-\pi) dR$. Furthermore, we observe that $\Delta r_x = \Delta r_{1-\pi}$ due to the symmetry between $c$ and $\pi$. Therefore, the integral representation in equation (11) can be equivalently written as:

$$I_f(Q, R) = \int_0^1 \Delta r_{1-\pi} \pi(w) \, dp = \int_0^1 \Delta r_{1-\pi} w(1-c) \, dc,$$

where the second equality is obtained by redefining $c := 1-\pi$. Last, $\gamma(c) := w(1-c) = \frac{2}{(1-c)^2} f''(\frac{1}{1-c})$ proves the theorem.


Proof of Lemma 2  For $c < \eta_0 < \frac{1}{2}$ we have:

\[
\Delta r_c = \frac{c}{2} - \frac{1}{\eta_c} \eta(1-c)dP - \frac{1}{\eta_c} c(1-\eta)dP \\
= \frac{c}{2} - \frac{1 - c}{\frac{1}{2}} dQ - \frac{c}{\frac{1}{2}} dR \\
= \frac{1}{2} \int_{\eta_c} \cd R - (1-\eta)dQ, \\
\]

(13)

We define the function where we use the fact $\pi = \frac{1}{2}$, $\eta dP = \frac{1}{2} dQ$ and $(1-\eta)dP = \frac{1}{2} dR$ in the second equality. Since we assume $\eta \geq \eta_0$, when $\eta_0 \in (c, \frac{1}{2})$ the set $\{X | \eta(X) \leq c\}$ is empty and the corresponding integral is 0.

Similarly, when $\frac{1}{2} < 1 - \eta_0 < c$, we will have $\Delta r_c = \frac{1}{2} \int_{\eta_0} (1-c)dQ - c dR$. Since we assume $\eta \leq 1 - \eta_0$, when $\frac{1}{2} < 1 - \eta_0 < c$ the set $\{X | \eta(X) > c\}$ is also empty and the corresponding integral is also 0. Combining these two parts proves the theorem.

Proof of Lemma 5  We will make use of the Koksma-Hlawka inequality from [11]. Given a function $g(u)$ and observations $u$ drawn iid. from a uniform distribution $U[0, 1]$, this inequality says:

\[
\left| \frac{1}{n} \sum_{i=1}^{n} g(u_i) - \text{E}_{u \sim U[0,1]}[g(u)] \right| \leq D_n^* ||g||_{TV},
\]

(14)

where $||g||_{TV} := \int_0^1 |g(u)| du$ is called the total variation of $g$, and $D_n^*$ is called the star discrepancy. The star discrepancy is the Kolmogorov-Smirnov distance between a uniform distribution and its empirical estimate

\[
\frac{1}{n} \sum_{i=1}^{n} [u_i \in [0, a]].
\]

Intuitively, the total variation of $g$ is the vertical component of the arc-length of its graph.

First, we normalize $\gamma(c)$ to $\hat{\gamma}(c) = \frac{\gamma(c)}{\Gamma(1-\eta_0)}$ and correspondingly define $\hat{\Gamma}(c) = \int_0^c \hat{\gamma}(c) dc$. Recall the definition of $I_f$ and Lemma 2, we have:

\[
I_f = \int_{\eta_0}^{1-\eta_0} \Delta r_c \hat{\gamma}(c) dc \\
= \Gamma(1-\eta_0) \int_0^{1} \Delta r_{\hat{\Gamma}^{-1}(u)} \hat{\gamma}(\hat{\Gamma}^{-1}(u)) \hat{\Gamma}^{-1}(u) du \\
= \Gamma(1-\eta_0) \int_0^{1} \Delta r_{\hat{\Gamma}^{-1}(u)} du
\]

which means $g(u) = \Delta r_{\hat{\Gamma}^{-1}(u)}$ in our case. Therefore, we only need to bound the star discrepancy of $u_j$ and the total variation of $\Delta r_{\hat{\Gamma}^{-1}(u)}$.

First, the star discrepancy is $D_n^* = \frac{\Gamma(1-\eta_0)}{n}$. Second, $\Delta r_c$ has a maximum slope of $\frac{1}{2}$, ie. $|\frac{\Delta r_c}{du}| \leq \frac{1}{2}$. Then using the chain rule we have:

\[
\left\| \Delta r_{\hat{\Gamma}^{-1}(u)} \right\|_{TV} = \int_0^1 \left\| \frac{d\Delta r_{\hat{\Gamma}^{-1}(u)}}{du} \right\| du
\]

(16)

\[
= \int_0^1 \left\| \frac{d\Delta r_c}{du} \right\| \left\| \frac{d\hat{\Gamma}^{-1}(u)}{du} \right\| du \leq \frac{1}{2} \left\| \hat{\Gamma}^{-1}(u) \right\|_{TV} = \frac{1}{2}
\]

The last equality holds because $\hat{\Gamma}^{-1}(u)$ is a monotonic function with range $[0,1]$. Therefore, we have a deterministic bound, $|I_f - I_f| \leq \frac{\Gamma(1-\eta_0)}{2n} = O(n^{-1})$, which proves the lemma.

Proof of Lemma 4  First, we will bound the error $||\theta - \hat{\theta}||$. Since $||\theta - \hat{\theta}|| \leq \epsilon_0$ and $||\phi(x)|| \leq 1$, we have $||\theta(x) - \hat{\theta}(x) + \phi(x)|| \leq \epsilon_0$ by Cauchy-Schwartz inequality. Note that $\eta(x) = (1 + \exp(-\theta(x)))^{-1}$ has a maximum slope of 1 in terms of $\theta(x)$. This means that a small error in $(\theta(x), \phi(x))$ also translates into a small error in $\eta(x)$, ie. $|\eta(x) - \hat{\eta}(x)| \leq \epsilon_0$.

Second, we observe that $r_c(\eta(x), X) - r_c(\hat{\eta}(X)) \leq \epsilon_0$ since $|\eta(x) - \hat{\eta}(x)| \leq \epsilon_0$. This implies:

\[
\left| r_c^{\theta} - r_c^{\hat{\theta}} \right| = \left| \text{E}_{X}[r_c(\eta(x), X) - r_c(\hat{\eta}(X))] \right| \\
\leq \text{E}_{X}[r_c(\eta(x), X) - r_c(\hat{\eta}(X))] \leq \epsilon_0.
\]

(17)

Last, applying similar reasoning proves the lemma:

\[
\left| I_f - I_f^{\hat{\theta}} \right| = \left| \frac{1}{n} \sum r_c^{\theta} - \frac{1}{n} \sum r_c^{\hat{\theta}} \right| \\
\leq \frac{1}{n} \sum r_c^{\theta} - r_c^{\hat{\theta}} \leq \epsilon_0 = O(m^{-\frac{1}{2}}).
\]

(18)

Proof of Lemma 3  First, $\hat{I}_f - I_f^{\hat{\theta}}$ can be expressed in an alternative form by reordering the sum:

\[
\left| \frac{1}{n} \sum \left( \frac{1}{m} \sum r_c(\hat{\eta}(x_i)) - \frac{1}{n} \sum \text{E}_{X}[r_c(\hat{\eta}(X))] \right) \\
= \frac{1}{m} \sum \left( \frac{1}{n} \sum r_c(\hat{\eta}(x_i)) - \text{E}_{X} \left[ \frac{1}{n} \sum r_c(\hat{\eta}(X)) \right] \right)
\]

Now we can treat $\frac{1}{n} \sum r_c(\hat{\eta}(X))$ as a random variable and its range is $[0, \frac{1}{4}]$. Applying Hoeffding’s bound, we have for all $d > 0$:

\[
\text{Pr} \left\{ |\hat{I}_f - I_f^{\hat{\theta}}| \geq d \right\} \leq 2 \exp \left( \frac{-32md^2}{n} \right).
\]

(19)

Rearranging the terms proves the lemma: for all $\delta > 0$, with probability $1 - \delta$,

\[
|\hat{I}_f - I_f^{\hat{\theta}}| \leq \sqrt{\frac{\ln(2/\delta)}{32m}} = O(m^{-\frac{1}{2}}).
\]

(20)