The marginal probability distributions are $i = 1, X, X'$ for finite alphabet it is also possible to find the lower bound of the mutual information of two random variables with finite alphabets are established. While they are not particularly tight, they are the first where assumptions on the distribution and the sample size. Different from previous results, these intervals do not need any assumptions on the distribution and the sample size.

I. INTRODUCTION

In this paper confidence intervals for the mutual information of two random variables with finite alphabets are established. While they are not particularly tight, they are the first where no further restrictions have to be considered, neither on being in an asymptotic regime nor on the underlying joint probability distribution. By quantization of random variables with a non finite alphabet it is also possible to find the lower bound of the confidence interval of the mutual information of such random variables. The simplicity of these confidence intervals also allows to give an upper bound on the necessary sample size when the confidence interval width, the confidence level, and the alphabet sizes are fixed.

II. NOTATIONAL SETUP

Let $X, X', Y, Y'$ be two pairs of finite discrete random variables, with joint probability distributions $p_X = \{p_X(i) : i = 1, 2, \ldots, M_x\}$, $p_{X'} = \{p_{X'}(i) : i = 1, 2, \ldots, M_{x'}\}$, $p_Y = \{p_Y(j) : j = 1, 2, \ldots, M_y\}$, $p_{Y'} = \{p_{Y'}(j) : j = 1, 2, \ldots, M_{y'}\}$.

The marginal probability distributions are $p_X = \{p_X(i) : i = 1, 2, \ldots, M_x\}$, $p_Y = \{p_Y(j) : j = 1, 2, \ldots, M_y\}$, $p_{X'} = \{p_{X'}(i) : i = 1, 2, \ldots, M_{x'}\}$ and $p_{Y'} = \{p_{Y'}(j) : j = 1, 2, \ldots, M_{y'}\}$, where the marginals are calculated from the joint probability distributions as usual. The Shannon entropy $\Pi$ is defined as

$$H(X) = H(p_X) = -\sum_{i=1}^{M_x} p_X(i) \log p_X(i)$$

and the joint entropy $\Pi$ as

$$H(XY) = H(p_{XY}) = -\sum_{i=1}^{M_x} \sum_{j=1}^{M_y} p_{XY}(i,j) \log p_{XY}(i,j).$$

All logs are natural if not stated otherwise. $H(\cdot)$ is defined as the binary entropy function $H(x) = -x \log x - (1-x) \log(1-x)$.

The mutual information $\Pi$ is defined as

$$I(X; Y) = I(p_{XY}) = H(X) + H(Y) - H(XY).$$

W.l.o.g. it is assumed, that $M_x \leq M_y$, what can be done because the mutual information is symmetric ($I(X; Y) = I(Y; X)$), and therefore by renaming the variables if necessary it can be assumed that $M_x \leq M_y$ always holds. The variational distance between two probability distributions is defined as

$$V(p_{XY}, p_{X'Y'}) = \|p_{XY} - p_{X'Y'}\|_1 = \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} |p_{XY}(i,j) - p_{X'Y'}(i,j)|,$$

and similarly for the marginal distributions. It can be easily seen, that $V(\cdot, \cdot) \in [0, 2]$ for any two probability distributions. The empirical joint distribution for an i.i.d. sequence of pairs $((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n))$, sampled from a distribution $p_{XY}$, is defined as

$$p_{xy^n} = \{p_{xy^n}(i,j) : i = 1, 2, \ldots, M_x; j = 1, 2, \ldots, M_y\},$$

where

$$p_{xy^n}(i,j) = \frac{1}{n} \sum_{k=1}^{n} \delta_{x_k,i} \delta_{y_k,j}$$

and $\delta_{ij}$ is the Kronecker delta.

III. RELATED WORK

The following two bounds will be used to construct the confidence interval for mutual information and are stated here as two Lemmas.

**Lemma 1**: Let $(X, Y)$ and $(X', Y')$ be two pairs of random variables taking values on the same range, with joint probability distributions $p_{XY}$ and $p_{X'Y'}$. Let

$$\epsilon = V(p_{XY}, p_{X'Y'}).$$

Confidence Intervals for the Mutual Information

Arno G. Stefani, Johannes B. Huber
Institute for Information Transmission (LIT)
FAU Erlangen-Nuremberg
Erlangen, Germany
Email: {stefani, huber}@LNT.de

Christophe Jardin, Heinrich Sticht
Bioinformatics, Institute for Biochemistry
FAU Erlangen-Nuremberg
Erlangen, Germany
Email: {christophe.jardin, h.sticht}@biochem.uni-erlangen.de

Abstract—"THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD"

By combining a bound on the absolute value of the difference of mutual information between two joint probability distributions with a fixed variational distance, and a bound on the probability of a maximal deviation in variational distance between a true joint probability distribution and an empirical joint probability distribution, confidence intervals for the mutual information of two random variables with finite alphabets are established. Different from previous results, these intervals do not need any assumptions on the distribution and the sample size.
If \( \epsilon \leq 2 - \frac{2}{M_x M_y} \), then it holds that
\[
|I(X; Y) - I(X'; Y')| \leq 3 \cdot \frac{\epsilon}{2} \log(M_x M_y - 1) + 3 \mathcal{H}(\frac{\epsilon}{2}). \tag{3}
\]

**Lemma 2:** For any \( \epsilon > 0 \)
\[
\Pr[V(p_{XY}, p_{X'Y'}) > \epsilon] < (2^{M_x M_y} - 2)e^{-\epsilon^2/2}. \tag{4}
\]

The first bound was found by Zhang \[2\] Theorem 2. In the next section this bound will be slightly improved and generalized for the usage here, using a result of Ho and Yeung \[3\] Theorem 6. The second bound was originally found by Weissman et al. \[5\] Theorem 2.1 and slightly modified by Ho and Yeung \[3\] Lemma 3 to have no dependence on the true distribution.

**IV. RESULTS**

First, \[3\] is improved to yield:

**Theorem 1:** Let \((X, Y)\) and \((X', Y')\) be two pairs of random variables taking values on the same range, with joint probability distributions \(p_{XY}\) and \(p_{X'Y'}\) and \(M_x \leq M_y\). Fix an \( \epsilon > 0 \). Let
\[
V(p_{XY}, p_{X'Y'}) \leq \epsilon.
\]

Then it holds that
\[
|I(X; Y) - I(X'; Y')| \leq \begin{cases} 
\frac{\epsilon}{2} \log[(M_x M_y - 1)(M_x - 1)(M_y - 1)] + 3 \mathcal{H}(\frac{\epsilon}{2}) & \text{for } \epsilon \leq 2 - \frac{2}{M_x} \\
\log(M_x) & \text{for } \epsilon > 2 - \frac{2}{M_x}.
\end{cases}
\]

**Proof:** The proof widely follows the lines of the proof of \[3\] in Zhang \[2\] Eq. (2), but replaces the entropy difference bound of Zhang \[2\] Eq. 4 by the corresponding bound in Ho and Yeung \[3\] Theorem 6, which makes the new bound valid for any \( \epsilon \) and also for any \( V(p_{XY}, p_{X'Y'}) \leq \epsilon \) instead of \( V(p_{XY}, p_{X'Y'}) = \epsilon \). Beyond this, some slight changes in the proof of Zhang lead to a tighter bound.

First it is shown that \( V(p_X, p_{X'}) \leq \epsilon \):
\[
V(p_X, p_{X'}) = \|p_X - p_{X'}\|_1 \\
= \sum_{i=1}^{M_x} |p_X(i) - p_{X'}(i)| \\
= M_x \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} |p_{XY}(i, j) - p_{X'Y'}(i, j)| \\
\leq \sum_{i=1}^{M_x} \sum_{j=1}^{M_y} |p_{XY}(i, j) - p_{X'Y'}(i, j)| \\
= V(p_{XY}, p_{X'Y'}) \\
\leq \epsilon.
\]

In an analogous way it can be shown that \( V(p_Y, p_{Y'}) \leq \epsilon \).

For \( \epsilon \leq 2 - \frac{2}{M_x} \) then it holds:
\[
|I(X; Y) - I(X'; Y')| \\
= |H(X) + H(Y) - H(XY) - H(X') - H(Y') + H(X'Y')| \\
\leq |H(X) - H(X')| + |H(Y) - H(Y')| + |H(XY) - H(X'Y')| \\
\leq \frac{\epsilon}{2} \log(M_x - 1) + \mathcal{H}(\frac{\epsilon}{2}) + \frac{\epsilon}{2} \log(M_y - 1) + \mathcal{H}(\frac{\epsilon}{2}) \\
\leq \frac{\epsilon}{2} \log[(M_x M_y - 1)(M_x - 1)(M_y - 1)] + 3 \mathcal{H}(\frac{\epsilon}{2}).
\]

In \[3\] eq. \( 1 \) was used. In \[7\] the bound of Ho and Yeung \[3\] Theorem 6 was applied together with the assumption \( M_x \leq M_y \) and therefore, by the assumption \( \epsilon \leq 2 - \frac{2}{M_x} \), with \( 2 - \frac{2}{M_x} \geq 2 - \frac{2}{M_y} \geq 2 - \frac{2}{M_x} \geq \epsilon \).

For \( \epsilon > 2 - \frac{2}{M_x} \) the well known bounds on mutual information and entropy \[11\], \( I(X; Y) \geq 0 \) and \( I(X; Y) \leq H(X) \leq \log M_x \) are first used to show that
\[
0 \leq I(X; Y), I(X'; Y') \leq \log M_x,
\]
what immediately implies
\[
|I(X; Y) - I(X'; Y')| \leq \log M_x, \tag{8}
\]
indeed of \( \epsilon \), what completes the proof.

**Remark:** The absolute entropy difference bound of Ho and Yeung \[3\] Theorem 6 could also be used to bound \( |I(X; Y) - I(X'; Y')| \) in the case \( \epsilon > 2 - \frac{2}{M_x} \), but here it can easily be seen that \( |I(X; Y) - I(X'; Y')| = |H(X) - H(X') + |H(Y) - H(Y')| + |H(XY) - H(X'Y')| \leq \log M_x + |H(Y) - H(Y')| + |H(XY) - H(X'Y')| \geq \log M_x \) and therefore the upper bound \( \log M_x \) is tighter for \( \epsilon > 2 - \frac{2}{M_x} \). From this argumentation it can also be seen that the upper bound for the case that \( \epsilon \) is smaller, but close to \( 2 - \frac{2}{M_x} \), is still greater than \( \log M_x \), and could therefore be improved by taking the minimum of this bound and \( \log M_x \), but for the sake of simplicity and applicability of this bound this improvement has not been applied in Theorem 1. This shows that this bound is only useful for sufficiently small \( \epsilon \), since \( \log M_x \) is a well known and in the context of confidence intervals trivial bound. Nevertheless \[5\] is everywhere tighter than \[3\], applicable for any \( \epsilon \), and the variational distance \( V(p_{XY}, p_{X'Y'}) \) has only to be less or equal \( \epsilon \) and not strictly equal to \( \epsilon \) for \[5\]. Therefore Theorem 1 is an improvement of the bound of Zhang (Lemma 3).

Finally the confidence interval is constructed by a combination of Theorem 1 and Lemma 2.

**Theorem 2:** For any \( \alpha \in (0, 1] \) and \( M_x, M_y \) with \( M_x \leq M_y \) let (where \( \ln \) is the natural logarithm)
\[
\epsilon = \sqrt{\frac{2 \ln 2 M_x M_y - 2}{\alpha}}
\]
and
\[ \Delta I(\epsilon) = \begin{cases} \frac{2}{n} \log([M_x M_y - 1](M_x - 1)(M_y - 1)) + 3\mathcal{H}(\frac{\epsilon}{2}) \\ \log(M_x) \end{cases} \]
for \( \epsilon \leq 2 - \frac{2}{M_x} \)
and \( \epsilon > 2 - \frac{2}{M_x} \)
then, for any two random variables \( X, Y \) with true joint probability distribution \( p_{XY} \) and empirical joint probability distribution \( p_{X \times Y \gamma} \) it holds that
\[ \Pr[I(p_{X \times Y \gamma}) - \Delta I(\epsilon) \leq I(p_{XY}) \leq I(p_{X \times Y \gamma} + \Delta I(\epsilon))] \geq 1 - \alpha. \]

**Proof:** Rewriting (4) as
\[ \Pr[V(p_{X \times Y \gamma}, p_{X \times Y \gamma}) \leq \epsilon] \geq 1 - (2^{M_x M_y} - 2)e^{-n\epsilon^2/2}, \quad (9) \]
and solving \( 1 - \alpha = 1 - (2^{M_x M_y} - 2)e^{-n\epsilon^2/2} \) yields (obviously only the positive solution is of interest)
\[ \epsilon = \sqrt{\frac{2}{n} \log(\frac{2^{M_x M_y} - 2}{\alpha})}. \]
Then it follows that
\[ 1 - \alpha \leq \Pr[V(p_{X \times Y}, p_{X \times Y \gamma}) \leq \epsilon] \leq \Pr[H(X) + H(Y)] \leq \Delta I(\epsilon)] \]
\[ = \Pr[I(p_{X \times Y \gamma}) - \Delta I(\epsilon) \leq I(p_{XY}) \leq I(p_{X \times Y \gamma} + \Delta I(\epsilon))], \]
where (10) is an application of Theorem 1.

The next theorem gives an upper bound on the necessary number of samples \( n \) to achieve a given confidence interval width at a given confidence level \( 1 - \alpha \).

**Theorem 3:** For any \( \alpha \in (0, 1], M_x, M_y, \) with \( M_x \leq M_y \), and \( \gamma \in (0, \log M_x) \) let \( \epsilon \) be the minimum root of
\[ \frac{\epsilon}{2} \log([M_x M_y - 1](M_x - 1)(M_y - 1)) + 3\mathcal{H}(\frac{\epsilon}{2}) = \gamma. \]
Then for \( [\cdot] \) is the ceiling operator
\[ n = \left\lceil \frac{2}{\epsilon^2} \log(\frac{2^{M_x M_y} - 2}{\alpha}) \right\rceil \]
it holds that
\[ \Pr[I(p_{X \times Y \gamma}) - \gamma \leq I(p_{XY}) \leq I(p_{X \times Y \gamma} + \gamma)] \geq 1 - \alpha. \]

**Proof:** If \( \gamma \geq \log M_x \) then the probability of being within the bounds is trivially one, therefore \( \gamma \) is restricted to be less \( \log M_x \). Then obviously only the first part of (4) applies, where \( \epsilon \leq 2 - \frac{2}{M_x} \). It is easy to show, that this term is strictly increasing for \( \epsilon \in (0, 2 - \frac{2}{M_x}) \). Therefore there is only one solution for \( \epsilon \in (0, 2 - \frac{2}{M_x}) \) of equation (11) which is just the desired maximal variational distance between the true and the empirical joint distribution. This \( \epsilon \) is also the minimum root as stated in the theorem. Then solving (9) for \( n \), after the substitution of \( \Pr[V(p_{XY}, p_{X \times Y \gamma}) \leq \epsilon] \) by \( 1 - \alpha \), yields
\[ n \geq \frac{2}{\epsilon^2} \log(\frac{2^{M_x M_y} - 2}{\alpha}) \]
and therefore
\[ n \geq \left\lceil \frac{2}{\epsilon^2} \log(\frac{2^{M_x M_y} - 2}{\alpha}) \right\rceil \]
clearly suffices to guarantee
\[ \Pr[I(p_{X \times Y \gamma}) - \gamma \leq I(p_{XY}) \leq I(p_{X \times Y \gamma} + \gamma)] \geq 1 - \alpha. \]

The next theorem is an improvement of Theorem 3 that uses the entropy optimization procedures of [3, Theorems 2 and 3], which depend on the actual empirical distribution, instead of the worst case entropy difference bound [3, Theorem 6].

**Theorem 4:** For any \( \alpha \in (0, 1], M_x, M_y \) with \( M_x \leq M_y \) let
\[ \epsilon = \sqrt{\frac{2}{n} \log(\frac{2^{M_x M_y} - 2}{\alpha})} \]
and let
\[ I_{min} = \min_{p_X: V(p_{X \times Y \gamma}, p_X) \leq \epsilon} H(X) + \min_{p_Y: V(p_{Y \times Y \gamma}, p_Y) \leq \epsilon} H(Y) \]
\[ I_{max} = \max_{p_X: V(p_{X \times Y \gamma}, p_X) \leq \epsilon} H(X) + \max_{p_Y: V(p_{Y \times Y \gamma}, p_Y) \leq \epsilon} H(Y) \]
where the solutions for the entropy optimization problems are given in [3, Theorems 2 and 3]. Then it holds that
\[ \Pr[I_{min} \leq I(p_{XY}) \leq I_{max}] \geq 1 - \alpha. \]

**Proof:** Since \( V(p_{X \times Y \gamma}, p_X) \) as well as \( V(p_{Y \times Y \gamma}, p_Y) \) are \( \leq V(p_{X \times Y \gamma}, p_{XY}) \leq \epsilon \), as shown in the proof of Theorem 1, it is obvious that
\[ \min_{p_X: V(p_{X \times Y \gamma}, p_X) \leq \epsilon} I(p_{XY}) \geq I_{min}, \]
\[ \max_{p_X: V(p_{X \times Y \gamma}, p_X) \leq \epsilon} I(p_{XY}) \leq I_{max}. \]
By the argumentation of the proof of Theorem 2 again
\[ \epsilon = \sqrt{\frac{2}{n} \log(\frac{2^{M_x M_y} - 2}{\alpha})} \]
is fixed, and it follows that
\[ 1 - \alpha \leq \Pr[V(p_{XY}, p_{X \times Y \gamma}) \leq \epsilon] \leq \Pr[I_{min} \leq I(p_{XY}) \leq I_{max}] \]
\[ \leq \Pr[I_{min} \leq I(p_{XY}) \leq I_{max}] . \]
V. Discussion

Theorem 3 can be seen as an upper bound for $n$ (the number of samples), which is tight when Theorem 2 is used to determine the confidence interval. This is explained by the fact, that the absolute entropy difference bound that was used to construct the confidence intervals is completely independent of the actual empirical distribution $p_{x^n}$. Also, by using the entropy difference bounds, the dependence between the entropies $H(X)$, $H(Y)$ and $H(XY)$ was ignored, since for example the worst case distribution $p_{x^n}$ is not necessarily the marginal of the worst case distribution $p_{x^n}$. What makes the mutual information difference bound less tight again.

Taken together, one can see that there is much room left for improvement. By this, $n$ of Theorem 3 is an upper bound on the necessary samples size.

A first improvement of this situation was given in Theorem 4. An approach for making also use of the dependence between the entropies is given as a conjecture and only for two binary random variables in [4].

Besides this in the preprint [7], an algorithm for finding the lower bound of the confidence interval for a binary and an arbitrary finite random variable is given. This bound is tight in terms of the maximal variational distance between the empirical and the true joint distribution.

VI. Numerical Examples

In this section the different possibilities for the construction of the confidence intervals, which just have been discussed are compared in two numerical examples. In these particular examples it can be seen that the lower bound conjectured in [4] (called Method 1) matches the lower bound of preprint [7] (called Method 2) which gives a further indication for the correctness of at least the lower bound in [4] (though there is still no proof available).

The following setup is used: A binary symmetric channel (BSC) with input variable $X$ and output variable $Y$ is given, where the bit error rate (BER) is equal to $0.1$ and the input probabilities $p_X = \{\frac{1}{2}, \frac{1}{2}\}$.

The joint probabilities therefore are

$p_{XY}(1,1) = 0.45$, $p_{XY}(1,2) = 0.05$, $p_{XY}(2,1) = 0.05$, $p_{XY}(2,2) = 0.45$.

In this case the true mutual information is known to be

$I(p_{XY}) = 1 - H(0.1) = 0.53100$

(unlike in the sections before, in this section all logs are to the base 2). Then, taking $n = 10^5$ samples from $p_{XY}$ yielded the following exemplary empirical distribution

$p_{x^n y^n}(1,1) = 0.44950$, $p_{x^n y^n}(1,2) = 0.5058$, $p_{x^n y^n}(2,1) = 0.04868$, $p_{x^n y^n}(2,2) = 0.45124$.

Now fixing the confidence level $1 - \alpha = 0.95$ the presedcribed methods could be used to estimate the confidence intervals.

Before this is done, a good approximation to the best possible confidence interval is determined, where best possible interval is defined as having minimal interval width. Therefore samples of size $n$ are sampled $10^5$ times from $p_{XY}$, yielding an exemplary empirical sampling cumulative distribution function (cdf) of $I(p_{x^n y^n})$ (shown in Fig. 1), which should be a sufficiently good approximation to the real sampling cdf of $I(p_{x^n y^n})$, due to the high number of samples.

Then, since it can be seen from the empirical sampling cdf $I(p_{x^n y^n})$ that the sampling probability density function (pdf) is close to being unimodal and symmetric, the approximation to the smallest possible confidence interval is given by the $\frac{1}{4}$-quantile $\approx 0.52517$ and the $(1 - \frac{1}{4})$-quantile $\approx 0.53699$ of the empirical sampling cdf of $I(p_{x^n y^n})$ (both marked in Fig. 1).

In Table 1 the results of the two methods described in Section IV (Theorem 2 and 4) and of Method 1 and 2, applied to $p_{x^n y^n}$, are given. Here it can be seen, that the independence of the empirical distribution in Theorem 2 makes the confidence interval pretty broad compared to the other methods. Besides this, one can see that the improved methods (Method 1 and 2 in Table 1) have nearly the same performance as Theorem 4. The situation rather changes when a true distribution with small mutual information is used (such a situation is prevalent in [6]). This is shown in the following example, where a BSC is used with BER $= 0.2$ and an unequally distributed input variable $X$ with

![Graph showing empirical sampling cdf](image-url)
TABLE I

| Method                | Confidence interval | Lower bound | Upper bound | Width  |
|-----------------------|---------------------|-------------|-------------|--------|
| approximated best possible | 0.52517             | 0.53699     | 0.01182     |        |
| Theorem 2             | 0.38170             | 0.68504     | 0.30334     |        |
| Theorem 4             | 0.51645             | 0.55091     | 0.03445     |        |
| Method 1              | 0.51666             | 0.55080     | 0.03414     |        |
| Method 2              | 0.51666             | —           | —           |        |

distribution \( p_X = \{0.1, 0.9\} \). The joint probabilities therefore are

\[
  p_{XY}(1, 1) = 0.08, \quad p_{XY}(1, 2) = 0.02, \\
  p_{XY}(2, 1) = 0.18, \quad p_{XY}(2, 2) = 0.72.
\]

Here the true mutual information

\[
  I(p_{XY}) \approx 0.10482.
\]

Again taking \( n = 10^5 \) samples from \( p_{XY} \) yielded the following exemplary empirical joint distribution

\[
  p_{X^nY^n}(1, 1) = 0.07996, \quad p_{X^nY^n}(1, 2) = 0.02023, \\
  p_{X^nY^n}(2, 1) = 0.18012, \quad p_{X^nY^n}(2, 2) = 0.71969.
\]

The sampling cdf of \( I(p_{X^nY^n}) \) in this case can be seen in Fig. 2. The approximation to the smallest possible confidence interval is determined by the same method as in the first example. The results are given in Table II

TABLE II

| Method                | Confidence interval | Lower bound | Upper bound | Width  |
|-----------------------|---------------------|-------------|-------------|--------|
| approximated best possible | 0.10143             | 0.10826     | 0.00683     |        |
| Theorem 2             | -0.04743            | 0.25591     | 0.30334     |        |
| Theorem 4             | 0.05269             | 0.15721     | 0.10452     |        |
| Method 1              | 0.08679             | 0.12402     | 0.03723     |        |
| Method 2              | 0.08679             | —           | —           |        |

REFERENCES

[1] T. M. Cover and J. A. Thomas, Elements of Information Theory, 2nd ed. New York: Wiley, 2006.
[2] Z. Zhang, “Estimating mutual information via Kolmogorov distance,” IEEE Trans. Inform. Theory, vol. 53, no. 9, pp. 3280–3282, Sep. 2007.
[3] S.-W. Ho and R. W. Yeung, “The interplay between entropy and variational distance,” IEEE Trans. Inform. Theory, vol. 56, no. 12, pp. 5906–5929, Dec. 2010.
[4] A. G. Stefani, J. B. Huber, C. Jardin and H. Sticht, “Towards confidence intervals for the mutual information between two binary random variables,” In Proc. Workshop Computational Systems Biology (WCSB 2012), Ulm, Germany, Jun. 4–6, 2012.
[5] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdú and M.J. Weiberger, “Inequalities for the L1 Deviation of the Empirical Distribution,” Tech. Rept., HP Laboratories Palo Alto, HPL-2003-97 (R.1), Jun. 2003.
[6] O. G. Othersen, A. G. Stefani, J. B. Huber and H. Stich, “Application of Information Theory to Feature Selection in Protein Docking,” J Mol Model., vol. 18, no. 4, pp. 1285–1297, Jul. 2012.
[7] A. G. Stefani, J. B. Huber, C. Jardin and H. Stich, “A tight lower bound on the mutual information of a binary and an arbitrary finite random variable in dependence of the variational distance,” available at http://arxiv.org/abs/1301.5937

ACKNOWLEDGMENT

The authors would like to thank the DFG for supporting their research with SPP1395 in the projects HU634-7 and STI1553.