Entropy Bounds for Discrete Random Variables via Coupling

Igal Sason

Department of Electrical Engineering
Technion - Israel Institute of Technology
Haifa 32000, Israel

Istanbul, Turkey
July 2013

2013 IEEE International Symposium on Information Theory (ISIT 2013).

Motivation

- In many interesting applications, the exact distribution of X is not available or is numerically hard to compute.

Motivation

- In many interesting applications, the exact distribution of X is not available or is numerically hard to compute.
- In such cases, having good bounds on certain probability distances between X and another RV Y with a known p.m.f. can be valuable to get a rigorous bound on $|H(X) - H(Y)|$.

Motivation

- In many interesting applications, the exact distribution of X is not available or is numerically hard to compute.
- In such cases, having good bounds on certain probability distances between X and another RV Y with a known p.m.f. can be valuable to get a rigorous bound on $|H(X) - H(Y)|$.
- This work is a follow-up of the papers:
  1. S. W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. on Info. Theory*, vol. 56, pp. 5906–5929, Dec. 2010.
  2. Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. on Info. Theory*, vol. 53, pp. 3280–3282, Sept. 2007.
  3. I. Kontoyiannis, P. Harremoës and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Info. Theory*, pp. 466–472, Feb. 2005.

Motivation

- In many interesting applications, the exact distribution of X is not available or is numerically hard to compute.
- In such cases, having good bounds on certain probability distances between X and another RV Y with a known p.m.f. can be valuable to get a rigorous bound on $|H(X) - H(Y)|$.
- This work is a follow-up of the papers:
  1. S. W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. on Info. Theory*, vol. 56, pp. 5906–5929, Dec. 2010.
  2. Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. on Info. Theory*, vol. 53, pp. 3280–3282, Sept. 2007.
  3. I. Kontoyiannis, P. Harremoës and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Info. Theory*, pp. 466–472, Feb. 2005.
- The new ingredient is a derivation of improved bounds on the entropy difference that rely on both the **local and total variation distances**; this is done via **maximal coupling** combined with **Stein's method**.

Coupling

A **coupling** of a pair of two RVs (X, Y) is a pair of two random variables $(\hat{X}, \hat{Y})$ with the same marginal probability distributions as of (X, Y).

Coupling

A **coupling** of a pair of two RVs (X, Y) is a pair of two random variables $(\hat{X}, \hat{Y})$ with the same marginal probability distributions as of (X, Y).

Maximal Coupling

For a pair of RVs (X, Y), a coupling $(\hat{X}, \hat{Y})$ is called a **maximal coupling** if $\mathbb{P}(\hat{X} = \hat{Y})$ is as large as possible among all the couplings of (X, Y).

Coupling

A **coupling** of a pair of two RVs (X, Y) is a pair of two random variables $(\hat{X}, \hat{Y})$ with the same marginal probability distributions as of (X, Y).

Maximal Coupling

For a pair of RVs (X, Y), a coupling $(\hat{X}, \hat{Y})$ is called a **maximal coupling** if $\mathbb{P}(\hat{X} = \hat{Y})$ is as large as possible among all the couplings of (X, Y).

Total Variation and Local Distances

Let X and Y be discrete RVs that take values in a set $\mathcal{A}$, and let P_X and P_Y be their p.m.f. The **local** and **total variation distances** are

$$d_{\mathsf{loc}}(X, Y) \triangleq \sup_{u \in \mathcal{A}} |P_X(u) - P_Y(u)|, \qquad d_{\mathsf{TV}}(X, Y) \triangleq \frac{1}{2} \sum_{u \in \mathcal{A}} |P_X(u) - P_Y(u)|.$$

The local distance is the l^∞ distance between the p.m.f, the total variation distance is half the l^1 distance, and $d_{\mathsf{loc}}(X, Y) \leq d_{\mathsf{TV}}(X, Y)$.

Link Between Maximal Coupling and Total Variation Distance

If $(\hat{X}, \hat{Y})$ is a maximal coupling of (X, Y) then $\mathbb{P}(\hat{X} \neq \hat{Y}) = d_{\mathsf{TV}}(X, Y)$.

Bound on the Entropy of Discrete Random Variables (Zhang, 07)

Theorem

Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{\mathsf{TV}}(X, Y) \log(M-1) + h\big(d_{\mathsf{TV}}(X, Y)\big)$$

where h denotes the binary entropy function.

Bound on the Entropy of Discrete Random Variables (Zhang, 07)

Theorem

Let X and Y be two discrete random variables that take values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{TV}(X,Y) \log(M-1) + h\big(d_{TV}(X,Y)\big)$$

where h denotes the binary entropy function.

Corollary

If $d_{TV}(X,Y) \leq \varepsilon$, then

$$|H(X) - H(Y)| \leq \begin{cases} \varepsilon \log(M-1) + h(\varepsilon) & \text{if } \varepsilon \in \left[0, 1 - \frac{1}{M}\right] \\ \log(M) & \text{if } \varepsilon > 1 - \frac{1}{M} \end{cases}$$

Simplified Proof of of Zhang's inequality

$$\begin{aligned}
&\left|H(X) - H(Y)\right| \\
&= \left|H(\hat{X}) - H(\hat{Y})\right| \\
&= \left|H(\hat{X}|\hat{Y}) - H(\hat{Y}|\hat{X})\right| \\
&\leq \max\Big\{H(\hat{X}|\hat{Y}), H(\hat{Y}|\hat{X})\Big\} \\
&\leq \mathbb{P}(\hat{X} \neq \hat{Y}) \, \log(M-1) + h\big(\mathbb{P}(\hat{X} \neq \hat{Y})\big) \\
&= d_{\mathsf{TV}}(X,Y) \, \log(M-1) + h\big(d_{\mathsf{TV}}(X,Y)\big).
\end{aligned}$$

Simplified Proof of of Zhang's inequality

$$\begin{aligned}
&\big|H(X) - H(Y)\big| \\
&= \big|H(\hat{X}) - H(\hat{Y})\big| \\
&= \big|H(\hat{X}|\hat{Y}) - H(\hat{Y}|\hat{X})\big| \\
&\leq \max\Big\{H(\hat{X}|\hat{Y}), H(\hat{Y}|\hat{X})\Big\} \\
&\leq \mathbb{P}(\hat{X} \neq \hat{Y}) \log(M-1) + h\big(\mathbb{P}(\hat{X} \neq \hat{Y})\big) \\
&= d_{\mathsf{TV}}(X,Y) \log(M-1) + h\big(d_{\mathsf{TV}}(X,Y)\big).
\end{aligned}$$

Example where Equality is Achieved

If $\varepsilon \in [0, 1 - \frac{1}{M}]$, the bound is tight when

$$X \sim P_X = \left(1 - \varepsilon, \frac{\varepsilon}{M-1}, \ldots, \frac{\varepsilon}{M-1}\right), \qquad Y \sim P_Y = (1, 0, \ldots, 0)$$

Note

In this example, $d_{\mathsf{loc}}(X, Y) = d_{\mathsf{TV}}(X, Y)$.

Note

In this example, $d_{\mathsf{loc}}(X,Y) = d_{\mathsf{TV}}(X,Y)$.

Main Observation I

If the local distance between two probability distributions on a finite alphabet is smaller than the total variation distance, then the bounds on the entropy difference can be significantly strengthened.

A Refinement of the Bound (Finite Alphabets)

Theorem

Let X and Y be discrete RVs taking values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{\mathsf{TV}}(X,Y) \log(M\alpha - 1) + h\big(d_{\mathsf{TV}}(X,Y)\big) \quad (1)$$

where $\alpha \triangleq \frac{d_{\mathsf{loc}}(X,Y)}{d_{\mathsf{TV}}(X,Y)}$ denotes the ratio of the local and total variation distances (so, $\alpha \in [\frac{2}{M}, 1]$), and h denotes the binary entropy function.

A Refinement of the Bound (Finite Alphabets)

Theorem

Let X and Y be discrete RVs taking values in a set $\mathcal{A}$, and let $|\mathcal{A}| = M$. Then,

$$|H(X) - H(Y)| \leq d_{\mathsf{TV}}(X, Y) \log(M\alpha - 1) + h\big(d_{\mathsf{TV}}(X, Y)\big) \quad (1)$$

where $\alpha \triangleq \frac{d_{\mathsf{loc}}(X,Y)}{d_{\mathsf{TV}}(X,Y)}$ denotes the ratio of the local and total variation distances (so, $\alpha \in [\frac{2}{M}, 1]$), and h denotes the binary entropy function. Furthermore, if $\frac{1}{2} \leq \frac{P_X}{P_Y} \leq 2$ whenever $P_X, P_Y > 0$, then the bound in (1) is tightened to

$$|H(X) - H(Y)| \leq d_{\mathsf{TV}}(X, Y) \log\left(\frac{M\alpha - 1}{4}\right) + h\big(d_{\mathsf{TV}}(X, Y)\big).$$

Concept of Proof of the New Theorem

The previous simplified proof only relies on the total variation distance. Not clear how the local distance can be helpful to improve the bound.

1. The proof relies on a specific construction of maximal coupling.
2. The derivation of the bound leads to a non-convex optimization problem of the form:

$$\text{maximize} \quad \left(-\sum_{i=1}^{M} s_i \log(s_i) + \sum_{i=1}^{M} t_i \log(t_i) \right)$$

subject to

$$\begin{cases} s_i, t_i \geq 0, \; s_i + t_i \leq \alpha \\ s_i t_i = 0, \quad \forall\, i \in \{1, \ldots, M\} \\ \displaystyle\sum_{i=1}^{M} s_i = \sum_{i=1}^{M} t_i = 1 \end{cases}$$

with the $2M$ variables $s_1, t_1, \ldots s_M, t_M$.

Concept of proof (Cont.)

Fortunately, this non-convex optimization problem admits the following closed-form solution:

$$g(\alpha) = \log\left(M - \left\lceil \frac{1}{\alpha} \right\rceil\right) + \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor \log \alpha + \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right) \log \left(1 - \alpha \left\lfloor \frac{1}{\alpha} \right\rfloor\right).$$

No need for Fano's inequality in this case. This proof is completely different from the previous (simplified) proof of Zhang's inequality.
Full details in the paper:
I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to *IEEE Trans. on Info. Theory*, Sept. 2012.
`http://arxiv.org/abs/1209.5259`.

Special Cases of the New Bound

- Since, in general, $\alpha \leq 1$ then the case where $\alpha = 1$ is the worst case for the new bound. In the latter case, it is particularized to the bound by Zhang (2007).

Special Cases of the New Bound

- Since, in general, $\alpha \leq 1$ then the case where $\alpha = 1$ is the worst case for the new bound. In the latter case, it is particularized to the bound by Zhang (2007).
- If $\alpha \leq \frac{1}{N}$ for some integer N (since $\alpha \in \left[\frac{2}{M}, 1\right]$ then it yields that $N \in \{1, \ldots, \lfloor \frac{M}{2} \rfloor\}$), the new bound implies that

$$|H(X) - H(Y)| \leq d_{\mathsf{TV}}(X,Y) \log\left(\frac{M-N}{N}\right) + h\big(d_{\mathsf{TV}}(X,Y)\big).$$

This inequality and Theorem 7 by Ho and Yeung (2010) are similar *but they hold under different conditions* where none of them implies the other.

Main Observation II

There is an extension of the new bound to countably infinite alphabets, where just knowing the total variation distance between two distributions does not imply anything about the difference of the respective entropies (i.e., one has discontinuity of entropy).

Main Observation II

There is an extension of the new bound to countably infinite alphabets, where just knowing the total variation distance between two distributions does not imply anything about the difference of the respective entropies (i.e., one has discontinuity of entropy).

Specifically, if one of the distributions is finitely supported, then knowing also something about the local distance and the tail behavior of the other distribution allows to bound the difference of entropies in this case.

The entropy difference for countably infinite alphabets - New Bound

Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countably infinite set. Let X and Y be discrete RVs where X takes values in the set $\mathcal{X} = \{a_1, \ldots, a_m\}$ for some $m \in \mathbb{N}$, and Y takes values in the set $\mathcal{A}$. Assume that for some $\eta_1, \eta_2, \eta_3 > 0$,

$$\eta_2 \leq d_{\mathsf{TV}}(X, Y) \leq \eta_1, \quad d_{\mathsf{loc}}(X, Y) \leq \eta_3$$

where $\eta_3 \leq \eta_2$.

The entropy difference for countably infinite alphabets - New Bound

Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countably infinite set. Let X and Y be discrete RVs where X takes values in the set $\mathcal{X} = \{a_1, \ldots, a_m\}$ for some $m \in \mathbb{N}$, and Y takes values in the set $\mathcal{A}$. Assume that for some $\eta_1, \eta_2, \eta_3 > 0$,

$$\eta_2 \leq d_{\mathsf{TV}}(X, Y) \leq \eta_1, \quad d_{\mathsf{loc}}(X, Y) \leq \eta_3$$

where $\eta_3 \leq \eta_2$. Let M be an integer such that

$$\sum_{i=M}^{\infty} P_Y(a_i) \leq \eta_3, \quad M \geq \max\left\{m + 1, \frac{\eta_2}{(1-\eta_1)\eta_3}\right\}$$

and let $\eta_4 > 0$ satisfy $-\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) \leq \eta_4$.

The entropy difference for countably infinite alphabets - New Bound

Let $\mathcal{A} = \{a_1, a_2, \ldots\}$ be a countably infinite set. Let X and Y be discrete RVs where X takes values in the set $\mathcal{X} = \{a_1, \ldots, a_m\}$ for some $m \in \mathbb{N}$, and Y takes values in the set $\mathcal{A}$. Assume that for some $\eta_1, \eta_2, \eta_3 > 0$,

$$\eta_2 \leq d_{\mathsf{TV}}(X, Y) \leq \eta_1, \quad d_{\mathsf{loc}}(X, Y) \leq \eta_3$$

where $\eta_3 \leq \eta_2$. Let M be an integer such that

$$\sum_{i=M}^{\infty} P_Y(a_i) \leq \eta_3, \quad M \geq \max\left\{m+1, \frac{\eta_2}{(1-\eta_1)\eta_3}\right\}$$

and let $\eta_4 > 0$ satisfy $-\sum_{i=M}^{\infty} P_Y(a_i) \log P_Y(a_i) \leq \eta_4$. Then, the following inequality holds:

$$|H(X) - H(Y)| \leq \eta_1 \log\left(\frac{M\eta_3}{\eta_2} - 1\right) + h(\eta_1) + \eta_4.$$

Poisson Approximation

- Example: The entropy of a sum of a large number (n) of Bernoulli RVs $(X_i \sim \text{Bern}(p_i))$ that none of them dominates the sum; their distribution is close to the Poisson distribution with parameter $\lambda = \sum_{i=1}^{n} p_i$ (Law of Small Numbers - Kontoyiannis et al., 2005).

Poisson Approximation

- Example: The entropy of a sum of a large number (n) of Bernoulli RVs $(X_i \sim \text{Bern}(p_i))$ that none of them dominates the sum; their distribution is close to the Poisson distribution with parameter $\lambda = \sum_{i=1}^{n} p_i$ (Law of Small Numbers - Kontoyiannis et al., 2005).
- In this work, we derive improved bounds on the entropy of a sum of independent Bernoulli RVs (not necessarily identically distributed).

Bounds on the Total Variation Distance (Barbour and Hall, 1984)

Let $W = \sum_{i=1}^n X_i$ be a sum of n independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$ for $i \in \{1, \dots, n\}$, and $\mathbb{E}(W) = \lambda$. Then, the total variation distance between the probability distribution of W and the Poisson distribution with mean λ satisfies

$$\frac{1}{32}\left(1 \wedge \frac{1}{\lambda}\right) \sum_{i=1}^n p_i^2 \leq d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^n p_i^2$$

where $a \wedge b \triangleq \min\{a, b\}$ for every $a, b \in \mathbb{R}$.

The derivation of the upper and lower bounds is based on the Chen-Stein method for Poisson approximation.

Improved Lower Bound on the Total Variation Distance (I.S., ITA '13)

Let $W = \sum_{i=1}^{n} X_i$ be a sum of n independent Bernoulli random variables with $\mathbb{E}(X_i) = p_i$ for $i \in \{1, \ldots, n\}$, and $\mathbb{E}(W) = \lambda$. Then, the following inequality holds:

$$\widetilde{K}_1(\lambda) \sum_{i=1}^{n} p_i^2 \leq d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq \left(\frac{1 - e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2$$

where

$$\widetilde{K}_1(\lambda) \triangleq \frac{e}{2\lambda} \, \frac{1 - \frac{1}{\theta} \, \left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}}$$

$$\theta \triangleq 3 + \frac{7}{\lambda} + \frac{1}{\lambda} \cdot \sqrt{(3\lambda + 7)\left[(3 + 2e^{-1/2})\lambda + 7\right]}.$$

Upper Bound on the Local Distance (Barbour et al., 1992)

$$d_{\mathsf{loc}}(P_W, \mathsf{Po}(\lambda)) \leq 4 \, \min\left\{\sqrt{\frac{2}{e\lambda}}, \, 2e^{-\lambda} \, I_0(\lambda)\right\} \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2$$

where I_0 denotes the modified Bessel function of order zero.

Application of the New Bound for the Poisson Approximation

The new bound on the entropy difference enables to get a rigorous bound on the entropy difference $H(\text{Po}(\lambda)) - H(W)$ with the constants

$$\eta_1 \triangleq \frac{\lambda(1-e^{-\lambda})}{n}$$

$$\eta_2 \triangleq \frac{e}{2} \, \frac{1 - \frac{1}{\theta}\,\left(3 + \frac{7}{\lambda}\right)}{\theta + 2e^{-1/2}} \, \frac{\lambda}{n}$$

$$\eta_3 \triangleq \min\left\{ 1, 4\sqrt{\frac{2}{\pi\lambda}}, \, 8e^{-\lambda} \, I_0(\lambda) \right\} \frac{\lambda\bigl(1-e^{-\lambda}\bigr)}{n}$$

$$\eta_4 \triangleq \left[\left(\lambda \log\left(\frac{e}{\lambda}\right)\right)_+ + \lambda^2 + \frac{6\log(2\pi)+1}{12}\right] \cdot \exp\left\{-\left[\lambda + (M-2)\, \log\left(\frac{M-2}{\lambda e}\right)\right]\right\}$$

$$M \triangleq \max\left\{n+2, \, \frac{\eta_2}{\eta_3(1-\eta_1)}, \, \lambda e^2, \, \ln\left(\frac{1}{\eta_3}\right) - \lambda\right\}.$$

Poisson Approximation

This leads to very accurate estimates of the entropy of sums of independent Bernoulli RVs (not necessarily i.i.d.). For details, see:
I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to *IEEE Trans. on Info. Theory*, Sept. 2012.
`http://arxiv.org/abs/1209.5259`.

Poisson Approximation

This leads to very accurate estimates of the entropy of sums of independent Bernoulli RVs (not necessarily i.i.d.). For details, see:
I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to *IEEE Trans. on Info. Theory*, Sept. 2012.
`http://arxiv.org/abs/1209.5259`.

Poisson Approximation (Cont.)

Weaker bounds on the entropy of sums of **dependent**, non-identically distributed Bernoulli RVs were derived (that only depend on the total variation distance), and their application was exemplified. See:
I. Sason, "On the entropy of sums of Bernoulli random variables via the Chen-Stein method," *Proceedings of ITW 2012*, pp. 542–546, Lausanne, Switzerland, Sept. 2012. `http://arxiv.org/abs/1207.0436`.

Summary and Conclusions

- This work refines bounds on the entropy difference of two discrete RVs via the use of maximal couplings, leading to sharpened bounds that depend on both the local and total variation distances.

Summary and Conclusions

- This work refines bounds on the entropy difference of two discrete RVs via the use of maximal couplings, leading to sharpened bounds that depend on both the local and total variation distances.
- The derivation of the new bounds relies on the notion of *maximal coupling*, which is also known to be useful for the derivation of error bounds via Stein's method.

Summary and Conclusions

- This work refines bounds on the entropy difference of two discrete RVs via the use of maximal couplings, leading to sharpened bounds that depend on both the local and total variation distances.
- The derivation of the new bounds relies on the notion of *maximal coupling*, which is also known to be useful for the derivation of error bounds via Stein's method.
- The link between Stein's method and information theory was pioneered by Barbour et al. (2010) in the context of the compound Poisson distribution. A recent work by Ley & Swan, (2012) further links between information theory and Stein's method.

Summary and Conclusions

- This work refines bounds on the entropy difference of two discrete RVs via the use of maximal couplings, leading to sharpened bounds that depend on both the local and total variation distances.
- The derivation of the new bounds relies on the notion of *maximal coupling*, which is also known to be useful for the derivation of error bounds via Stein's method.
- The link between Stein's method and information theory was pioneered by Barbour et al. (2010) in the context of the compound Poisson distribution. A recent work by Ley & Swan, (2012) further links between information theory and Stein's method.
- The new bounds were exemplified in the context of the Poisson approximation, showing remarkable improvement in their tightness.

Related Papers

1. A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford University Press, 1992.
2. A. D. Barbour, O. Johnson, I. Kontoyiannis and M. Madiman, "Compound Poisson approximation via information functionals," EJP, vol. 15, pp. 1344–1368, 2010.
3. S. W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. on Info. Theory*, vol. 56, pp. 5906–5929, Dec. 2010.
4. I. Kontoyiannis, P. Harremoës and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Info. Theory*, vol. 51, pp. 466–472, Feb. 2005.
5. C. Ley and Y. Swan, "Stein's density approach for discrete distributions and information inequalities," accepted to the *IEEE Trans. on Info. Theory*, 2013.
6. I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to *IEEE Trans. on Info. Theory*, Sept. 2012. `http://arxiv.org/abs/1209.5259`.
7. I. Sason, "Improved lower bounds on the total variation distance and relative entropy for the Poisson approximation," ITA 2013 Workshop, Feb. '13.
8. Z. Zhang, "Estimating mutual information via Kolmogorov distance," *IEEE Trans. on Info. Theory*, vol. 53, pp. 3280–3282, Sept. 2007.