On the Entropy of Sums of Bernoulli Random Variables via the Chen-Stein Method

Igal Sason

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

ETH, Zurich, Switzerland
August 20, 2012.

Problem Statement

- Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

Problem Statement

- Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

- Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

where it is assumed that $\lambda \in (0, \infty)$.

Problem Statement

- Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

- Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

  where it is assumed that $\lambda \in (0, \infty)$.
- Let $\mathrm{Po}(\lambda)$ denote the Poisson distribution with parameter λ.

Problem Statement

- Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

- Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

  where it is assumed that $\lambda \in (0, \infty)$.
- Let $\mathrm{Po}(\lambda)$ denote the Poisson distribution with parameter λ.

Problem: Derive tight bounds for the entropy of W.

Problem Statement

- Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

- Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

  where it is assumed that $\lambda \in (0, \infty)$.

- Let $\mathrm{Po}(\lambda)$ denote the Poisson distribution with parameter λ.

Problem: Derive tight bounds for the entropy of W.
In this talk, error bounds on the entropy difference $H(W) - H(\mathrm{Po}(\lambda))$ are introduced, providing rigorous bounds for the Poisson approximation of the entropy $H(W)$. These bounds are exemplified.

Poisson Approximation

Binomial Approximation to the Poisson

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli random variables with parameter $\frac{\lambda}{n}$, then for large n, the distribution of their sum is close to a Poisson distribution with parameter λ: $S_n \triangleq \sum_{i=1}^{n} X_i \approx \mathrm{Po}(\lambda)$.

Poisson Approximation

Binomial Approximation to the Poisson

If $X_1, X_2, \ldots, X_n$ are i.i.d. Bernoulli random variables with parameter $\frac{\lambda}{n}$, then for large n, the distribution of their sum is close to a Poisson distribution with parameter λ: $S_n \triangleq \sum_{i=1}^{n} X_i \approx \mathrm{Po}(\lambda)$.

General Poisson Approximation (Law of Small Numbers)

For random variables $\{X_i\}_{i=1}^{n}$ on $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$, the sum $\sum_{i=1}^{n} X_i$ is approximately Poisson distributed with mean $\lambda = \sum_{i=1}^{n} p_i$ as long as

- $\mathbb{P}(X_i = 0)$ is close to 1,
- $\mathbb{P}(X_i = 1)$ is uniformly small,
- $\mathbb{P}(X_i > 1)$ is negligible as compared to $\mathbb{P}(X_i = 1)$,
- $\{X_i\}_{i=1}^{n}$ are weakly dependent.

Information-Theoretic Results for Poisson Approximation

Maximum Entropy Result for Poisson Approximation

The $\mathrm{Po}(\lambda)$ distribution has maximum entropy among all probability distributions that can be obtained as sums of independent Bernoulli RVs:

$$H(\mathrm{Po}(\lambda)) = \sup_{S \in B_\infty(\lambda)} H(S)$$

$$B_\infty(\lambda) \triangleq \bigcup_{n \in \mathbb{N}} B_n(\lambda)$$

$$B_n(\lambda) \triangleq \left\{ S : \; S = \sum_{i=1}^{n} X_i, \; X_i \sim \mathrm{Bern}(p_i) \text{ independent}, \; \sum_{i=1}^{n} p_i = \lambda \right\}.$$

Information-Theoretic Results for Poisson Approximation

Maximum Entropy Result for Poisson Approximation

The $\mathrm{Po}(\lambda)$ distribution has maximum entropy among all probability distributions that can be obtained as sums of independent Bernoulli RVs:

$$H(\mathrm{Po}(\lambda)) = \sup_{S \in B_\infty(\lambda)} H(S)$$

$$B_\infty(\lambda) \triangleq \bigcup_{n \in \mathbb{N}} B_n(\lambda)$$

$$B_n(\lambda) \triangleq \left\{ S : \; S = \sum_{i=1}^{n} X_i, \; X_i \sim \mathrm{Bern}(p_i) \text{ independent}, \; \sum_{i=1}^{n} p_i = \lambda \right\}.$$

Due to a monotonicity property, then

$$H(\mathrm{Po}(\lambda)) = \lim_{n \to \infty} \sup_{S \in B_n(\lambda)} H(S).$$

Maximum Entropy Result for Poisson Approximation (Cont.)

For $n \in \mathbb{N}$, the maximum entropy distribution in the class $B_n(\lambda)$ is $\text{Binomial}\left(n, \frac{\lambda}{n}\right)$, so

$$H(\text{Po}(\lambda)) = \lim_{n\to\infty} H\left(\text{Binomial}\left(n, \frac{\lambda}{n}\right)\right).$$

Maximum Entropy Result for Poisson Approximation (Cont.)

For $n \in \mathbb{N}$, the maximum entropy distribution in the class $B_n(\lambda)$ is $\text{Binomial}\left(n, \frac{\lambda}{n}\right)$, so

$$H(\text{Po}(\lambda)) = \lim_{n\to\infty} H\left(\text{Binomial}\left(n, \frac{\lambda}{n}\right)\right).$$

Proofs rely on convexity arguments a la Mateev (1978), Shepp & Olkin (1978), Karlin & Rinott (1981), Harremoës (2001).

Maximum Entropy Result for Poisson Approximation (Cont.)

For $n \in \mathbb{N}$, the maximum entropy distribution in the class $B_n(\lambda)$ is $\text{Binomial}\left(n, \frac{\lambda}{n}\right)$, so

$$H(\text{Po}(\lambda)) = \lim_{n\to\infty} H\left(\text{Binomial}\left(n, \frac{\lambda}{n}\right)\right).$$

Proofs rely on convexity arguments a la Mateev (1978), Shepp & Olkin (1978), Karlin & Rinott (1981), Harremoës (2001).

Recent generalizations and extensions by Johnson et al. (2007–12):

- Extension of this maximum entropy result to the larger set of ultra-log-concave probability mass functions.
- Generalization to maximum entropy results for discrete compound Poisson distributions.

Information-Theoretic Ideas in Poisson Approximation

Nice surveys on the information-theoretic approach for Poisson approximation are available at:

1. I. Kontoyiannis, P. Harremoës, O. Johnson and M. Madiman, "Information-theoretic ideas in Poisson approximation and concentration," slides of a short course (the slides are available at the homepage of I. Kontoyiannis), September 2006.
2. O. Johnson, *Information Theory and the Central Limit Theorem*, Imperial College Press, 2004.

Total Variation Distance

Let P and Q be two probability measures defined on a set $\mathcal{X}$. Then, the total variation distance between P and Q is defined by

$$d_{\mathsf{TV}}(P,Q) \triangleq \sup_{\text{Borel } A \subseteq \mathcal{X}} |P(A) - Q(A)|$$

where the supermum is taken w.r.t. all the Borel subsets A of $\mathcal{X}$.

Total Variation Distance

Let P and Q be two probability measures defined on a set $\mathcal{X}$. Then, the total variation distance between P and Q is defined by

$$d_{\mathsf{TV}}(P,Q) \triangleq \sup_{\text{Borel } A \subseteq \mathcal{X}} |P(A) - Q(A)|$$

where the supermum is taken w.r.t. all the Borel subsets A of $\mathcal{X}$.
If $\mathcal{X}$ is a countable set then this definition is simplified to

$$d_{\mathsf{TV}}(P,Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| = \frac{||P - Q||_1}{2}$$

so the total variation distance is equal to one-half of the L_1-distance between the two probability distributions.

Total Variation Distance

Let P and Q be two probability measures defined on a set $\mathcal{X}$. Then, the total variation distance between P and Q is defined by

$$d_{\mathsf{TV}}(P,Q) \triangleq \sup_{\text{Borel } A \subseteq \mathcal{X}} |P(A) - Q(A)|$$

where the supermum is taken w.r.t. all the Borel subsets A of $\mathcal{X}$.
If $\mathcal{X}$ is a countable set then this definition is simplified to

$$d_{\mathsf{TV}}(P,Q) = \frac{1}{2} \sum_{x \in \mathcal{X}} |P(x) - Q(x)| = \frac{||P - Q||_1}{2}$$

so the total variation distance is equal to one-half of the L_1-distance between the two probability distributions.

Question: How to get bounds on the total variation distance and the entropy difference for the Poisson approximation ?

Chen-Stein Method

The Chen-Stein method forms a powerful probabilistic tool to calculate error bounds for the Poisson approximation (Chen 1975).

This method is based on the simple property of the Poisson distribution where $Z \sim \mathrm{Po}(\lambda)$ with $\lambda \in (0, \infty)$ if and only if

$$\lambda \, \mathbb{E}[f(Z+1)] - \mathbb{E}[Z \, f(Z)] = 0$$

for all bounded functions f that are defined on $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$.

Chen-Stein Method

The Chen-Stein method forms a powerful probabilistic tool to calculate error bounds for the Poisson approximation (Chen 1975).

This method is based on the simple property of the Poisson distribution where $Z \sim \mathrm{Po}(\lambda)$ with $\lambda \in (0, \infty)$ if and only if

$$\lambda \, \mathbb{E}[f(Z+1)] - \mathbb{E}[Z \, f(Z)] = 0$$

for all bounded functions f that are defined on $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$.

This method provides a rigorous analytical treatment, via error bounds, to the case where W has approximately a Poisson distribution $\mathrm{Po}(\lambda)$. The idea behind this method is to treat analytically (by bounds)

$$\lambda \, \mathbb{E}[f(W+1)] - \mathbb{E}[W \, f(W)]$$

which is shown to be close to zero for a function f as above.

The following theorem relies on the Chen-Stein method:

Bounds on the Total Variation Distance for Poisson Approximation, [Barbour & 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 $i \in \{1, \ldots, n\}$, 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 $a, b \in \mathbb{R}$.

Generalization of the Upper Bound on the Total Variation Distance for a Sum of Dependent Bernoulli Random Variables

Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

where it is assumed that $\lambda \in (0, \infty)$. Note that W is a sum of (possibly dependent and non-identically distributed) Bernoulli RVs $\{X_\alpha\}_{\alpha \in I}$.

Generalization of the Upper Bound on the Total Variation Distance for a Sum of Dependent Bernoulli Random Variables

Let I be a countable index set, and for $\alpha \in I$, let X_α be a Bernoulli random variable with

$$p_\alpha \triangleq \mathbb{P}(X_\alpha = 1) = 1 - \mathbb{P}(X_\alpha = 0) > 0.$$

Let

$$W \triangleq \sum_{\alpha \in I} X_\alpha, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{\alpha \in I} p_\alpha$$

where it is assumed that $\lambda \in (0, \infty)$. Note that W is a sum of (possibly dependent and non-identically distributed) Bernoulli RVs $\{X_\alpha\}_{\alpha \in I}$.

For every $\alpha \in I$, let B_α be a subset of I that is chosen such that $\alpha \in B_\alpha$. This subset is interpreted [Arratia et al., 1989] as the neighborhood of dependence for α where X_α is independent or weakly dependent of all of the X_β for $\beta \notin B_\alpha$.

Generalization (Cont.)

Furthermore, the following coefficients were defined by [Arratia et al., 1989]:

$$b_1 \triangleq \sum_{\alpha\in I}\sum_{\beta\in B_\alpha} p_\alpha p_\beta$$

$$b_2 \triangleq \sum_{\alpha\in I}\sum_{\alpha\neq\beta\in B_\alpha} p_{\alpha,\beta}, \quad p_{\alpha,\beta} \triangleq \mathbb{E}(X_\alpha X_\beta)$$

$$b_3 \triangleq \sum_{\alpha\in I} s_\alpha, \quad s_\alpha \triangleq \mathbb{E}\Big|\mathbb{E}(X_\alpha - p_\alpha \,|\, \sigma(\{X_\beta\})_{\beta\in I\setminus B_\alpha})\Big|$$

where $\sigma(\cdot)$ in the conditioning of last equality denotes the σ-algebra that is generated by the random variables inside the parenthesis.

Generalization (Cont.)

Then, the following upper bound on the total variation distance holds:

$$d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq (b_1 + b_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b_3\left(1 \wedge \frac{1.4}{\sqrt{\lambda}}\right).$$

Generalization (Cont.)

Then, the following upper bound on the total variation distance holds:

$$d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq (b_1 + b_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b_3\left(1 \wedge \frac{1.4}{\sqrt{\lambda}}\right).$$

Proof Methodology: The Chen-Stein method [Arratia et al., 1989].

Generalization (Cont.)

Then, the following upper bound on the total variation distance holds:

$$d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq (b_1 + b_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right) + b_3\left(1 \wedge \frac{1.4}{\sqrt{\lambda}}\right).$$

Proof Methodology: The Chen-Stein method [Arratia et al., 1989].

Special Case

If $\{X_\alpha\}_{\alpha \in I}$ are independent and $\alpha \triangleq \{1, \ldots, n\}$, then $b_1 = \sum_{i=1}^{n} p_i^2$ and $b_2 = b_3 = 0$, which gives the previous upper bound on the total variation distance.

$\Rightarrow p_i = \frac{\lambda}{n}$ for all $i \in \{1, \ldots, n\}$ (i.e., a sum of n i.i.d. Bernoulli random variables with parameter $\frac{\lambda}{n}$) gives that

$$d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \leq \frac{\lambda^2}{n} \searrow 0.$$

Total Variation Distance vs. Entropy Difference

An upper bound on the total variation distance between W and the Poisson random variable $Z \sim \mathrm{Po}(\lambda)$ was introduced. This bound was derived by Arratia et al. (1989) via the Chen-Stein method.

Total Variation Distance vs. Entropy Difference

An upper bound on the total variation distance between W and the Poisson random variable $Z \sim \mathrm{Po}(\lambda)$ was introduced. This bound was derived by Arratia et al. (1989) via the Chen-Stein method.

Questions

1. Does a small total variation distance between two probability distributions ensure that their entropies are close ?

Total Variation Distance vs. Entropy Difference

An upper bound on the total variation distance between W and the Poisson random variable $Z \sim \mathrm{Po}(\lambda)$ was introduced. This bound was derived by Arratia et al. (1989) via the Chen-Stein method.

Questions

1. Does a small total variation distance between two probability distributions ensure that their entropies are close ?
2. Can one get, under some conditions, a bound on the difference between the entropies in terms of the total variation distance ?

Total Variation Distance vs. Entropy Difference

An upper bound on the total variation distance between W and the Poisson random variable $Z \sim \mathrm{Po}(\lambda)$ was introduced. This bound was derived by Arratia et al. (1989) via the Chen-Stein method.

Questions

1. Does a small total variation distance between two probability distributions ensure that their entropies are close ?
2. Can one get, under some conditions, a bound on the difference between the entropies in terms of the total variation distance ?
3. Can one get a bound on the difference between the entropy $H(W)$ and the entropy of the Poisson RV $Z \sim \mathrm{Po}(\lambda)$ (with $\lambda = \sum p_i$) in terms of the bound on the total variation distance ?

Total Variation Distance vs. Entropy Difference

An upper bound on the total variation distance between W and the Poisson random variable $Z \sim \mathrm{Po}(\lambda)$ was introduced. This bound was derived by Arratia et al. (1989) via the Chen-Stein method.

Questions

1. Does a small total variation distance between two probability distributions ensure that their entropies are also close ?
2. Can one get, under some conditions, a bound on the difference between the entropies in terms of the total variation distance ?
3. Can one get a bound on the difference between the entropy $H(W)$ and the entropy of the Poisson RV $Z \sim \mathrm{Po}(\lambda)$ (with $\lambda = \sum p_i$) in terms of the bound on the total variation distance ?
4. How the entropy of the Poisson RV can be calculated efficiently (by definition, it becomes an infinite series) ?

Question 1

Does a small total variation distance between two probability distributions ensure that their entropies are also close ?

Question 1

Does a small total variation distance between two probability distributions ensure that their entropies are also close ?

Answer 1

Answer: In general, NO. The total variation distance between two probability distributions may be arbitrarily small whereas the difference between the two entropies is arbitrarily large.

Question 1

Does a small total variation distance between two probability distributions ensure that their entropies are also close ?

Answer 1

Answer: In general, NO. The total variation distance between two probability distributions may be arbitrarily small whereas the difference between the two entropies is arbitrarily large.

A Possible Example [Ho & Yeung, ISIT 2007]

For a fixed $L \in \mathbb{N}$, let $P = (p_1, p_2, \ldots, p_L)$ be a probability mass function. For $M \geq L$, let

$$Q = \left(p_1 - \frac{p_1}{\sqrt{\log M}}, p_2 + \frac{p_1}{M\sqrt{\log M}}, \ldots, p_L + \frac{p_1}{M\sqrt{\log M}}, \frac{p_1}{M\sqrt{\log M}}, \ldots, \frac{p_1}{M\sqrt{\log M}} \right).$$

Example (cont.)

Then

$$d_{\mathsf{TV}}(P,Q) = \frac{2\,p_1}{\sqrt{\log M}} \searrow 0$$

when $M \to \infty$. On the other hand, for a sufficiently large M,

$$H(Q) - H(P) \approx p_1 \left(1 - \frac{L}{M}\right) \sqrt{\log M} \nearrow \infty.$$

Example (cont.)

Then

$$d_{\mathsf{TV}}(P,Q) = \frac{2\,p_1}{\sqrt{\log M}} \searrow 0$$

when $M \to \infty$. On the other hand, for a sufficiently large M,

$$H(Q) - H(P) \approx p_1 \left(1 - \frac{L}{M}\right) \sqrt{\log M} \nearrow \infty.$$

Conclusion

It is easy to construct two random variables X and Y whose supports are finite although one of their supports is not bounded such that the total variation distance between their distributions is arbitrarily close to zero whereas the difference between their entropies is arbitrarily large.

$\Rightarrow$ If the alphabet size of X or Y is not bounded, then

$d_{\mathsf{TV}}(P_X, P_Y) \to 0$ does **NOT** imply that $|H(P_X) - H(P_Y)| \to 0$.

Question 2

Can one get, under some conditions, a bound on the difference between the entropies in terms of the total variation distance ?

Question 2

Can one get, under some conditions, a bound on the difference between the entropies in terms of the total variation distance ?

Answer 2

Yes, it is true when the alphabet sizes of X and Y are finite and fixed. The following theorem is an L_1 bound on the entropy (see [Cover and Thomas, Theorem 17.3.3] or [Csiszár and Körner, Lemma 2.7]):

Let P and Q be two probability mass functions on a finite set $\mathcal{X}$ such that

$$||P - Q||_1 \triangleq \sum_{x \in \mathcal{X}} |P(x) - Q(x)| \leq \frac{1}{2}.$$

Then, the difference between their entropies to the base e satisfies

$$|H(P) - H(Q)| \leq ||P - Q||_1 \ \log\left(\frac{|\mathcal{X}|}{||P - Q||_1}\right).$$

Answer 2 (cont.)

This inequality can be written in the form

$$|H(P) - H(Q)| \leq 2d_{\mathsf{TV}}(P,Q) \ \log\left(\frac{|\mathcal{X}|}{2d_{\mathsf{TV}}(P,Q)}\right)$$

when $d_{\mathsf{TV}}(P,Q) \leq \frac{1}{4}$, thus bounding the difference of the entropies in terms of the total variation distance when the alphabet sizes are finite and fixed.

Answer 2 (cont.)

This inequality can be written in the form

$$|H(P) - H(Q)| \leq 2d_{\mathsf{TV}}(P,Q) \ \log\left(\frac{|\mathcal{X}|}{2d_{\mathsf{TV}}(P,Q)}\right)$$

when $d_{\mathsf{TV}}(P,Q) \leq \frac{1}{4}$, thus bounding the difference of the entropies in terms of the total variation distance when the alphabet sizes are finite and fixed.

Further Improvements of the Bound on the Entropy Difference in Terms of the Total Variation Distance

This inequality was further improved by S. Wai Ho & R. Yeung (see Theorem 6 and its refinement in Theorem 7 in their paper entitled "The interplay between entropy and variation distance," IEEE Trans. on Information Theory, Dec. 2010.)

Question 3

Can one get a bound on the difference between the entropy $H(W)$ and the entropy of the Poisson RV $Z \sim \mathrm{Po}(\lambda)$ (with $\lambda = \sum p_i$) in terms of the bound on the total variation distance ?

Question 3

Can one get a bound on the difference between the entropy $H(W)$ and the entropy of the Poisson RV $Z \sim \mathrm{Po}(\lambda)$ (with $\lambda = \sum p_i$) in terms of the bound on the total variation distance ?

Answer 3

From the reply to the first question, since the Poisson distribution is defined over an infinitely countable set, it is not clear that the difference

$$H(W) - H(\mathrm{Po}(\lambda))$$

can be bounded in terms of the total variation distance $d_{\mathrm{TV}}(W, \mathrm{Po}(\lambda))$. But, we provide an affirmative answer and derive such a bound.

New Result (Answer 3 (cont.))

Let I be an arbitrary finite index set with $m \triangleq |I|$. Under the setting of this work and the notation used in slides 9–10, let

$$a(\lambda) \triangleq 2\left[(b_1+b_2)\left(\frac{1-e^{-\lambda}}{\lambda}\right)+b_3\left(1 \wedge \frac{1.4}{\sqrt{\lambda}}\right)\right]$$

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

where, in the last equality, $(x)_+ \triangleq \max\{x,0\}$ for every $x \in \mathbb{R}$. Let $Z \sim \mathrm{Po}(\lambda)$ be a Poisson random variable with mean λ. If $a(\lambda) \leq \frac{1}{2}$ and $\lambda \triangleq \sum_{\alpha \in I} p_\alpha \leq m-1$, then the difference between the entropies (to the base e) of Z and W satisfies the following inequality:

$$|H(Z)-H(W)| \leq a(\lambda)\,\log\left(\frac{m+2}{a(\lambda)}\right)+b(\lambda).$$

A Tighter Bound for Independent Summands

If the summands $\{X_\alpha\}_{\alpha\in I}$ are also independent, then

$$0 \leq H(Z) - H(W) \leq g(\underline{p}) \log\left(\frac{m+2}{g(\underline{p})}\right) + b(\lambda)$$

if $g(\underline{p}) \leq \frac{1}{2}$ and $\lambda \leq m-1$, where

$$g(\underline{p}) \triangleq 2\theta \min\left\{1 - e^{-\lambda}, \frac{3}{4e(1-\sqrt{\theta})^{3/2}}\right\}$$

$$\underline{p} \triangleq \{p_\alpha\}_{\alpha\in I}, \quad \lambda \triangleq \sum_{\alpha\in I} p_\alpha$$

$$\theta \triangleq \frac{1}{\lambda}\sum_{\alpha\in I} p_\alpha^2.$$

The proof relies on the upper bounds on the total variation distance by Barbour and Hall (1984) and by Cekanavičius and Roos (2006), the maximum entropy result, and the derivation of the general bound.

Question 4

How the entropy of the Poisson RV can be calculated efficiently ?

Question 4

How the entropy of the Poisson RV can be calculated efficiently ?

Answer 4

The entropy of a random variable $Z \sim \mathrm{Po}(\lambda)$ is equal to

$$H(Z) = \lambda \log\left(\frac{e}{\lambda}\right) + \sum_{k=1}^{\infty} \frac{\lambda^k e^{-\lambda} \log k!}{k!}$$

so the entropy of the Poisson distribution (in nats) is expressed in terms of an infinite series that has no closed form.

Answer 4 (Cont.)

Sequences of simple upper and lower bounds on this entropy, which are asymptotically tight, were derived by Adell et al. [IEEE Trans. on IT, May 2010]. In particular, from Theorem 2 in this paper

$$-\frac{31}{24\lambda^2}-\frac{33}{20\lambda^3}-\frac{1}{20\lambda^4}\leq H(Z)-\frac{1}{2}\log(2\pi e\lambda)+\frac{1}{12\lambda}\leq\frac{5}{24\lambda^2}+\frac{1}{60\lambda^3}$$

which gives tight bounds on the entropy of $Z\sim\mathrm{Po}(\lambda)$ for $\lambda\gg 1$.

- For $\lambda\geq 20$, the entropy of Z is approximated by the average of its above upper and lower bounds, asserting that the relative error of this approximation is less than 0.1% (and it scales like $\frac{1}{\lambda^2}$).
- For $\lambda\in(0,20)$, a truncation of the above infinite series after its first $\lceil 10\lambda\rceil$ terms gives an accurate approximation.

Example: Random Graphs

The setting of this problem was introduced by Arratia et al. (1989) in the context of applications of the Chen-Stein method.

Example: Random Graphs

The setting of this problem was introduced by Arratia et al. (1989) in the context of applications of the Chen-Stein method.

Problem Setting

- On the cube $\{0,1\}^n$, assume that each of the $n2^{n-1}$ edges is assigned a random direction by tossing a fair coin.

Example: Random Graphs

The setting of this problem was introduced by Arratia et al. (1989) in the context of applications of the Chen-Stein method.

Problem Setting

- On the cube $\{0,1\}^n$, assume that each of the $n2^{n-1}$ edges is assigned a random direction by tossing a fair coin.
- Let $k \in \{0,1,\ldots,n\}$ be fixed, and denote by $W \triangleq W(k,n)$ the random variable that is equal to the number of vertices at which exactly k edges point outward (so $k=0$ corresponds to the event where all n edges, from a certain vertex, point inward).

Example: Random Graphs

The setting of this problem was introduced by Arratia et al. (1989) in the context of applications of the Chen-Stein method.

Problem Setting

- On the cube $\{0,1\}^n$, assume that each of the $n2^{n-1}$ edges is assigned a random direction by tossing a fair coin.
- Let $k \in \{0,1,\ldots,n\}$ be fixed, and denote by $W \triangleq W(k,n)$ the random variable that is equal to the number of vertices at which exactly k edges point outward (so $k=0$ corresponds to the event where all n edges, from a certain vertex, point inward).
- Let I be the set of all 2^n vertices, and X_α be the indicator that vertex $\alpha \in I$ has exactly k of its edges directed outward. Then $W = \sum_{\alpha \in I} X_\alpha$ with $X_\alpha \sim \mathrm{Bern}(p), \quad p = 2^{-n}\binom{n}{k}, \quad \forall \alpha \in I.$

Example: Random Graphs

The setting of this problem was introduced by Arratia et al. (1989) in the context of applications of the Chen-Stein method.

Problem Setting

- On the cube $\{0,1\}^n$, assume that each of the $n2^{n-1}$ edges is assigned a random direction by tossing a fair coin.
- Let $k \in \{0,1,\ldots,n\}$ be fixed, and denote by $W \triangleq W(k,n)$ the random variable that is equal to the number of vertices at which exactly k edges point outward (so $k=0$ corresponds to the event where all n edges, from a certain vertex, point inward).
- Let I be the set of all 2^n vertices, and X_α be the indicator that vertex $\alpha \in I$ has exactly k of its edges directed outward. Then $W = \sum_{\alpha \in I} X_\alpha$ with $X_\alpha \sim \mathrm{Bern}(p), \quad p = 2^{-n}\binom{n}{k}, \quad \forall \alpha \in I.$
- Problem: Estimate the entropy $H(W)$.

Example: Random Graphs (Cont.)

- The problem setting implies that $\lambda = \binom{n}{k}$ (since $|I| = 2^n$).
- The neighborhood of dependence of a vertex $\alpha \in I$, denoted by B_α, is the set of vertices that are directly connected to α (including α itself since it is required that $\alpha \in B_\alpha$). Hence, $b_1 = 2^{-n}(n+1)\binom{n}{k}^2$.
- If α and β are two vertices that are connected by an edge, then a conditioning on the direction of this edge gives that

$$p_{\alpha,\beta} \triangleq \mathbb{E}(X_\alpha X_\beta) = 2^{2-2n} \binom{n-1}{k} \binom{n-1}{k-1}$$

for every $\alpha \in I$ and $\beta \in B_\alpha \setminus \{\alpha\}$, so by definition (see slide 10),

$$b_2 = n\, 2^{2-n} \binom{n-1}{k} \binom{n-1}{k-1}.$$

Example: Random Graphs (Cont.)

- $b_3 = 0$ (since the conditional expectation of X_α given $(X_\beta)_{\beta \in I \setminus B_\alpha}$ is, similarly to the un-conditional expectation, equal to p_α). The directions of the edges outside the neighborhood of dependence of α are irrelevant to the directions of the edges connecting the vertex α.
- In the following, the bound on the entropy difference is applied to get a rigorous error bound on the Poisson approximation of the entropy $H(W)$.
- By symmetry, the cases with $W(k,n)$ and $W(n-k,n)$ are equivalent, so

$$H\big(W(k,n)\big) = H\big(W(n-k,n)\big).$$

Numerical Results for the Example of Random Graphs

Table: Numerical results for the Poisson approximations of the entropy $H(W)$ ($W = W(k,n)$) by the entropy $H(Z)$ where $Z \sim \mathrm{Po}(\lambda)$, jointly with the associated error bounds of these approximations. These error bounds are calculated from the new theorem (see slide 18).

n	k	$\lambda = \binom{n}{k}$	$H(W) \approx$	Maximal relative error
30	26	$2.741 \cdot 10^4$	6.528 nats	0.94%
30	25	$1.425 \cdot 10^5$	7.353 nats	4.33%
100	95	$7.529 \cdot 10^7$	10.487 nats	$1.6 \cdot 10^{-19}$
100	85	$2.533 \cdot 10^{17}$	21.456 nats	$2.6 \cdot 10^{-10}$
100	75	$2.425 \cdot 10^{23}$	28.342 nats	$1.9 \cdot 10^{-4}$
100	70	$2.937 \cdot 10^{25}$	30.740 nats	2.1%

Generalization: Bounds on the Entropy for a Sum of Non-Negative, Integer-Valued and Bounded Random Variables

- A generalization of the earlier bound was derived in the full-paper version, considering the accuracy of the Poisson approximation for the entropy of a sum of non-negative, integer-valued and bounded random variables.
- This generalization is enabled via the combination of the proof of the previous bound, considering the entropy of the sum of Bernoulli random variables, with the approach of Serfling (Section 7 in his paper from 1978).

Full Paper Version

- This talk presents in part the first half of the paper:
  I. Sason, "An information-theoretic perspective of the Poisson approximation via the Chen-Stein method," submitted to the IEEE Trans. on Information Theory, June 2012. [Online]. Available: `http://arxiv.org/abs/1206.6811`.
- A generalization of the bounds that considers the accuracy of the Poisson approximation for the entropy of a sum of non-negative, integer-valued and bounded random variables is introduced in the full paper.
- The second part of this paper derives lower bounds on the total variation distance, relative entropy and other measures that are not covered in this talk.

Bibliography

- J. A. Adell, A. Lekouna and Y. Yu, "Sharp bounds on the entropy of the Poisson law and related quantities," *IEEE Trans. on Information Theory*, vol. 56, no. 5, pp. 2299–2306, May 2010.
- R. Arratia, L. Goldstein and L. Gordon, "Poisson approximation and the Chen-Stein method," *Statistical Science*, vol. 5, no. 4, pp. 403–424, November 1990.
- A. D. Barbour and P. Hall, "On the rate of Poisson Convergence," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 95, no. 3, pp. 473–480, 1984.
- A. D. Barbour, L. Holst and S. Janson, *Poisson Approximation*, Oxford University Press, 1992.
- A. D. Barbour and L. H. Y. Chen, *An Introduction to Stein's Method*, Lecture Notes Series, Institute for Mathematical Sciences, Singapore University Press and World Scientific, 2005.

Bibliography (Cont.)

- V. Čekanavičius and B. Roos, "An expansion in the exponent for compound binomial approximations," *Lithuanian Mathematical Journal*, vol. 46, no. 1, pp. 54–91, 2006.
- L. H. Y. Chen, "Poisson approximation for dependent trials," *Annals of Probability*, vol. 3, no. 3, pp. 534–545, June 1975.
- T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, second edition, 2006.
- I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
- P. Harremoës, "Binomial and Poisson distributions as maximum entropy distributions," *IEEE Trans. on Information Theory*, vol. 47, no. 5, pp. 2039–2041, July 2001.
- P. Harremoës, O. Johnson and I. Kontoyiannis, "Thinning, entropy and the law of thin numbers," *IEEE Trans. on Information Theory*, vol. 56, no. 9, pp. 4228–4244, September 2010.

Bibliography (Cont.)

- S. W. Ho and R. W. Yeung, "The interplay between entropy and variational distance," *IEEE Trans. on Information Theory*, vol. 56, no. 12, pp. 5906–5929, December 2010.
- O. Johnson, *Information Theory and the Central Limit Theorem*, Imperial College Press, 2004.
- O. Johnson, "Log-concavity and maximum entropy property of the Poisson distribution," *Stochastic Processes and their Applications*, vol. 117, no. 6, pp. 791–802, November 2006.
- O. Johnson, I. Kontoyiannis and M. Madiman, "A criterion for the compound Poisson distribution to be maximum entropy," *Proceedings 2009 IEEE International Symposium on Information Theory*, pp. 1899–1903, Seoul, South Korea, July 2009.
- S. Karlin and Y. Rinott, "Entropy inequalities for classes of probability distributions I: the univariate case," *Advances in Applied Probability*, vol. 13, no. 1, pp. 93–112, March 1981.

Bibliography (Cont.)

- I. Kontoyiannis, P. Harremoës and O. Johnson, "Entropy and the law of small numbers," *IEEE Trans. on Information Theory*, vol. 51, no. 2, pp. 466–472, February 2005.
- I. Kontoyiannis, P. Harremoës, O. Johnson and M. Madiman, "Information-theoretic ideas in Poisson approximation and concentration," slides of a short course, September 2006.
- R. J. Serfling, "Some elementary results on Poisson approximation in a sequence of Bernoulli trials," *Siam Review*, vol. 20, pp. 567–579, July 1978.
- L. A. Shepp and I. Olkin, "Entropy of the sum of independent Bernoulli random variables and the multinomial distribution," *Contributions to Probability*, pp. 201–206, Academic Press, New York, 1981.
- Y. Yu, "Monotonic convergence in an information-theoretic law of small numbers," *IEEE Trans. on Information Theory*, vol. 55, no. 12, pp. 5412–5422, December 2009.