New Lower Bounds on the Total Variation Distance and Relative Entropy for the Poisson Approximation

Igal Sason

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

ITA 2013, San-Diego, CA, USA
February 2013.

Problem Statement

- Let X_i be n Bernoulli random variables with

$$p_i \triangleq \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) > 0.$$

Problem Statement

- Let X_i be n Bernoulli random variables with

$$p_i \triangleq \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) > 0.$$

- Let

$$W \triangleq \sum_{i=1}^{n} X_i, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{i=1}^{n} p_i$$

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

Problem Statement

- Let X_i be n Bernoulli random variables with

$$p_i \triangleq \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) > 0.$$

- Let

$$W \triangleq \sum_{i=1}^{n} X_i, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{i=1}^{n} p_i$$

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

Problem Statement

- Let X_i be n Bernoulli random variables with

$$p_i \triangleq \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) > 0.$$

- Let

$$W \triangleq \sum_{i=1}^{n} X_i, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{i=1}^{n} p_i$$

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

Problem: Getting tight bounds for the total variation distance and relative entropy between W and $\text{Po}(\lambda)$.

Problem Statement

- Let X_i be n Bernoulli random variables with

$$p_i \triangleq \mathbb{P}(X_i = 1) = 1 - \mathbb{P}(X_i = 0) > 0.$$

- Let

$$W \triangleq \sum_{i=1}^{n} X_i, \quad \lambda \triangleq \mathbb{E}(W) = \sum_{i=1}^{n} p_i$$

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

Problem: Getting tight bounds for the total variation distance and relative entropy between W and $\text{Po}(\lambda)$.

In this talk: New (improved) lower bounds on the total variation distance and relative entropy are derived, and their possible use is outlined.

Related Work

1. A. D. Barbour and P. Hall, "On the rate of Poisson convergence," *Math. Proc. of the Cambridge Philosophical Society*, 1984.
2. Barbour et al., "Compound Poisson approximation via information functionals," *EJP*, August 2010.
3. Kontoyiannis et al., "Entropy and the law of small numbers," *IEEE Trans. on IT*, Feb. 2005.
4. Harremoës and Ruzankin, "Rate of convergence to Poisson law in terms of information divergence," *IEEE Trans. on IT*, Sept. 2004.
5. P. Harremoës and C. Vignat, "Lower bounds on information divergence," Feb. 2011 (`http://arxiv.org/pdf/1102.2536.pdf`).
6. 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 $\mathcal{X}$.
- 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 it is simplified to

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

$\Rightarrow$ The total variation distance is equal to one-half of the L_1-distance between the two probability distributions.

Theorem 1 - 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 $i \in \{1, \ldots, n\}$, 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 $a, b \in \mathbb{R}$.

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

Chen-Stein Method

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

Key Idea:
$Z \sim \text{Po}(\lambda)$ with $\lambda \in (0, \infty)$ if and only if

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

for <u>all bounded functions</u> 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).

Key Idea:
$Z \sim \text{Po}(\lambda)$ 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 $\mathbb{N}_0 \triangleq \{0, 1, \ldots\}$.

The idea behind this method is to treat analytically (by bounds) the functional

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

$\Rightarrow$ leads to some rigorous bounds on $d_{\text{TV}}(W, Z)$ as a function of the choice of the bounded function $f : \mathbb{N}_0 \to \mathbb{R}$.

Chen-Stein Method (2)

The proof of Theorem 1 (Barbour and Hall, 1984) relies on the choice

$$f(k) \triangleq (k-\lambda) \exp\left(-\frac{(k-\lambda)^2}{\lambda}\right), \quad \forall\, k \in \mathbb{N}_0.$$

Chen-Stein Method (2)

The proof of Theorem 1 (Barbour and Hall, 1984) relies on the choice

$$f(k) \triangleq (k-\lambda)\, \exp\left(-\frac{(k-\lambda)^2}{\lambda}\right), \quad \forall\, k \in \mathbb{N}_0.$$

The new (and rather surprising) result is that the lower bound on the total variation distance is much improved by generalizing f to be

$$f(k) \triangleq (k-\alpha_1)\, \exp\left(-\frac{(k-\alpha_2)^2}{\theta\lambda}\right), \quad \forall\, k \in \mathbb{N}_0$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\theta \in \mathbb{R}^+$ are optimized to get the tightest lower bound on $d_{\mathsf{TV}}(W, Z)$. But, it complicates the analysis (Th. 2).

Chen-Stein Method (2)

The proof of Theorem 1 (Barbour and Hall, 1984) relies on the choice

$$f(k) \triangleq (k-\lambda)\,\exp\left(-\frac{(k-\lambda)^2}{\lambda}\right), \quad \forall\, k \in \mathbb{N}_0.$$

The new (and rather surprising) result is that the lower bound on the total variation distance is much improved by generalizing f to be

$$f(k) \triangleq (k-\alpha_1)\,\exp\left(-\frac{(k-\alpha_2)^2}{\theta\lambda}\right), \quad \forall\, k \in \mathbb{N}_0$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ and $\theta \in \mathbb{R}^+$ are optimized to get the tightest lower bound on $d_{\mathsf{TV}}(W,Z)$. But, it complicates the analysis (Th. 2).

Special case: $\alpha_1 = \alpha_2 \triangleq \lambda$ and optimizing θ leads to a simplified lower bound (Th. 3) that achieves almost the same improvement for $\lambda \geq 10$.

Theorem 2 - Improved Lower Bound on the Total Variation Distance

In the setting of Theorem 1, the total variation distance between the probability distribution of W and $\mathsf{Po}(\lambda)$ satisfies

$$K_1(\lambda) \sum_{i=1}^{n} p_i^2 \le d_{\mathsf{TV}}(P_W, \mathsf{Po}(\lambda)) \le \left(\frac{1-e^{-\lambda}}{\lambda}\right) \sum_{i=1}^{n} p_i^2$$

where
$$K_1(\lambda) \triangleq \sup_{\substack{\alpha_1, \alpha_2 \in \mathbb{R}, \\ \alpha_2 \le \lambda + \frac{3}{2}, \\ \theta > 0}} \left(\frac{1 - h_\lambda(\alpha_1, \alpha_2, \theta)}{2\, g_\lambda(\alpha_1, \alpha_2, \theta)}\right)$$

for some functions h_λ, g_λ that are given in Th. 2 of the conference paper.

Theorem 3 - Simple Lower Bound on the Total Variation Distance

Under the assumptions in Theorem 1, 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

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

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

Theorem 4 - Improved Lower Bound on the Relative Entropy

In the setting of Th. 1, the divergence between the probability distribution of W and the Poisson distribution with mean $\lambda = \mathbb{E}(W)$ satisfies:

$$K_2(\lambda) \left(\sum_{i=1}^{n} p_i^2\right)^2 \leq D\big(P_W || \text{Po}(\lambda)\big) \leq \frac{1}{\lambda} \sum_{i=1}^{n} \frac{p_i^3}{1-p_i}$$

where

$$K_2(\lambda) \triangleq m(\lambda) \left(K_1(\lambda)\right)^2$$

$$m(\lambda) \triangleq \begin{cases} \left(\frac{1}{2e^{-\lambda}-1}\right) \log\left(\frac{1}{e^{\lambda}-1}\right) & \text{if } \lambda \in (0, \log 2) \\ 2 & \text{if } \lambda \geq \log 2. \end{cases}$$

Theorem 4 - Proof Outline

- The lower bound on the relative entropy is based on new lower bound on the total variation distance and the distribution-dependent refinement of Pinsker's inequality (Ordentlich & Weinberger, IEEE Trans. on IT, 2005).

Theorem 4 - Proof Outline

- The lower bound on the relative entropy is based on new lower bound on the total variation distance and the distribution-dependent refinement of Pinsker's inequality (Ordentlich & Weinberger, IEEE Trans. on IT, 2005).
- The upper bound on the relative entropy was derived by Kontoyiannis et al. (IEEE Trans. on IT, 2005).

Theorem 4 - Proof Outline

- The lower bound on the relative entropy is based on new lower bound on the total variation distance and the distribution-dependent refinement of Pinsker's inequality (Ordentlich & Weinberger, IEEE Trans. on IT, 2005).
- The upper bound on the relative entropy was derived by Kontoyiannis et al. (IEEE Trans. on IT, 2005).

Corollary

If $\{X_i\}$ are i.i.d. binary RVs, $W \triangleq \sum_{i=1}^{n} X_i$, and $Z \sim \text{Po}(\lambda)$ with $\lambda = \sum_{i=1}^{n} p_i$, then

$$d_{\text{TV}}(W, Z) = O\left(\frac{1}{n}\right), \qquad D\big(P_W || \text{Po}(\lambda)\big) = O\left(\frac{1}{n^2}\right).$$

Improved Lower Bound on the Relative Entropy (2)

- The combination of the original lower bound on $d_{\mathsf{TV}}(W,Z)$ (see Theorem 1) with Pinsker's inequality gives:

$$D\big(P_W \,||\, \mathsf{Po}(\lambda)\big) \geq \frac{1}{512}\left(1 \wedge \frac{1}{\lambda^2}\right) \left(\sum_{i=1}^{n} p_i^2\right)^2.$$

- The improvement of the new lower bound on the relative entropy is by a factor of $179.7 \log\big(\frac{1}{\lambda}\big)$ for $\lambda \approx 0$, by a factor of 9.22 for $\lambda \to \infty$, and at least by a factor of 6.14 for all $\lambda \in (0, \infty)$.

Personal Communications with P. Harremoës

- There exists another recent lower bound on the relative entropy (a work in preparation by Kontoyiannis et al.). The bounds are derived by different approaches.
- The two lower bounds on the relative entropy scale like $\left(\sum_{i=1}^{n} p_i^2\right)^2$ but with a different scaling factor.

Possible Uses of the New Bounds

The paper (whose shortened version is submitted to ISIT 2013):

I. Sason, "Entropy bounds for discrete random variables via coupling," submitted to the *IEEE Trans. on Information Theory*, September 2012. [Online]. Available: `http://arxiv.org/abs/1209.5259`

introduces new entropy bounds for discrete random variables via maximal coupling, providing bounds on the difference between the entropies of two discrete random variables in terms of the local and total variation distances between their probability mass functions.

Possible Uses of the New Bounds

The paper (whose shortened version is submitted to ISIT 2013):

I. Sason, “Entropy bounds for discrete random variables via coupling,” submitted to the *IEEE Trans. on Information Theory*, September 2012. [Online]. Available: `http://arxiv.org/abs/1209.5259`

introduces new entropy bounds for discrete random variables via maximal coupling, providing bounds on the difference between the entropies of two discrete random variables in terms of the local and total variation distances between their probability mass functions.

The new lower bound on the total variation distance $d_{\mathsf{TV}}(W, Z)$ was involved in a rigorous estimation of the entropy of W (via bounds).

Application: Getting numerical bounds on the sum-rate capacity of a noiseless K-user multiple-access channel with binary inputs subject to a constraint on the total power.

Possible Uses of the New Bounds (2)

- The use of the new lower bound on the relative entropy for the Poisson approximation of a sum of Bernoulli random variables is exemplified in the full paper version of this work (see Section 4.E) in the context of a problem in binary hypothesis testing.

Possible Uses of the New Bounds (2)

- The use of the new lower bound on the relative entropy for the Poisson approximation of a sum of Bernoulli random variables is exemplified in the full paper version of this work (see Section 4.E) in the context of a problem in binary hypothesis testing.
- The upper bound on the error probability depends exponentially on the relative entropy between the two probability distributions.

Possible Uses of the New Bounds (2)

- The use of the new lower bound on the relative entropy for the Poisson approximation of a sum of Bernoulli random variables is exemplified in the full paper version of this work (see Section 4.E) in the context of a problem in binary hypothesis testing.
- The upper bound on the error probability depends exponentially on the relative entropy between the two probability distributions.
- The improvement of the lower bound on the relative entropy $\Rightarrow$ Remarkable improvement in the minimal block length that ensures a fixed error probability (numerical results appear in the full paper).

Conclusions

- New lower bounds on the total variation distance between the distribution of a sum of independent Bernoulli random variables and the Poisson random variable (with the same mean) were introduced via the Chen-Stein method.
- Corresponding lower bounds on the relative entropy were introduced, based on the lower bounds on the total variation distance and an existing distribution-dependent refinement of Pinsker's inequality.
- Two uses of these bounds were outlined.
- The full paper version is available at http://arxiv.org/abs/1206.6811.
- Conference paper is available at http://arxiv.org/abs/1301.7504.