Information-Theoretic Ideas in Poisson Approximation and Concentration

Ioannis Kontoyiannis
Athens Univ Econ & Business

joint work with
P. Harremoës, O. Johnson, M. Madiman

LMS/EPSRC Short Course
Stability, Coupling Methods and Rare Events

Heriot-Watt University, Edinburgh, September 2006

Outline

1. Poisson Approximation in Relative Entropy

Motivation: Entropy and the central limit theorem
Motivation: Poisson as a maximum entropy distribution
A very simple general bound; **Examples**

2. Analogous Bounds in Total Variation

Suboptimal Poisson approximation
Optimal Compound Poisson approximation

3. Tighter Poisson Bounds for Independent Summands

A (new) discrete Fisher information; subadditivity
A log-Sobolev inequality

4. Measure Concentration and Compound Poisson Tails

The compound Poisson distributions
A log-Sobolev inequality and its info-theoretic proof
Compound Poisson concentration

Motivation: The Central Limit Theorem

Recall

$N(0,\sigma^2)$ has maximum entropy among all distributions with variance $\leq \sigma^2$ where the entropy of a RV Z with density f is

$$h(Z) := h(f) := -\int f \log f$$

The Central Limit Theorem

For IID RVs $X_1, \ldots, X_n$ with zero mean, variance σ^2, and a 'nice' density, not only $\hat{S}_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \xrightarrow{\mathcal{D}} N(0,\sigma^2)$ but in fact $h(\hat{S}_n) \uparrow h(N(0,\sigma^2))$

- $\rightsquigarrow$ Accumulation of many, small, independent random effects is maximally random (cf. second law of thermodynamics)
- $\rightsquigarrow$ Monotonicity in n indicates that the entropy is a *natural measure* for the convergence of the CLT
- $\rightsquigarrow$ This powerful intuition comes with powerful new techniques

[Linnik (1959), Brown (1982), Barron (1985), Ball-Barthe-Naor (2003),...]

Poisson Approximation: Generalities

Binomial convergence to the Poisson

If $X_1, X_2, \ldots, X_n$ are IID Bern(λ/n) [Bernoulli with parameter λ/n]

then, for large n, the distr'n of $S_n := \sum_{i=1}^{n} X_i$ is $\approx$ Po(λ) [Poisson with param λ]

General Poisson approximation

If the X_i are (possibly dependent) Bern(p_i) random variables,
then the distribution of their sum S_n is $\approx$ Po(λ) as long as:

(a) Each $E(X_i) = p_i$ is small

(b) The overall mean $E(S_n) = \sum_{i=1}^{n} p_i \approx \lambda$

(c) The X_i are weakly dependent

$\rightsquigarrow$ *Information-theoretic interpretation of this phenomenon*?

The Poisson Distribution and Entropy

Recall: the **entropy** of a discrete random variable X with distribution P is

$$H(X) = H(P) = -\sum_x P(x) \log P(x)$$

Theorem 0: Maximum Entropy

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

$$H(\text{Po}(\lambda)) = \sup \left\{ H(S_k) \; : \; S_k = \sum_{i=1}^{k} X_i, \; X_i \sim \text{indep Bern}(p_i), \; \sum_{i=1}^{k} p_i = \lambda, \; {\scriptstyle k \geq 1} \right\}$$

Proof. Messy but straightforward convexity arguments *a la*
[Mateev 1978] [Shepp & Olkin 1978] [Harremoës 2001] [Topsøe 2002] □

Measuring Distance Between Probability Distributions

Recall

The **total variation distance** between two distributions P and Q on the same discrete set S is

$$\|P - Q\|_{TV} = \frac{1}{2} \sum_{x \in S} |P(x) - Q(x)|$$

The **entropy** of a discrete random variable X with distribution P is

$$H(X) = H(P) = -\sum_{x} P(x) \log P(x)$$

The **relative entropy** (or Kullback-Leibler divergence) is

$$D(P\|Q) = \sum_{x \in S} P(x) \log \frac{P(x)}{Q(x)}$$

Pinsker's ineq: $\frac{1}{2}\|P - Q\|_{TV}^2 \le D(P\|Q)$

A Simple Poisson Approximation Bound

Theorem 1: Poisson Approximation [KHJ 05]

Suppose the X_i are (possibly dependent) Bern(p_i) random variables such that the mean of $S_n = \sum_{i=1}^n X_i$ is $E(S_n) = \sum_{i=1}^n p_i = \lambda$. Then:

The distribution P_{S_n} of S_n satisfies

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \leq \sum_{i=1}^n p_i^2 \,+\, D(P_{X_1,\ldots,X_n}\|P_{X_1}\times\cdots\times P_{X_n})$$

Note

- $D(P_{X_1,\ldots,X_n}\|P_{X_1}\times\cdots\times P_{X_n}) \geq 0$ with "=" iff the X_i are independent
- More generally, the bound is "small" iff (a)–(c) are satisfied!
- Alternatively,

$$D(P_{X_1,\ldots,X_n}\|P_{X_1}\times\cdots\times P_{X_n}) = \sum_{i=1}^n H(X_i) - H(X_1,\ldots,X_n)$$

Elementary Properties of $D(P\|Q)$

Properties

i. Data processing inequality: $D(P_{g(X)}\|P_{g(Y)}) \leq D(P_X\|P_Y)$

Proof. By Jensen's inequality:

$$\begin{aligned}
D(P_{g(X)}\|P_{g(Y)}) &= \sum_z P_{g(X)}(z) \log \frac{P_{g(X)}(z)}{P_{g(Y)}(z)} \\
&= \sum_z \Big[\sum_{x:g(x)=z} P_X(x) \Big] \log \frac{\left[\sum_{x:g(x)=z} P_X(x)\right]}{\left[\sum_{x:g(x)=z} P_Y(x)\right]} \\
&\leq \sum_z \sum_{x:g(x)=z} P_X(x) \log \frac{P_X(x)}{P_Y(x)} \\
&= D(P_X\|P_Y)
\end{aligned}$$

□

ii. $D(\mathsf{Bern}(p)\|\mathsf{Po}(p)) \leq p^2$

Proof. Elementary calculus □

Proof of Theorem 1

Letting $Z_1, Z_2, \ldots, Z_n$ be independent $\mathsf{Po}(p_i)$ and $T_n = \sum_{i=1}^n Z_i$:

$$
\begin{aligned}
&D\Big(P_{S_n}\Big\|\mathsf{Po}(\lambda)\Big) \\
&\quad = D(P_{S_n} \| P_{T_n}) \\
&\quad \leq D(P_{X_1,\ldots,X_n} \| P_{Z_1,\ldots,Z_n}) && \text{(data processing, i.)} \\
&\quad = \sum_{i=1}^n D(P_{X_i} \| P_{Z_i}) \;+\; D(P_{X_1,\ldots,X_n} \| P_{X_1} \times \cdots \times P_{X_n}) \\
&\qquad\qquad && \text{(``chain rule'': } \log(ab) = \log a + \log b) \\
&\quad \leq \sum_{i=1}^n p_i^2 \;+\; D(P_{X_1,\ldots,X_n} \| P_{X_1} \times \cdots \times P_{X_n}) && \text{(calculus, ii.)}
\end{aligned}
$$

□

Example: Independent Bernoullis

If $X_1, X_2, \ldots, X_n$ are indep Bern(p_i), Theorem 1 gives

$$D(P_{S_n} \| \mathsf{Po}(\lambda)) \;\le\; \sum_{i=1}^{n} p_i^2$$

Convergence: In view of Barbour-Hall (1984) this is
necessary and sufficient for convergence

Rate: Pinsker's ineq gives $\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \;\le\; \sqrt{2}\Big[\sum_{i=1}^n p_i^2\Big]^{1/2}$
but Le Cam (1960) gives the optimal TV rate as $O\Big(\sum_{i=1}^n p_i^2\Big)$

Question: *Can we get the optimal TV rate with IT methods??*

Two Examples

The classical Binomial/Poisson example

If $X_1, X_2, \ldots, X_n$ are IID Bern(λ/n), Theorem 1 gives

$$D(P_{S_n} \| \text{Po}(\lambda)) \;\leq\; \sum_{i=1}^{n} (\lambda/n)^2 \;=\; \lambda^2/n$$

Sufficient for convergence, but the actual rate is $O(1/n^2)$

A Markov chain example

Suppose $X_1, X_2, \ldots, X_n$ is a stationary Markov chain with transition matrix

$\begin{pmatrix} \frac{n}{n+1} & \frac{1}{n+1} \\ \frac{n-1}{n+1} & \frac{2}{n+1} \end{pmatrix}$ and each X_i having (the stationary) Bern($\frac{1}{n}$) distribution

Theorem 1 $\Rightarrow$ $\quad D(P_{S_n} \| \text{Po}(1)) \;\leq\; \dfrac{3 \log n}{n} + \dfrac{1}{n}$

Pinsker $\Rightarrow$ $\quad \| P_{S_n} - \text{Po}(1) \|_{TV} \;\leq\; 4 \left[\dfrac{\log n}{n} \right]^{1/2}$ but optimal rate is $O(1/n)$

Elementary Properties of Total Variation

TV Properties

i. TV and relative entropy are both "f-divergences"

$$D_f(P\|Q) := \sum_x Q(x)\, f\left(\frac{P(x)}{Q(x)}\right)$$

ii. Data processing ineq holds for both, same proof as before

iii. Chain rule for TV:

$$\|P \times P' - Q \times Q'\|_{TV} \leq \|P - Q\|_{TV} + \|P' - Q'\|_{TV}$$

Proof. Triangle inequality □

iv. $\|\text{Bern}(p) - \text{Po}(p)\|_{TV} \leq p^2$

Proof. Simple calculus □

v. TV is an actual norm

A Simple Poisson Approximation Bound in TV

Theorem 2: Poisson Approximation in TV [K-Madiman 06]

Suppose the X_i are *independent* Bern(p_i) random variables such that the mean of $S_n = \sum_{i=1}^n X_i$ is $E(S_n) = \sum_{i=1}^n p_i = \lambda$.

Then the distribution P_{S_n} of S_n satisfies

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \le \sum_{i=1}^n p_i^2$$

Proof. Letting $Z_1, Z_2, \ldots, Z_n$ be independent Po(p_i) and $T_n = \sum_{i=1}^n Z_i$:

$$\begin{aligned}
\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} & \\
&= \|P_{S_n} - P_{T_n}\|_{TV} \\
&\le \|P_{X_1,\ldots,X_n} - P_{Z_1,\ldots,Z_n}\|_{TV} && \text{(data processing)} \\
&\le \sum_{i=1}^n \|P_{X_i} - P_{Z_i}\|_{TV} && \text{(chain rule)} \\
&\le \sum_{i=1}^n p_i^2 && \text{(calculus)} \qquad \square
\end{aligned}$$

Example Revisited: Independent Bernoullis

Recall: If $X_1, \ldots, X_n$ are indep Bern(p_i) with $\lambda = \sum_{i=1}^n p_i$ then Thm 2 says

$$\|P_{S_n} - \text{Po}(\lambda)\|_{TV} \leq \sum_{i=1}^n p_i^2$$

& from Barbour-Hall (1984): $C_1 \sum_{i=1}^n p_i^2 \leq \|P_{S_n} - \text{Po}(\lambda)\|_{TV} \leq C_2 \sum_{i=1}^n p_i^2$
so we have the right convergence rate!

For finite n: Stein's method actually yields

$$\|P_{S_n} - \text{Po}(\lambda)\|_{TV} \leq \min\left\{1, \frac{1}{\lambda}\right\} \sum_{i=1}^n p_i^2,$$

which is much better for large λ

E.g. if all $p_i = \frac{1}{\sqrt{n}}$ then $\lambda = \sqrt{n}$ and our bound $= 1$
whereas Stein's method yields the bound $1/\sqrt{n}$

Corollary: Generalization to Dependent RVs

Corollary: General Poisson Approximation in TV [K-Madiman 06]

Suppose the X_i are (*possibly dependent*) $\mathbb{Z}_+$-valued random variables with $p_i = \Pr\{X_i = 1\}$, and let $\lambda = \sum_{i=1}^n p_i$. Then the distribution P_{S_n} of $S_n = \sum_{i=1}^n X_i$ satisfies

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \leq \sum_{i=1}^n p_i^2 + \sum_{i=1}^n E|p_i - q_i| + \sum_{i=1}^n \Pr\{X_i \geq 2\}$$

where $q_i = \Pr\{X_i = 1 | X_1, \ldots, X_{i-1}\}$

Proof of Corollary

To show: $\|P_{S_n} - \text{Po}(\lambda)\|_{TV} \leq \sum_{i=1}^{n} p_i^2 + \sum_{i=1}^{n} E|p_i - q_i| + \sum_{i=1}^{n} \Pr\{X_i \geq 2\}$

As before (data processing+chain rule):

$$\begin{aligned}\|P_{S_n} - \text{Po}(\lambda)\|_{TV} &\leq \|P_{X_1,\ldots,X_n} - P_{Z_1,\ldots,Z_n}\|_{TV} \\ &\leq \sum_{i=1}^{n} E\Big[\|P_{X_i|X_1,\ldots,X_{i-1}} - P_{Z_i}\|_{TV}\Big]\end{aligned}$$

Letting $I_i = \mathbb{I}_{\{X_i=1\}}$, by the triangle ineq:

$$\begin{aligned}\|P_{S_n} - \text{Po}(\lambda)\|_{TV} \leq &\sum_{i=1}^{n} \|P_{Z_i} - P_{I_i}\|_{TV} \\ &+ \sum_{i=1}^{n} E\Big[\|P_{I_i} - P_{I_i|X_1,\ldots,X_{i-1}}\|_{TV}\Big] \\ &+ \sum_{i=1}^{n} E\Big[\|P_{I_i|X_1,\ldots,X_{i-1}} - P_{X_i|X_i,\ldots,X_{i-1}}\|_{TV}\Big] \quad \square\end{aligned}$$

Compound Poisson Approximation

Can IT methods actually yield *optimal* bounds?
We turn to a more general problem:

Compound Binomial convergence to the compound Poisson

If $X_1, X_2, \ldots, X_n$ are IID $\sim Q$ and $I_1, I_2, \ldots, I_n$ are IID Bern(λ/n)
then, for large n, the distr'n of

$$S_n := \sum_{i=1}^{n} I_i X_i = \sum_{i=1}^{\mathrm{Bin}(n,\lambda/n)} X_i \approx \sum_{i=1}^{\mathrm{Po}(\lambda)} X_i$$

which is the **compound Poisson distr CP(λ, Q)**

General Compound Poisson approximation

For a general sum $S_n = \sum_{i=1}^{n} Y_i$ of (possibly dependent) $\mathbb{R}^d$-valued RVs Y_i
we may *hope* that the distribution of S_n is $\approx$ CP(λ, Q) as long as:

(a) Each $p_i := \Pr\{Y_i \neq 0\}$ is small

(b) The Y_i are weakly dependent

(c) The distr Q is chosen appropriately

A General Compound Poisson Approximation Result

Notes

$\rightsquigarrow$ Interpretation: Events occurring at random and in clusters

$\rightsquigarrow$ The class of dist's CP(λ, Q) is much richer that the Poisson

$\rightsquigarrow$ Depending on the choice of Q, MUCH wider class of tails, etc

$\rightsquigarrow$ CP approximation a harder problem, especially in $\mathbb{R}^d$

$\rightsquigarrow$ Same method yields a general bound in relative entropy

$\rightsquigarrow$ In search of optimality, look directly at TV bounds

Theorem 3: Compound Poisson Approximation [K-Madiman 06]

Suppose the Y_i are independent $\mathbb{R}^d$-valued RVs

Write $p_i = \Pr\{Y_i \neq 0\}$ and Q_i for the distr of $Y_i | \{Y_i \neq 0\}$

Then the distribution P_{S_n} of $S_n = \sum_{i=1}^n Y_i$ satisfies

$$\|P_{S_n} - \text{CP}(\lambda, \bar{Q})\|_{TV} \leq \sum_{i=1}^n p_i^2$$

where $\lambda = \sum_{i=1}^n p_i$ and $\bar{Q} = \sum_{i=1}^n \frac{p_i}{\lambda} Q_i$

Proof of Theorem 3

Let $Z_1, Z_2, \ldots, Z_n$ be indep CP(p_i, Q_i), so that $T_n = \sum_{i=1}^n Z_i \sim$ CP$(\lambda, \bar{Q})$

Proof of Theorem 3

Let $Z_1, Z_2, \ldots, Z_n$ be indep CP(p_i, Q_i), so that $T_n = \sum_{i=1}^n Z_i \sim$ CP$(\lambda, \bar{Q})$
By the CP defn, each Z_i can be expressed as $Z_i = \sum_{j=1}^{W_i} X_{i,j}$
where $W_i \sim$ Po(p_i) and $X_{i,j} \sim Q_i$ are all indep.

Proof of Theorem 3

Let $Z_1, Z_2, \ldots, Z_n$ be indep CP(p_i, Q_i), so that $T_n = \sum_{i=1}^n Z_i \sim$ CP$(\lambda, \bar{Q})$
By the CP defn, each Z_i can be expressed as $Z_i = \sum_{j=1}^{W_i} X_{i,j}$
where $W_i \sim$ Po(p_i) and $X_{i,j} \sim Q_i$ are all indep. Hence:

$$T_n = \sum_{i=1}^n Z_i = \sum_{i=1}^n \sum_{j=1}^{W_i} X_{i,j}$$

Proof of Theorem 3

Let $Z_1, Z_2, \ldots, Z_n$ be indep CP(p_i, Q_i), so that $T_n = \sum_{i=1}^n Z_i \sim$ CP($\lambda, \bar{Q}$)
By the CP defn, each Z_i can be expressed as $Z_i = \sum_{j=1}^{W_i} X_{i,j}$
where $W_i \sim$ Po(p_i) and $X_{i,j} \sim Q_i$ are all indep. Hence:

$$T_n = \sum_{i=1}^n Z_i = \sum_{i=1}^n \sum_{j=1}^{W_i} X_{i,j}$$

Similarly let $I_1, I_2, \ldots, I_n$ be indep Bern(p_i) and write $Y_i = I_i X_{i,1}$. Hence:

$$S_n = \sum_{i=1}^n Y_i = \sum_{i=1}^n \sum_{j=1}^{I_i} X_{i,j}$$

Proof of Theorem 3

Let $Z_1, Z_2, \ldots, Z_n$ be indep CP(p_i, Q_i), so that $T_n = \sum_{i=1}^n Z_i \sim$ CP$(\lambda, \bar{Q})$
By the CP defn, each Z_i can be expressed as $Z_i = \sum_{j=1}^{W_i} X_{i,j}$
where $W_i \sim$ Po(p_i) and $X_{i,j} \sim Q_i$ are all indep. Hence:

$$T_n = \sum_{i=1}^n Z_i = \sum_{i=1}^n \sum_{j=1}^{W_i} X_{i,j}$$

Similarly let $I_1, I_2, \ldots, I_n$ be indep Bern(p_i) and write $Y_i = I_i X_{i,1}$. Hence:

$$S_n = \sum_{i=1}^n Y_i = \sum_{i=1}^n \sum_{j=1}^{I_i} X_{i,j}$$

Then:

$$\begin{aligned}
\|P_{S_n} - \mathrm{CP}(\lambda, \bar{Q})\|_{TV} &= \|P_{S_n} - P_{T_n}\|_{TV} \\
&\leq \|P_{\{I_i\},\{X_{i,j}\}} - P_{\{W_i\},\{X_{i,j}\}}\|_{TV} && \text{(data processing)} \\
&\leq \sum_{i=1}^n \|P_{I_i} - P_{W_i}\|_{TV} && \text{(chain rule)} \\
&\leq \sum_{i=1}^n p_i^2 && \text{(calculus)} \qquad \square
\end{aligned}$$

Comments

- $\rightsquigarrow$ In general, the bound of Theorem 3 $\|P_{S_n} - \mathsf{CP}(\lambda, \bar{Q})\|_{TV} \leq \sum_{i=1}^n p_i^2$ can*not* be improved
- $\rightsquigarrow$ Here, the IT method gives the optimal rate *and* optimal constants
- $\rightsquigarrow$ Can we refine our IT methods to recover the optimal $1/\lambda$ factor in the simple Poisson case?
- $\rightsquigarrow$ Recall the earlier example: If $X_1, \ldots, X_n$ are i.i.d. Bern($\frac{1}{\sqrt{n}}$) with $\lambda = \sqrt{n}$, Stein's method gives

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \leq \frac{1}{\sqrt{n}}$$

whereas we got

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \leq 1$$

- $\rightsquigarrow$ To obtain tighter bounds, take a hint from corresponding work for the CLT [Barron, Johnson, Ball-Barthe-Naor, ...] and turn to Fisher information

A Discrete Version of Fisher Information

By analogy to the continuous case, the Fisher information of a $\mathbb{Z}_+$-valued random variable $X \sim P$ is usually defined as

$$J(X) = E\Big[\Big(\frac{P(X) - P(X-1)}{P(X)}\Big)^2\Big] = E\Big[\Big(\frac{P(X-1)}{P(X)} - 1\Big)^2\Big]$$

Problem: $J(X) = +\infty$ whenever X has finite support

Recall: $(k+1)P(k+1) = \lambda P(k)$ **iff** $P = \text{Po}(\lambda)$

Define: the **Fisher information of** $X \sim P$ via

$$J(X) = \lambda E\Big[\Big(\frac{(X+1)P(X+1)}{\lambda P(X)} - 1\Big)^2\Big]$$

and note that $J(X) \geq 0$ with equality iff $X \sim$ Poisson

A New Bound in Terms of Relative Entropy

Theorem 4: Poisson Approximation via Fisher Information [KHJ 05]

If the X_i are independent Bern(p_i) with $E(S_n) = \sum_{i=1}^n p_i = \lambda$, then

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \;\leq\; \sum_{i=1}^n \frac{p_i^3}{\lambda(1-p_i)}$$

Note. This bound is of order $\approx \sum p_i^3$ compared to the earlier $\sum p_i^2$

A New Bound in Terms of Relative Entropy

Theorem 4: Poisson Approximation via Fisher Information [KHJ 05]

If the X_i are independent Bern(p_i) with $E(S_n) = \sum_{i=1}^n p_i = \lambda$, then

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \;\leq\; \sum_{i=1}^n \frac{p_i^3}{\lambda(1-p_i)}$$

Note. This bound is of order $\approx \sum p_i^3$ compared to the earlier $\sum p_i^2$

Proof.

Three steps:

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \quad \stackrel{(a)}{\leq} \quad J(S_n)$$

(a) follows from an application of a recent log-Sobolev inequality due to Bobkov and Ledoux (more later)

A New Bound in Terms of Relative Entorpy

Theorem 4: Poisson Approximation via Fisher Information [KHJ 05]

If the X_i are independent Bern(p_i) with $E(S_n) = \sum_{i=1}^{n} p_i = \lambda$, then

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \;\leq\; \sum_{i=1}^{n} \frac{p_i^3}{\lambda(1-p_i)}$$

Note. This bound is of order $\approx \sum p_i^3$ compared to the earlier $\sum p_i^2$

Proof.

Three steps:

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \quad \overset{(a)}{\leq} \quad J(S_n) \quad \overset{(b)}{\leq} \quad \sum_{i=1}^{n} \frac{p_i}{\lambda} J(X_i)$$

(a) follows from an application of a recent log-Sobolev inequality due to Bobkov and Ledoux (more later)

A New Bound in Terms of Relative Entropy

Theorem 4: Poisson Approximation via Fisher Information [KHJ 05]

If the X_i are independent Bern(p_i) with $E(S_n) = \sum_{i=1}^n p_i = \lambda$, then

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \;\leq\; \sum_{i=1}^n \frac{p_i^3}{\lambda(1-p_i)}$$

Note. This bound is of order $\approx \sum p_i^3$ compared to the earlier $\sum p_i^2$

Proof.

Three steps:

$$D\Big(P_{S_n}\Big\|\text{Po}(\lambda)\Big) \;\overset{(a)}{\leq}\; J(S_n) \;\overset{(b)}{\leq}\; \sum_{i=1}^n \frac{p_i}{\lambda} J(X_i) \;\overset{(c)}{\leq}\; \sum_{i=1}^n \frac{p_i^3}{\lambda(1-p_i)}$$

(a) follows from an application of a recent log-Sobolev inequality due to Bobkov and Ledoux (more later)

(c) is a simple evaluation of $J(\text{Bern}(p))$

Subadditivity of Fisher Information

Proof cont'd.

$$D\Big(P_{S_n}\Big\|\mathrm{Po}(\lambda)\Big) \overset{(a)}{\leq} J(S_n) \overset{(b)}{\leq} \sum_{i=1}^{n} \frac{p_i}{\lambda} J(X_i) \overset{(c)}{\leq} \sum_{i=1}^{n} \frac{p_i^3}{\lambda(1-p_i)}$$

(b) is based on the more general subadditivity property

$$J(S_n) \leq \sum_{i=1}^{n} \frac{E(X_i)}{E(S_n)} J(X_i) \qquad (*)$$

Recall

$$J(X) = \lambda E\Big[\Big(\frac{(X+1)P(X+1)}{\lambda P(X)} - 1\Big)^2\Big]$$

$(*)$ is proved by writing $\Big[\frac{(z+1)P*Q(z+1)}{P*Q(z)} - 1\Big]$ as a conditional expectation and using ideas about L^2 projections of convolutions

Ineq $(*)$ is the natural discrete analog of Stam's Fisher information ineq (in the continuous case), used to prove the *entropy power inequality* □

Example Revisited: Independent Bernoullis

Recall the earlier example

Suppose $X_1, \ldots, X_n$ are i.i.d. Bern($\frac{1}{\sqrt{n}}$) and let $\lambda = \sqrt{n}$

Our earlier bound was

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \leq 1$$

Stein's method gives

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \leq \frac{1}{\sqrt{n}}$$

Theorem 4 combined with Pinsker's ineq gives

$$\|P_{S_n} - \mathsf{Po}(\lambda)\|_{TV} \;\leq\; \sqrt{2}\Big[D(P_{S_n}\|\mathsf{Po}(\lambda))\Big]^{1/2} \;\leq\; \frac{1}{\sqrt{n}}\sqrt{\frac{5}{2}}$$

Moreover, Theorem 4 gives a strong **new** bound in terms of relative entropy!

Outline

1. Poisson Approximation in Relative Entropy

Motivation: Entropy and the central limit theorem
Motivation: Poisson as a maximum entropy distribution
A very simple general bound; **Examples**

2. Analogous Bounds in Total Variation

Suboptimal Poisson approximation
Optimal Compound Poisson approximation

3. Tighter Poisson Bounds for Independent Summands

A (new) discrete Fisher information; subadditivity
A log-Sobolev inequality

4. Measure Concentration and Compound Poisson Tails

The compound Poisson distributions
A log-Sobolev inequality and its info-theoretic proof
Compound Poisson concentration

Motivation: The Concentration Phenomenon

An Example [Bobkov & Ledoux (1998)]

If $W \sim \text{Po}(\lambda)$ and $f(i)$ is 1-Lipschitz, i.e., $|f(i+1) - f(i)| \leq 1$

$$\Pr\Big\{ f(W) - E[f(W)] > t \Big\} \leq \exp\Big\{ -\frac{t}{4} \log\Big(1 + \frac{t}{2\lambda}\Big)\Big\}$$

for all $t > 0$

Note

- ⇝ Sharp bound, valid for all t and *all* such f
- ⇝ One example from a very large class of such results
- ⇝ Many different methods of proof
  dominant one probably the “entropy method”

Proof by the Entropy Method: First Step

Define

The **relative entropy** of a function $g > 0$ w.r.t. a prob distr P

$$\mathrm{Ent}_P(g) = \sum_i P(i)g(i)\log g(i) - \Big[\sum_i P(i)g(i)\Big]\log\Big[\sum_i P(i)g(i)\Big]$$

e.g., if $g(i) = Q(i)/P(i)$, then $\mathrm{Ent}_P(g) = D(Q\|P) =$ relative entropy

A Logarithmic Sobolev Inequality

Our earlier log-Sobolev ineq $D(P\|\mathrm{Po}(\lambda)) \leq \lambda E\Big[\Big(\frac{(X+1)P(X+1)}{\lambda P(X)} - 1\Big)^2\Big]$

is equivalent to: If $W \sim \mathrm{Po}(\lambda)$, then for any function $g > 0$:

$$\mathrm{Ent}_{\mathrm{Po}(\lambda)}(g) \leq \lambda E\Big[\frac{|Dg(W)|^2}{g(W)}\Big]$$

where $Dg(i) = g(i+1) - g(i)$

Proof: Information-theoretic tools

Use the **tensorization property** of relative entropy – more later...

Proof Second Step: The Herbst Argument

Given f, substitute $g(i) = e^{\theta f(i)}$ in the log-Sobolev ineq

$$\mathrm{Ent}_{\mathrm{Po}(\lambda)}(g) \leq \lambda E\Big[\frac{|Dg(W)|^2}{g(W)}\Big]$$

This yields a bound on the log-moment generating fn of $f(W)$

$$L(\theta) = E\Big[e^{\theta f(W)}\Big], \quad W \sim \mathsf{Po}(\lambda)$$

and combining with Chernoff's bound,

$$\begin{aligned} \Pr\Big\{f(W) - E[f(W)] > t\Big\} &\leq L(\theta) \exp\Big\{-\theta\big(t + E[f(W)]\big)\Big\} \\ &\leq \exp\Big\{-\frac{t}{4}\log\Big(1 + \frac{t}{2\lambda}\Big)\Big\} \end{aligned}$$

Remarks

Note

- $\rightsquigarrow$ General, powerful inequality, proved by info-theoretic techniques
- $\rightsquigarrow$ Proof heavily dependent on existence of log-moment generating fn
- $\rightsquigarrow$ Domain of application restricted to a small family (Poisson distr)

Generalize to Compound Poisson Distrs on $\mathbb{Z}_+$

- $\rightsquigarrow$ The asymptotic tails of $Z \sim \text{CP}(\lambda, Q)$ are determined by those of Q

  e.g., if $Q(i) \sim e^{-\alpha i}$ then $\mathrm{CP}_{\lambda,Q}(i) \sim e^{-\alpha i}$

  if $Q(i) \sim 1/i^\beta$ then $\mathrm{CP}_{\lambda,Q}(i) \sim 1/i^\beta$, etc

  Versatility of tail behavior is attractive for modelling

  Concentration? If Q has *sub*-exponential tails the Herbst argument fails

- $\rightsquigarrow$ The $\text{CP}(\lambda, Q)$ distribution can be built up from "small Poissons"

$$\text{CP}(\boldsymbol{\lambda}, \boldsymbol{Q}) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\text{Po}(\lambda)} \boldsymbol{X}_i \stackrel{\mathcal{D}}{=} \sum_{j=1}^{\infty} \boldsymbol{j} \cdot \text{Po}(\boldsymbol{\lambda Q}(\boldsymbol{j}))$$

A Compound Poisson Log-Sobolev Inequality

Theorem 5: Log-Sobolev Inequality for CP Distrs [Wu 00, K-Madiman 05]

Let $X \sim P$ be an arbitrary RV with values in $\mathbb{Z}_+$

For any $\lambda > 0$, any distr Q on the natural nos, any $g > 0$

$$\mathrm{Ent}_{\mathrm{CP}(\lambda,Q)}(g) \leq \lambda \sum_{j \geq 1} Q(j) E\Big[\frac{|D^j g(Z)|^2}{g(Z)}\Big]$$

where $Z \sim \mathrm{CP}(\lambda, Q)$ and $D^j g(i) = g(i+j) - g(i)$

Proof Idea

Use the **tensorization property** of the relative entropy

$$\mathrm{Ent}_{\mathrm{Po}(\lambda_1)\times\mathrm{Po}(\lambda_2)\times\cdots\times\mathrm{Po}(\lambda_n)}(g) \leq \sum_{j=1}^{n} E\Big[\mathrm{Ent}_{\mathrm{Po}(\lambda_j)}\big(g(W_1^{j-1}, \cdot, W_{j+1}^n)\big)\Big]$$

to get a vector version of the Poisson LSI

Apply it to $g(w_1, w_2, \ldots, w_n) = \sum_j j \cdot w_j$ and let $n \to \infty$, using

$$\mathrm{CP}(\lambda, Q) \;=\; \lim_n \sum_{j=1}^{n} j \cdot \mathrm{Po}(\lambda Q(j))$$

New Measure Concentration Bounds

Theorem 6: Measure Concentration for CP Distributions [K-Madiman 05]

(i) Suppose $Z \sim \text{CP}(\lambda, Q)$ and Q has finite Kth moment

$$\sum_j j^K Q(j) < \infty$$

If f is 1-Lipschitz, i.e., $|f(i+1) - f(i)| \leq 1$ for all i
then for $t > 0$

$$\Pr\Big\{\big|f(Z) - E[f(Z)]\big| > t\} \leq A\Big(\frac{B}{t}\Big)^K$$

where the constants A, B are explicit and depend only
on $\lambda, K, |f(0)|$, and on the integer moments of Q

(ii) An analogous bound holds for any RV Z whose distr satisfies
the log-Sobolev ineq of Thm 5

The Constants in Theorem 2

Let

$$q(r) = \sum_j j^r\, Q(j)$$

Then

$$\Pr\Big\{\big|f(Z) - E[f(Z)]\big| > t\Big\} \le A\Big(\frac{B}{t}\Big)^K$$

where

$$\begin{aligned} A &= \exp\Big\{\sum_{r=1}^{K}\binom{K}{r} q(r)\Big\} \\ B &= 2|f(0)| + 2\lambda q(1) + 1 \end{aligned}$$

Proof Outline

Modification of Herbst argument: Given f, let $G_\theta(i) = |f(i) - E[f(Z)]|^\theta$ and define the "polynomial" moment-generating fn

$$M(\theta) = E[G_\theta(Z)]$$

Substitute $g = G_\theta$ in the log-Sobolev ineq

$$\mathrm{Ent}_{\mathrm{CP}(\lambda,Q)}(g) \leq \lambda \sum_{j\geq 1} Q(j) E\Big[\frac{|D^j g(Z)|^2}{g(Z)}\Big]$$

to get the differential inequality

$$\theta M'(\theta) - M(\theta)\log M(\theta) \leq \lambda M(\theta) \sum_j Q(j) \Big[\text{terms involving } \theta \log(C + Dj)\Big]$$

Solving, yields a bound on $M(\theta)$, and combining with Markov's ineq,

$$\Pr\Big\{\big|f(Z) - E[f(Z)]\big| > t\Big\} \;\leq\; \frac{M(\theta)}{t^\theta} \;\leq\; \cdots \;\leq\; A\Big(\frac{B}{t}\Big)^K$$

$\square$

Final Remarks

Information-theoretic approach to (Compound-)Poisson approximation

Two approaches

$\rightsquigarrow$ A simple, very general one

$\rightsquigarrow$ A tight one for the independent Poisson case

Non-asymptotic, strong *new* bounds, intuitively satisfying

Ideas

A new version of Fisher information

L^2-theory and log-Sobolev inequalities for discrete random variables

Concentration

A simple, general CP-approximation bound

A log-Sobolev ineq for the CP dist

New non-exponential measure concentration bounds

Information-Theoretic Interpretation

$$D\Big(P_{\hat{S}_n}\Big\|N(0,\sigma^2)\Big) \downarrow 0 \iff h(\hat{S}_n) \uparrow h(N(0,\sigma^2)) \text{ as } n \to \infty$$

(i) The accumulation of many, small, independent random effects is maximally random

(ii) The monotonicity in n indicates that the entropy is a natural measure for the convergence of the CLT

More generally the CLT holds as long as

(a) Each $E(X_i)$ is small

(b) The overall variance $\mathrm{Var}(\hat{S}_n) \approx \sigma^2$

(c) The X_i are weakly dependent

$\rightsquigarrow$ Next look at the other central result on the distribution of the sum of many small random effects: Poisson approximation

Two Examples

The defining compound Poisson example

If $X_1, X_2, \ldots, X_n$ are IID $\sim Q$ on $\mathbb{N}$ and $I_1, I_2, \ldots, I_n$ are IID Bern(λ/n) then for $S_n = \sum_{i=1}^n I_i X_i$ Theorem 3 gives

$$D(P_{S_n} \| \mathsf{CP}(\lambda, Q)) \leq \sum_{i=1}^n (\lambda/n)^2 = \lambda^2/n$$

Again, sufficient for convergence, but the optimal rate is $O(1/n^2)$

A Markov chain example

Let $S_n = \sum_{i=1}^n I_i X_i$ where $X_1, \ldots, X_n$ are IID $\sim Q$ on $\mathbb{N}$ and $I_1, \ldots, I_n$ is a stationary Markov chain with transition matrix

$$\begin{pmatrix} \frac{n}{n+1} & \frac{1}{n+1} \\ \frac{n-1}{n+1} & \frac{2}{n+1} \end{pmatrix}$$

Theorem 3 easily gives $D(P_{S_n} \| \mathsf{CP}(1, Q)) \leq \dfrac{3 \log n}{n} + \dfrac{1}{n}$

Another Example

Theorem 2 easily generalizes to non-binary X_i, as long as $J(X_i)$ can be evaluated or estimated. E.g.:

Sum of Small Geometrics

Suppose $X_1, X_2, \ldots, X_n$ are indep Geom(q_i)
let $\lambda = E(S_n) = \sum_{i=1}^n [(1-q_i)/q_i]$

Then $J(X_i) = (1-q_i)^2/q_i$ and proceeding as in the proof of Theorem 2

$$D(P_{S_n} \| \text{Po}(\lambda)) \;\leq\; \sum_{i=1}^n \frac{(1-q_i)^3}{\lambda q_i^2}$$

In the case when all $q_i = n/(n+\lambda) \approx 1 - \lambda/n$ this takes the elegant form

$$D(P_{S_n} \| \text{Po}(\lambda)) \;\leq\; \frac{\lambda^2}{n^2}$$

Tighter Bounds Compound Poisson Approximation?

Recall the proof of Theorem 2 in the Poisson case:

$$D\Big(P_{S_n}\Big\|\mathsf{Po}(\lambda)\Big) \quad \overset{(a)}{\leq} \quad J(S_n) \quad \overset{(b)}{\leq} \quad \sum_{i=1}^{n} \frac{p_i}{\lambda} J(X_i) \quad \overset{(c)}{\leq} \quad \sum_{i=1}^{n} \frac{p_i^3}{\lambda(1-p_i)}$$

$\leadsto$ In order to generalize this approach we first need a new version of the Fisher information, and a corresponding log-Sobolev ineq for the compound Poisson measure . . .

Properties of the Compound Poisson Distribution

- The $\mathrm{CP}(\lambda, Q)$ laws are the *only* infinitely divisible distr's on $\mathbb{Z}_+$
- The asymptotic tails of $Z \sim \mathrm{CP}(\lambda, Q)$ are determined by those of Q

  e.g., if $Q(i) \sim e^{-\alpha i}$ then $\mathrm{CP}_{\lambda,Q}(i) \sim e^{-\alpha i}$

  if $Q(i) \sim 1/i^{\beta}$ then $\mathrm{CP}_{\lambda,Q}(i) \sim 1/i^{\beta}$, etc

  Versatility of tail behavior is attractive for modelling

  Concentration? If Q has *sub*-exponential tails the Herbst argument fails

- The $\mathrm{CP}(\lambda, Q)$ distribution can be built up from "small Poissons"

$$\mathrm{CP}(\boldsymbol{\lambda}, \boldsymbol{Q}) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\mathrm{Po}(\boldsymbol{\lambda})} \boldsymbol{X}_i \stackrel{\mathcal{D}}{=} \sum_{j=1}^{\infty} \boldsymbol{j} \cdot \mathrm{Po}(\boldsymbol{\lambda Q}(\boldsymbol{j}))$$

A New Log-Sobolev Inequality

Let $C_{\lambda,Q}(k)$ denote the compound Poisson probabilities $\Pr\{\mathsf{CP}(\lambda,Q)=k\}$

Theorem 4: Log-Sobolev Inequality for the Compound Poisson Measure

Let $X \sim P$ be an arbitrary $\mathbb{Z}_+$-valued RV

(a) [Bobkov-Ledoux (1998)] For any $\lambda > 0$:

$$D\Big(P\Big\|\mathsf{Po}(\lambda)\Big) \;\le\; \lambda E\Big[\Big(\frac{(X+1)}{\lambda}\frac{P(X+1)}{P(X)}-1\Big)^2\Big]$$

(b) For any $\lambda > 0$ and any measure Q on $\mathbb{N}$:

$$D\Big(P\Big\|\mathsf{CP}(\lambda,Q)\Big) \;\le\; \lambda\sum_{j=1}^{\infty} Q(j)E\Big[\Big(\frac{C_{\lambda,Q}(X)}{C_{\lambda,Q}(X+j)}\frac{P(X+j)}{P(X)}-1\Big)^2\Big]$$

Proof of Theorem 4 (a)

Step 1. Derive a simple log-Sobolev ineq for the Bernoulli measure $B_p(k)$
For any binary RV $X \sim P$:

$$D\Big(P\Big\|\mathrm{Bern}(p)\Big) \;\le\; p(1-p)E\Big[\Big(\frac{B_p(X)}{B_p(X+1)}\frac{P(X+1)}{P(X)}-1\Big)^2\Big]$$

Step 2. Recall the "tensorization" property of relative entropy
Whenever $X=(X_1,\ldots,X_n)\sim P_n$:

$$D\Big(P_n\Big\|\prod_{i=1}^n \nu_i\Big) \;\le\; \sum_{i=1}^n E_{P_n}\Big[D\Big(P_n(\cdot|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n)\Big\|\nu_i\Big)\Big]$$

Use this to extend step 1 to products of Bernoullis:

$$D\Big(P_n\Big\|\prod_{i=1}^n \mathrm{Bern}(p)\Big) \;\le\; p(1-p)E\Big[\sum_{i=1}^n\Big(\frac{B_p^n(X)}{B_p^n(X+e_i)}\frac{P_n(X+e_i)}{P_n(X)}-1\Big)^2\Big]$$

Step 3. Since Po(λ) $\stackrel{\mathcal{D}}{=} \lim_n \sum_{i=1}^n$ Bern(λ/n), applying step 2 to a P_n
that only depends on $X_1+\cdots+X_n$ and taking $n\to\infty$:

$$D\Big(P\Big\|\mathrm{Po}(\lambda)\Big) \;\le\; \lambda E\Big[\Big(\frac{(X+1)}{\lambda}\frac{P(X+1)}{P(X)}-1\Big)^2\Big]$$

Proof of Theorem 4 (b)

In (a), the key was the representation of $\mathrm{Po}(\lambda)$ in terms of indep Bernoullis

$$\mathrm{Po}(\lambda) \stackrel{\mathcal{D}}{=} \lim_n \sum_{i=1}^n \mathrm{Bern}(\lambda/n)$$

Here use an alternative representation of $\mathrm{CP}(\lambda, Q)$ in terms of indep Poissons

$$\mathrm{CP}(\lambda, Q) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\mathrm{Po}(\lambda)} X_i \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\infty} j \cdot \mathrm{Po}(\lambda Q(j)) \stackrel{\mathcal{D}}{=} \lim_n \sum_{i=1}^{n} j \cdot \mathrm{Po}(\lambda Q(j)) \qquad (*)$$

Step 1. Start with the Poisson log-Sobolev ineq of (a)

Step 2. Tensorize to obtain an ineq for products of Poissons
Whenever $X = (X_1, \ldots, X_n) \sim P_n$:

$$D\Big(P_n \Big\| \prod_{i=1}^n \mathrm{Po}(\lambda_i)\Big) \le \Big[\cdots\Big]$$

Step 3. Apply step 2 to a P_n that only depends on $\sum_{j=1}^n j \cdot X_j$
and take $n \to \infty$ using $(*)$ □

Measure Concentration Bounds

Instead of continuing with CP-approximation, take a detour

$\rightsquigarrow$ Suppose, for simplicity, that Q has finite support $\{1, 2, \ldots, m\}$

$\rightsquigarrow$ Write as before $C_{\lambda,Q}(k) = \Pr\{\text{CP}(\lambda, Q) = k\}$

Theorem 5: Measure Concentration for CP-like Measures

(i) Let $Z \sim \text{CP}(\lambda, Q)$ and f be a Lipschitz-1 function on $\mathbb{Z}_+$ $\big[|f(k+1) - f(k)| \leq 1 \text{ for all } k\big]$. For $t > 0$:

$$\Pr\{f(Z) \geq E[f(Z)] + t\} \leq \exp\Big\{-\frac{t}{2m}\log(1 + \frac{t}{\lambda m^2})\Big\}$$

(ii) An analogous bound holds for any $Z \sim \mu$ that satisfies the log-Sobolev ineq of Thm 4

Remarks

Proof. Follows Herbst's Gaussian argument: Apply the log-Sobolev ineq to $f = e^{\theta g}$ for a Lipschitz g. Expand to get a differential inequality for the M.G.F. $L(\theta) = E[e^{\theta g(Z)}]$. Use the bound and apply Chebychev

The finite-support assumption. Can be relaxed at the price of technicalities. More general bounds, much more general class of tails

Poisson tails. From Theorem 5 we see that Lipschitz-1 functions of CP-like RVs have Poisson tails. In particular:

Corollary: Poisson Tails for Lipschitz Functions

Let $Z \sim \text{CP}(\lambda, Q)$ or any other distr satisfying the assumptions of Thm 5
For any Lipschitz-1 function f on $\mathbb{Z}_+$ we have:

$$E\left[e^{\theta |f(Z)| \log^+ |f(Z)|}\right] < \infty \qquad \text{for all } \theta > 0 \text{ small enough}$$