Probabilistic Method and Random Graphs

Lecture 5. Bins&Balls: Law of Small Numbers, Poisson Approximation [1]

Xingwu Liu

Institute of Computing Technology
Chinese Academy of Sciences, Beijing, China

[1] The slides are mainly based on Chapter 5 of *Probability and Computing*.

Questions, comments, or suggestions?

Review: Large Deviation Theory

Central limit theorem: $O(\sqrt{n})$ deviation, no rate information

Chernoff bounds: large deviation, but loose

Large deviation theorem: asymptotical, tight vanishing rate

By courtesy of Cramer (1944).
Let $X_1, \ldots X_n, \ldots \in \mathbb{R}$ be **i.i.d.** r.v. which satisfy $\mathbb{E}[e^{tX_1}] < \infty$ for $t \in \mathbb{R}$. Then for any $t > \mathbb{E}[X_1]$, we have

$$\lim_{n \to \infty} \frac{1}{n} \ln \Pr\left(\sum_{i=1}^{n} X_i \geq tn\right) = -\sup_{\lambda > 0}(\lambda t - \ln \mathbb{E}[e^{\lambda X_1}]).$$

Review: bins-and-balls

General model: m balls independently randomly placed in n bins

Distribution of the load X of a bin: $Bin(m, 1/n)$

When $m, n \gg r$, $\Pr(X = r) \approx e^{-\mu}\frac{\mu^r}{r!}$ with $\mu = \frac{m}{n}$.

Poisson distribution

Poisson distribution: $\Pr(X_\mu = r) = e^{-\mu}\frac{\mu^r}{r!}$.

Law of rare events

Rooted at Law of Small Numbers

Review: Basic Properties of Poisson distribution

Low-order moments

$\mathbb{E}[X_\mu] = Var[X_\mu] = \mu.$

Moment generation function

$M_{X_\mu}(t) = \mathbb{E}[e^{tX_\mu}] = \sum_k \frac{e^{-\mu}\mu^k}{k!} e^{tk} = e^{\mu(e^t - 1)}.$

Additive

By uniqueness of moment generation functions,
$X_{\mu_1} + X_{\mu_2} = X_{\mu_1 + \mu_2}$ if independent.

Chernoff-like bounds

1. If $x > \mu$, then $\Pr(X_\mu \geq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$.
2. If $x < \mu$, then $\Pr(X_\mu \leq x) \leq \frac{e^{-\mu}(e\mu)^x}{x^x}$.

Review: Joint Distribution of Bin Loads

Basic observation

Loads of multiple bins are not independent.
Hard to handle

Maximum load

- $\Pr(L \geq 2) \geq 0.5$ if $m \geq \sqrt{2n \ln 2}$
  - Birthday paradox
- $\Pr(L \geq 3\frac{\ln n}{\ln \ln n}) \leq \frac{1}{n}$ if $m = n$

Is there a closed form of $\Pr(X_1 = k_1, ...X_n = k_n)$?

Law of Small Numbers (Poisson Convergence)

Poisson convergence of binomial distribution

Assume that $X_n \sim Bin(n, p_n)$ with $\lim_{n\to\infty} np_n = \lambda$. For any fixed k, $\lim_{n\to\infty} \Pr(X_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}$.

It is intuitively acceptable (by their figures)

It can be used to approximately calculate Binomial distribution $Bin(n, p)$, but take care.
$n > 100, p < 0.01, np < 20.$

Error bounds implies the convergence

$$e^{\frac{p(k-np)}{1-p} - \frac{k(k-1)}{2(n-k+1)}} \le \frac{\Pr(Bin(n,p)=k)}{\Pr(Poi(np)=k)} \le e^{kp - \frac{k(k-1)}{2n}}.$$

Proof of the error bounds

Error bounds

$$e^{\frac{p(k-np)}{1-p}-\frac{k(k-1)}{2(n-k+1)}} \leq \frac{\Pr(Bin(n,p)=k)}{\Pr(Poi(np)=k)} \leq e^{kp-\frac{k(k-1)}{2n}}.$$

Proof

$A_{n,p,k} \triangleq \frac{\Pr(Bin(n,p)=k)}{\Pr(Poi(np)=k)} = \prod_{j=1}^{k-1}\left(1-\frac{j}{n}\right)e^{np}(1-p)^{n-k}$ for $0 \leq k \leq n$ and it's 0 otherwise.

Upper bound

$$A_{n,p,k} \leq e^{-\sum_{j=1}^{k-1}\frac{j}{n}+np-(n-k)p} \leq e^{kp-\frac{k(k-1)}{2n}}.$$

Lower bound

$$\begin{aligned} A_{n,p,k} &\geq e^{-\sum_{j=1}^{k-1}\frac{j/n}{1-j/n}+np-(n-k)\frac{p}{1-p}} \\ &= e^{-\sum_{j=1}^{k-1}\frac{j}{n-j}-\frac{p(np-k)}{1-p}} \geq e^{\frac{p(k-np)}{1-p}-\frac{k(k-1)}{2(n-k+1)}}. \end{aligned}$$

Generalize LSN to weak dependence

Poisson convergence with weak dependence

For each n, Bernoulli experiments $B_1^n, \ldots B_n^n$ with indicators X_i^n, if

- $\lim_{n\to\infty} \mathbb{E}[Y_n] = \lambda$ for $Y_n = \sum_{i=1}^n X_i^n$
- For any k, $\lim_{n\to\infty} \sum_{1\le i_1<\ldots<i_k\le n} \Pr(\bigcap_{r=1}^k B_{i_r}^n) = \frac{\lambda^k}{k!}$

Then $Y_n \to Poi(\lambda)$, i.e. $\Pr(Y_n = j) \to \frac{e^{-\lambda}\lambda^j}{j!}$ for any $j \ge 0$

Basic idea of the proof for $j = 0$:
Use Taylor series of $e^{-\lambda}$ and Bonferroni inequalities

- $\Pr(\bigcup_{i\ge 1}^{n} B_i^n) \le \sum_{l=1}^{r} (-1)^{l-1} \sum_{i_1<i_2<\ldots<i_l} \Pr(\bigcap_{r=1}^{l} B_{i_r}^n)$ for odd r
- $\Pr(\bigcup_{i\ge 1}^{n} B_i^n) \ge \sum_{l=1}^{r} (-1)^{l-1} \sum_{i_1<i_2<\ldots<i_l} \Pr(\bigcap_{r=1}^{l} B_{i_r}^n)$ for even r

Remarks on the case of weak dependence

Intuitive explanation

If X is the number of a large collection of nearly independent events that rarely occur, the $X \sim Poi(\mathbb{E}[X])$

Application

- The number of people who get their own hats back after a random permutation of the hats
- The number of pairs having the same birthday
- The number of isolated vertices in random graph $G(n, \frac{\ln n + c}{n})$

It can be further generalized

Generalize LSN to strong dependence

Poisson convergence with strong dependence, 1975

Stein-Chen Theorem: If $Y_n = \sum_{i=1}^n X_i, X_i \sim Ber(p_i)$ and $\lambda = \sum_{i=1}^n p_i$, then for any $A \subseteq \mathbb{Z}_+$,

$$|\Pr(Y_n \in A) - \Pr(Poi(\lambda) \in A)| \leq \min\left\{1, \frac{1}{\lambda}\right\} \sum_{i=1}^n p_i \mathbb{E}[|U_i - V_i|].$$

where $U_i \sim Y_n, 1 + V_i \sim Y_n | X_i = 1$.

Intuitive explanation

Poisson approximation remains valid even if the Bernoulli r.v.s are strongly dependent and have different expectations.

Remarks on the law of small numbers

Law of small numbers vs Law of large numbers (CLT)

- Poisson approximation vs Normal approximation
- Small number vs arbitrary number
- Summation on different sets vs summation on a single sequence

Relation between Poisson and Normal distribution

Should be related since both approximate binomial distribution.
When $\lambda \to \infty$, Poisson converges to Normal.
Specifically, $\lim_{\lambda\to\infty} \sum_{\alpha<k<\beta} \frac{\lambda^k e^{-\lambda}}{k!} = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-\frac{x^2}{2}} dx$.
Where $a = (\alpha - \lambda)/\sqrt{\lambda}, b = (\beta - \lambda)/\sqrt{\lambda}$ are fixed.

Intuitive argument

Uniqueness+continuity of moment generating functions.

Joint Distribution of Bin Loads

Theorem

$\Pr(X_1 = k_1, ...X_n = k_n) = \frac{m!}{k_1!k_2!\cdots k_n!n^m}$

Proof.

By the chain rule, $\Pr(X_1 = k_1, ...X_n = k_n)$

$= \prod_{i=0}^{n-1} \Pr(X_{i+1} = k_{i+1} | X_1 = k_1, ...X_i = k_i)$.

Note that $X_{i+1} | (X_1 = k_1, ...X_i = k_i)$ is a binomial r.v. of $m - (k_1 + \cdots + k_i)$ trials with success probability $\frac{1}{n-i}$.

$\square$

Remark

- You can also prove by counting
- Multinomial coefficient $\frac{m!}{k_1!k_2!\cdots k_n!}$: the number of ways to allocate m distinct balls into groups of sizes $k_1, \cdots, k_n$

Silver bullet for Bins&Balls problems?

In principle

Yes, since it can be computed

In practice

Usually No, since too hard to compute.
Example: what's the probability of having empty bins?

In need

Approximation for computing or insights for analysis

Poisson Approximation

At the first glance

The (marginal) load $X_i \sim Bin(m, \frac{1}{n})$ for each bin i .
$\{X_1, ...X_n\}$ are not independent.
But seemingly the only dependence is that their sum is m. So,

A applausible conjecture

The joint distribution $(X_1, ...X_n) \sim (Y_1, ...Y_n | \sum Y_i = m)$, where $Y_i \sim Bin(m, \frac{1}{n})$ are mutually independent.

If this is true, good simplification is obtained.

However

It is NOT the case!

Why is it not true?

General Fact

Given joint distribution $\mathcal{J}$ with marginal distribution $\mathcal{M}_1, ..., \mathcal{M}_n$ independent except $\mathcal{M}_1 + ... + \mathcal{M}_n = m$, then the marginals of $(\mathcal{M}_1, ..., \mathcal{M}_n | \mathcal{M}_1 + ... + \mathcal{M}_n = m)$ are not $\mathcal{M}_1, ..., \mathcal{M}_n$, i.e. $(\mathcal{M}_1, ..., \mathcal{M}_n | \mathcal{M}_1 + ... + \mathcal{M}_n = m) \nsim \mathcal{J}$

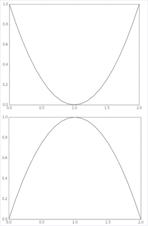

Figure: f_X and f_Y

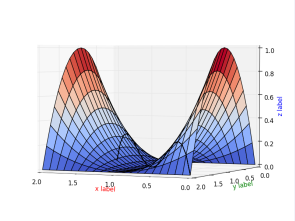

Figure: The joint distribution $f_X * f_Y$ conditioned on $X + Y = 1$ (the sick line)

But is the conjecture true for any distribution other than binomial?

Yes!

Poisson distribution again. (Better than the conjecture)

Poisson Approximation Theorem

Notation

$X_i^{(m)}$: the load of bin i in (m,n)-model, $1 \le i \le n$.
$Y_i^{(\mu)}$: independent Poisson r.v.s with expectation μ, $1 \le i \le n$.

Theorem

$(X_1^{(m)}, X_2^{(m)}, ...X_n^{(m)}) \sim (Y_1^{(\mu)}, Y_2^{(\mu)}, ...Y_n^{(\mu)} | \sum Y_i^{(\mu)} = m)$.

Remarks

- The equation is independent of μ: For any m, the same Poisson distribution works.
- Since $\Pr(X_1^{(m)}, X_2^{(m)}, ...X_n^{(m)}) \propto \Pr(Y_1^{(\mu)}, Y_2^{(\mu)}, ...Y_n^{(\mu)})$, the X_i's are **decoupled**.
- The two distributions are exactly equal, not approximate.

Proof

By straightforward calculation.

Example

Coupon Collector Problem

Let X be the number of purchases by n types are collected. Then for any constant c, $\lim_{n\to\infty} \Pr(X > n \ln n + cn) = 1 - e^{-e^{-c}}$.

Remark: $\Pr(n \ln n - 4n \le X \le n \ln n + 4n) \ge 0.98$

Basic idea of the proof

Use bins-and-balls model and the Poisson approximation.
It holds under the Poisson approximation.
The approximation is actually accurate.

Proof

Modeling

$X > n\ln n + cn$ means that there are empty bins in the $(n\ln n + cn, n)$-Bins&Balls model.

It holds under the Poisson approximation

Approximation experiment: n bins, each having a Poisson number of balls with the expectation $\ln n + c$.
Event $\mathcal{E}$: No bin is empty.
$\Pr(\mathcal{E}) = \left(1 - e^{-(\ln n + c)}\right)^n = \left(1 - \frac{e^{-c}}{n}\right)^n \to e^{-e^{-c}}$.

The approximation is accurate

Obj.: Asymptotically, $\Pr(\mathcal{E}) = \Pr(\mathcal{E}') = \Pr(\mathcal{E} | X = n\ln n + cn)$, where X is the totally number of balls in the approximation experiment while $\mathcal{E}'$ means no bin is empty in the $(n\ln n + cn, n)$-Bins&Balls model.

Proof: $\Pr(\mathcal{E}) = \Pr(\mathcal{E} | X = n \ln n + cn)$

Further reduction

Since $\Pr(\mathcal{E}) = \Pr(\mathcal{E} | X \in \mathbb{Z})$, there should be a neighborhood $\mathcal{N} \subset \mathbb{Z}$ s.t. $n \ln n + cn \in \mathcal{N}$ and $\Pr(\mathcal{E}) \approx \Pr(\mathcal{E} | X \in \mathcal{N})$.
If $\mathcal{N}$ is not too small or too big, i.e.

- $\Pr(X \in \mathcal{N}) \approx 1$;
- $\Pr(\mathcal{E} | X \in \mathcal{N}) \approx \Pr(\mathcal{E} | X = n \ln n + cn)$.

We finish the proof by total probability formula.

Does such $\mathcal{N}$ exist?

Yes! Try the $\sqrt{2m \ln m}$-neighborhood of $m = n \ln n + cn$.

Proof: $\Pr(|X - m| \le \sqrt{2m \ln m}) \to 1$

$X \sim Poi(m)$.

By Chernoff bound $\Pr(X \ge x) \le \frac{e^{-m}(em)^x}{x^x} = e^{x-m-x\ln\frac{x}{m}}$,

$$\Pr(X > m + \sqrt{2m\ln m}) \le e^{\sqrt{2m\ln m}-(m+\sqrt{2m\ln m})\ln(1+\sqrt{\frac{2\ln m}{m}})}$$

$$\text{by } \ln(1+z) \ge z - z^2/2 \text{ for } z \ge 0$$

$$\le e^{-\ln m + \frac{\ln^{3/2} m}{\sqrt{m}}} \to 0.$$

Likewise, $\Pr(X < m - \sqrt{2m\ln m}) \to 0$.

Proof: $\Pr(\mathcal{E} \mid |X-m| \le \sqrt{2m\ln m}) \approx \Pr(\mathcal{E} \mid X=m)$

$\Pr(\mathcal{E}|X=k)$ increases with k, so

$$\begin{aligned}\Pr(\mathcal{E}|X=m-\sqrt{2m\ln m}) &\le \Pr(\mathcal{E}||X-m| \le \sqrt{2m\ln m}) \\ &\le \Pr(\mathcal{E}|X=m+\sqrt{2m\ln m}).\end{aligned}$$

$$\begin{aligned}&|\Pr(\mathcal{E}||X-m| \le \sqrt{2m\ln m}) - \Pr(\mathcal{E}|X=m)| \\ &\le \Pr(\mathcal{E}|X=m+\sqrt{2m\ln m}) - \Pr(\mathcal{E}|X=m-\sqrt{2m\ln m}).\end{aligned}$$

The last formula means the probability that there is at least one empty bin after throwing $m-\sqrt{2m\ln m}$ balls but at least one among the next $2\sqrt{2m\ln m}$ balls goes into this bin, hence $\le \frac{2\sqrt{2m\ln m}}{n} \to 0$.

References

1. https://www.math.illinois.edu/~psdey/414CourseNotes.pdf
2. http://willperkins.org/6221/slides/poisson.pdf