Probabilistic Method and Random Graphs

Lecture 5. Bins&Balls: Poisson Approximation and Applications [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*.

Preface

Questions, comments, or suggestions?

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

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

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 $m = n$

Let's be ambitious

Is there a closed form of $\Pr(X_1 = k_1, ..., X_n = k_n)$?
Hard? Easy when $n = 2$.

Joint Distribution of Bin Loads

Theorem

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

Proof.

By the chain rule,

$$\begin{aligned} & \Pr(X_1 = k_1, \cdots, X_n = k_n) \\ = & \prod_{i=0}^{n-1} \Pr(X_{i+1} = k_{i+1} | X_1 = k_1, \cdots, X_i = k_i) \end{aligned}$$

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

□

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 .
$\{X_1, \cdots, X_n\}$ are not independent.
But seemingly the only dependence is that their sum is m. So,

A applausible conjecture

The joint distribution $(X_1, \cdots, X_n) \sim (Y_1, \cdots, Y_n | \sum Y_i = m)$, 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!
$(Y_1, \cdots, Y_n | \sum Y_i = m)$ doesn't have marginal distr. as Y_i's.

General Fact

Y_i: mutual independent, $1 \leq i \leq n$.
$(Y_1, ..., Y_n | g(\overrightarrow{Y}))$ doesn't have marginal distr. as Y_i's.

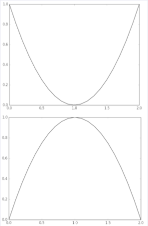

Figure: f_X and f_Y

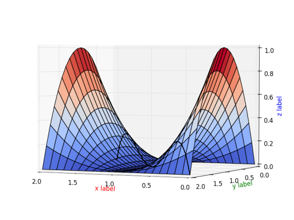

Figure: The joint distribution $f_X * f_Y$ conditioned on $X + Y = 1$ (the sick line)

Recall the false conjecture

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

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

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

Remarks

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

Proof

By straightforward calculation.

Example

Coupon Collector Problem

X: the number of purchases until n types are collected.
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

$X > n \ln n + cn$ is equivalent to event $\overline{\mathcal{E}}$, where $\mathcal{E}$ means that there is no empty bin 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 Y_i 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, $\Pr(\mathcal{E}) = \Pr(\mathcal{E}')$.
By Poisson Approximation, $\Pr(\mathcal{E}) = \Pr(\mathcal{E}' | \sum_{i=1}^n Y_i = n \ln n + cn)$, so we prove $\Pr(\mathcal{E}') = \Pr(\mathcal{E}' | Y = n \ln n + cn)$ with $Y = \sum_{i=1}^n Y_i$.

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

Further reduction

Since $\Pr(\mathcal{E}') = \Pr(\mathcal{E}'|Y \in \mathbb{Z})$, there should be $\mathcal{N} \subset \mathbb{Z}$ s.t $n\ln n + cn \in \mathcal{N}$ and $\Pr(\mathcal{E}') \approx \Pr(\mathcal{E}'|Y \in \mathcal{N})$.

If $\mathcal{N}$ is not too small or too big, i.e.

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

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

$Y \sim Poi(m)$.
By Chernoff bound $\Pr(Y \ge y) \le \frac{e^{-m}(em)^y}{y^y} = e^{y-m-y\ln\frac{y}{m}}$,

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

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

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

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

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

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

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

$$
\begin{aligned}
& |\Pr(\mathcal{E}' \,|\, |Y-m| \le \sqrt{2m\ln m}) - \Pr(\mathcal{E}'|Y=m)| \\
\le\ & \Pr(\mathcal{E}'|Y = m + \sqrt{2m\ln m}) - \Pr(\mathcal{E}'|Y = m - \sqrt{2m\ln m}) \\
=\ & \Pr(A)\text{(By Poisson approximation)}.
\end{aligned}
$$

Event A: In the $(m + \sqrt{2m\ln m})$-Bins&Balls model, the first $m - \sqrt{2m\ln m}$ balls leave a bin empty, but at least one among the next $2\sqrt{2m\ln m}$ balls goes into this bin.

$\Pr(A) \le \frac{2\sqrt{2m\ln m}}{n} \to 0$

Poisson approximation is nice but ...

Hard to use due to conditioning.

Can we remove the condition?

Condition-free Poisson Approximation

Notation

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

Theorem

For any non-negative n-ary function f, we have
$\mathbb{E}\left[f\left(X_1^{(m)}, ...X_n^{(m)}\right)\right] \leq e\sqrt{m}\mathbb{E}\left[f\left(Y_1^{(m)}, ...Y_n^{(m)}\right)\right]$.

Remark

Unlike $\left(X_1^{(m)}, X_2^{(m)}, ...X_n^{(m)}\right) \sim \left(Y_1^{(\mu)}, Y_2^{(\mu)}, ...Y_n^{(\mu)} | Y = m\right)$, the mean of the Poisson distribution is $\frac{m}{n}$, not arbitrary.

Condition-freedom at the cost of approximation.

Proof

$$\begin{aligned}
&\mathbb{E}[f(Y_1^{(m)}, ...Y_n^{(m)})] \\
&\quad = \sum_k \mathbb{E}[f(Y_1^{(m)}, ...Y_n^{(m)}) | \sum_i Y_i^{(m)} = k] \Pr(\sum_i Y_i^{(m)} = k) \\
&\quad \geq \mathbb{E}[f(Y_1^{(m)}, ...Y_n^{(m)}) | \sum_i Y_i^{(m)} = m] \Pr(\sum_i Y_i^{(m)} = m) \\
&\quad = \mathbb{E}[f(X_1^{(m)}, ...X_n^{(m)})] \Pr(\sum_i Y_i^{(m)} = m).
\end{aligned}$$

$\sum_i Y_i^{(m)} \sim Poi(m) \Rightarrow \Pr(\sum_i Y_i^{(m)} = m) = \frac{m^m e^{-m}}{m!} \geq \frac{1}{e\sqrt{m}}$ since $m! < e\sqrt{m}(me^{-1})^m$.

Remark

$\mathbb{E}[f(X_1^{(m)}, ...X_n^{(m)})] \leq 2\mathbb{E}[f(Y_1^{(m)}, ...Y_n^{(m)})]$ if f is monotonic in m

In Terms of Probability

Any event that takes place with probability p in the independent Poisson approximation experiment takes places in Bins&Balls setting with probability at most $pe\sqrt{m}$

If the probability of an event in Bins&Balls is monotonic in m, it is at most twice of that in the independent Poisson approximation experiment

Remark

Powerful in bounding the probability of rare events in Bins&Balls.

Application

Lower bound of max load in (n, n)-model

Asymptotically, $\Pr(\mathcal{E}) \leq \frac{1}{n}$, where $\mathcal{E}$ is the event that the max load in the (n, n)-Bins&Balls model is smaller than $\frac{\ln n}{\ln \ln n}$.

Remark: In fact, the max load is $\Theta\left(\frac{\ln n}{\ln \ln n}\right)$ w.h.p.

Proof

$\mathcal{E}'$: Poisson approx. experiment has max load $\leq M = \frac{\ln n}{\ln \ln n}$.
$\Pr(\mathcal{E}') \leq \left(1 - \frac{1}{eM!}\right)^n \leq e^{-\frac{n}{eM!}}$.

$M! \leq e\sqrt{M}(e^{-1}M)^M \leq M(e^{-1}M)^M$
$\Rightarrow \ln M! \leq \ln n - \ln \ln n - \ln(2e) \Rightarrow M! \leq \frac{n}{2e \ln n}$.

Altogether, $\Pr(\mathcal{E}) \leq e\sqrt{n}\Pr(\mathcal{E}') \leq \frac{e\sqrt{n}}{n^2} \leq \frac{1}{n}$.