Advanced Algorithms IV (Fall 2020)

Instructor: Chihao Zhang
Scribed by: Yunqing Li, Hongyi Jin

Last modified on Oct 2, 2020

In this lecture, we first introduce the discrete Poisson distribution and examine some of its properties. We then focus on its approximation of the binmoial distribution. We will show the approximation is accurate, with only linear loss. At last we show the utility of the approximation via the max load problem and coupon collector's problem.

1 The m-balls-into-n-bins model

The m-balls-into-n-bins model is the following simple random process: throwing m balls into n bins uniformly at random. We already met the model in previous lectures. Today we continue to discuss the model with some new tools.

Let X_i be the indicator of the event that the i-th bin is empty. Let us compute the probability that $X_i = 0$. In this case, each throw should miss the i-th bin. Since the probability that one throw hits the i-th bin is $\frac{1}{n}$, the probability of $X_i = 0$ can be calculated as:

$$\mathbf{Pr}\left[X_i = 0\right] = \left(1 - \frac{1}{n}\right)^m \approx e^{-\frac{m}{n}}$$

Define X as the number of empty bins among n bins, then $X = \sum_{i=1}^n X_i$. By the linearity of expectation, $\mathbf{E}\left[X\right] = \sum_{i=1}^n \mathbf{E}\left[X_i\right] = ne^{-\frac{m}{n}}$. It means that about $e^{-\frac{m}{n}}$ fraction of bins are empty, and this fraction decreases exponentially with m.

We can then generalize the argument to find out the probability that one bin has r balls for any constant r. It can be calculated as follows:

$$\begin{aligned}\mathbf{Pr}\left[\text{the } i\text{-th bin has } r \text{ balls}\right] &= \binom{m}{r}\left(\frac{1}{n}\right)^r\left(1 - \frac{1}{n}\right)^{m-r} \\ &\approx \frac{m!}{(m-r)!r!}n^{-r}e^{-\frac{m}{n}}\end{aligned}$$

Since m and n are much larger compared to r, we can approximate $\frac{m!}{(m-r)!}$ to m^r and get

$$\mathbf{Pr}\left[\text{the } i\text{-th bin has } r \text{ balls}\right] \approx \frac{1}{r!}\left(\frac{m}{n}\right)^r e^{-\frac{m}{n}}$$

The conditions for this approximation are $r = o(n)$ and $r = o(m)$.

If we define $\lambda = \frac{m}{n}$, then the result can also be written as:

$$\mathbf{Pr}\left[\text{the } i\text{-th bin has } r \text{ balls}\right] \approx e^{-\lambda}\frac{\lambda^r}{r!}$$

The result is exactly the p.d.f of Poisson distribution with mean $\lambda = \frac{m}{n}$.

2 Poisson Distribution

Definition 1. *A discrete random variable Y satisfies Poisson distribution with mean $\lambda > 0$ïijŇwritten as $Y \sim \mathrm{Pois}(\lambda)$, if for $r = 0, 1, \cdots$, $\mathbf{Pr}\left[Y = r\right] = e^{-\lambda}\frac{\lambda^r}{r!}$*

We first verify that the definition is indeed a distribution.

Theorem 2. $\mathrm{Pois}(\lambda)$ *is a probability distribution.*

Proof.

$$\sum_{r=0}^{\infty} \mathbf{Pr}\left[Y = r\right] = \sum_{r=0}^{\infty} e^{-\lambda}\frac{\lambda^r}{r!} = e^{-\lambda}\sum_{r=0}^{\infty}\frac{\lambda^r}{r!} = e^{-\lambda}\cdot e^{\lambda} = 1,$$

where we used the Taylor expansion of $e^x = \sum_{i=0}^{\infty}\frac{x^i}{i!}$. □

Secondly, we verify that the expectation of a r.v. with distribution $Pois(\lambda)$ is exactly its mean λ.

Theorem 3. *Suppose $Y \sim Pois(\lambda)$, then $\mathbf{E}\left[Y\right] = \lambda$.*

Proof.

$$\mathbf{E}\left[Y\right] = \sum_{r=0}^{\infty} r\cdot\mathbf{Pr}\left[Y = r\right] = \sum_{r=0}^{\infty} r\cdot e^{-\lambda}\frac{\lambda^r}{r!} = \lambda\cdot\sum_{r=1}^{\infty} e^{-\lambda}\frac{\lambda^{(r-1)}}{(r-1)!} = \lambda.$$

□

Poisson distribution has an important property:

Lemma 4. *Assuming for every $1 \le i \le n$, $X_i \sim \mathrm{Pois}(\lambda_i)$, then $\sum_{i=1}^{n} X_i \sim \mathrm{Pois}\left(\sum_{i=1}^{n}\lambda_i\right)$.*

Proof. We only need to prove the lemma for $n = 2$. The larger n case follows from an induction argument. Assuming $X \sim \mathrm{Pois}(\lambda_1)$ and $Y \sim \mathrm{Pois}(\lambda_2)$. Then for every $r \ge 0$,

$$\begin{aligned}
\mathbf{Pr}\left[X + Y = r\right] &= \sum_{k=0}^{r}\mathbf{Pr}\left[(X = k)\cap(Y = r-k)\right] \\
&= \sum_{k=0}^{r} e^{-\lambda_1}\frac{\lambda_1^k}{k!}\cdot e^{-\lambda_2}\frac{\lambda_2^{r-k}}{(r-k)!} \\
&= e^{-\lambda_1-\lambda_2}\sum_{k=0}^{r}\frac{\lambda_1^k\lambda_2^{r-k}}{k!(r-k)!} \\
&= e^{-\lambda_1-\lambda_2}\frac{1}{r!}\sum_{k=0}^{r}\binom{r}{k}\lambda_1^k\lambda_2^{r-k} \\
&= e^{-(\lambda_1+\lambda_2)}\frac{1}{r!}(\lambda_1+\lambda_2)^r
\end{aligned}$$

□

As hinted in the m-balls-into-n-bins model, Poisson distribution can be used to approximate binomial distribution when the mean λ is small. Now we prove this rigorously.

Theorem 5. $\mathrm{Bin}(n,p) \approx \mathrm{Pois}(np)$ *when* $np = O(1)$ *for sufficiently large* n*.*

Proof. Let $X \sim \mathrm{Bin}(n,p)$ and $\lambda = np$. We first give an upper bound of $\mathbf{Pr}[X = r]$ for any $r \geq 0$.

$$\begin{aligned}
\mathbf{Pr}[X = r] &= \binom{n}{r} p^r (1-p)^{n-r} \\
&= \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \\
&\leq \frac{n^r}{r!} p^r \frac{(1-p)^n}{(1-p)^r} \\
&\leq \frac{(np)^r}{r!} \cdot \frac{e^{-pn}}{1-pr} && \text{(because } (1-p)^r > 1-pr\text{)}
\end{aligned}$$

When $n \longrightarrow \infty, p \longrightarrow 0$, $\frac{(np)^r}{r!} \cdot \frac{e^{-pn}}{1-pr} = \frac{(np)^r}{r!} \cdot e^{-pn} = \frac{\lambda^r}{r!} e^{-\lambda}$. Hence the upper bound is $\frac{\lambda^r}{r!} e^{-\lambda}$. We then lower bound $\mathbf{Pr}[X = r]$.

$$\begin{aligned}
\mathbf{Pr}[X = r] &= \binom{n}{r} p^r (1-p)^{n-r} \\
&= \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \\
&\geq \frac{(n-r+1)^r}{r!} p^r (1-p)^n
\end{aligned}$$

From $1 + p \leq e^p$, we can derive that $1 - p \geq e^{-p}(1-p^2)$. Using this inequality,

$$\begin{aligned}
\mathbf{Pr}[X = r] &\geq \frac{((n-r+1)p)^r}{r!} e^{-pn} (1-p^2)^n \\
&\geq \frac{((n-r+1)p)^r}{r!} e^{-pn} (1-np^2)
\end{aligned}$$

When $n \longrightarrow \infty$,$((n-r+1)p)^r \longrightarrow (np)^r$ and $1 - np^2 \longrightarrow 1$. Hence $\frac{((n-r+1)p)^r}{r!} e^{-pn}(1-np^2) \longrightarrow \frac{\lambda^r}{r!} e^{-\lambda}$. The lower bound is also $\frac{\lambda^r}{r!} e^{-\lambda}$. □

3 Poisson Approximation

For every $i \in [n]$, let $X_i^{(m)}$ be the number of balls in the i-th bin in the m-balls-into-n-bins model. It is clear that $X_i^{(m)}$ and $X_j^{(m)}$ are correlated, although every X_i has identical distribution $\mathrm{Bin}(m, \frac{1}{n})$. This is the main difficulty to analyze the model in many situation. Surprisingly, if we replace these binomial random variables with *independent* Poisson variables, the distributions turn out to be the same under some condition.

Let $Y_i^{(m)} \sim \mathrm{Pois}(\frac{m}{n})$.

Theorem 6. *The distribution of* $\left(X_1^{(k)}, ..., X_n^{(k)}\right)$ *is the same as the distribution of* $\left(Y_1^{(m)}, ..., Y_n^{(m)}\right)$ *conditional on that* $\sum_{i=1}^n Y_i^{(m)} = k$*.*

Proof. Fix a nonnegative number $k \geq 0$. Let $X^{(k)} = (X_1^{(k)}, ..., X_n^{(k)})$ and $Y^{(m)} = (Y_1^{(m)}, ..., Y_n^{(m)})$. Then for all $\mathbf{k} = (k_1, ..., k_n)$ with $\sum_{i=1}^n k_i = k$:

$$\begin{aligned}
\mathbf{Pr}\left[Y^{(m)} = \mathbf{k} \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = k\right] &= \frac{\prod_{i=1}^n \mathbf{Pr}\left[Y_1^{(m)} = k_i\right]}{\mathbf{Pr}\left[\sum_{i=1}^n Y_i^{(m)} = k\right]} && \text{(independence of } Y_i^{(m)}\text{)} \\
&= \frac{\prod_{i=1}^n e^{-\frac{m}{n}} (\frac{m}{n})^{k_i} \frac{1}{k_i!}}{e^{-m} m^k \frac{1}{k!}} && \text{(sum of Poisson r.v. is a Poisson r.v.)} \\
&= n^{-k} \frac{k!}{k_1!...k_n!} \\
&= \mathbf{Pr}\left[X^{(k)} = \mathbf{k}\right]
\end{aligned}$$

□

The theorem tells us that one might try to use independent Poisson variables to replace correlated binomial variables when studying the balls-into-bins model. In this case, we have $k = m$. However, in order to apply the theorem, we need to handle the condition $\sum_{i=1}^n Y_i^{(m)} = k$. Sometimes we can simply drop the condition since the condition $\sum_{i=1}^n Y_i^{(m)} = m$ happens with reasonable probability.

Suppose we have a non-negative function $f : \mathbb{N}^n \to \mathbb{N}$, and we want to compute its expectation $\mathbf{E}\left[f(X_1^{(m)}, \cdots, X_n^{(m)})\right]$. This is challenging in general since $X_1^{(m)}, \cdots, X_n^{(m)}$ have a complex joint distribution. However, using the approximation $\mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)})\right]$, it becomes easier due to the independence of $Y_1^{(m)}, \cdots, Y_n^{(m)}$. The following theorem indicates that the approximation is good.

Theorem 7. $\mathbf{E}\left[f(X_1^{(m)}, \cdots, X_n^{(m)})\right] \leq e\sqrt{m} \cdot \mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)})\right]$

Proof.

$$\begin{aligned}
\mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)})\right] &= \sum_{k=0}^m \mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)}) \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = k\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i^{(m)} = k\right] \\
&\geq \mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)}) \,\middle|\, \sum_{i=1}^n Y_i^{(m)} = m\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i^{(m)} = m\right] \\
&= \mathbf{E}\left[f(X_1^{(m)}, \cdots, X_n^{(m)})\right] \mathbf{Pr}\left[\sum_{i=1}^n Y_i^{(m)} = m\right]
\end{aligned}$$

The last equality comes from the fact that the joint distribution of $Y_1^{(m)}, \cdots, Y_n^{(m)}$ given $\sum_{i=1}^n Y_i^{(m)} = m$ is the same as that of $X_1^{(m)}, \cdots, X_n^{(m)}$, according to Theorem 6. Since $\sum_{i=1}^n Y_i^{(m)}$ follows the Poisson distribution with mean m,

$$\mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)})\right] \geq \mathbf{E}\left[f(X_1^{(m)}, \cdots, X_n^{(m)})\right] e^{-m} \frac{1}{m!} m^m$$

By Stirling's formula, we know $m! = \sqrt{2\pi m}(\frac{m}{e})^m(1 + o(1))$. Here we use a looser bound: $m! \le e\sqrt{m}(\frac{m}{e})^m$, which can be verified by integration.

We can now derive

$$\mathbf{E}\left[f(Y_1^{(m)}, \cdots, Y_n^{(m)})\right] \ge \frac{1}{e\sqrt{m}}\mathbf{E}\left[f(X_1^{(m)}, \cdots, X_n^{(m)})\right],$$

which finishes the proof. □

Remark. *If f is* monotone*, the coefficient $e\sqrt{m}$ can be improved to 2.*

Usually we let f be the indicator of certain "bad event" B, which is 1 if the event B happens and 0 otherwise. Therefore $\mathbf{E}[f]$ is the probability that the bad event happens and we often want to argue that this probability is not too large. We can upper bound it in the independent Poisson world and the theorem implies an upper bound in the real world.

Corollary 8. *If B happens w.p. p in the Poisson case, then B happens w.p. at most $e\sqrt{m}p$ in the real model.*

We will show you two applications of Corollary 8 in the next section.

4 Applications of Poisson Approximation

4.1 Max Load

In previous classes, we have seen the max load problem: In the "m balls into n bins" model, let X_i be the number of balls in the i-th bin. We want to compute the number of balls in the fullest bin, that is $X = \max\limits_{i \in [n]} X_i$.

Assume $m = n$ as we did in the last class. We already showed via the union bound argument that for some constant c,

$$\mathbf{Pr}\left[X \ge \frac{c \log n}{\log\log n}\right] = O\left(\frac{1}{n}\right)$$

.

Now, using the Poisson approximation, we can show that for some other constant c',

$$\mathbf{Pr}\left[X < \frac{c' \log n}{\log\log n}\right] = O\left(\frac{1}{n}\right)$$

Denote $k = \frac{c' \log n}{\log\log n}$. Let $Y_1, \cdots, Y_n$ be the Poisson approximation of $X_1, \cdots, X_n$. It is clear that $Y_i \sim \text{Pois}(1)$. Denote $Y = \max\limits_{i \in [n]} Y_i$. If we can bound the probability of the event $Y < k$, then we can obtain a bound for $\mathbf{Pr}[X < k]$ according to Corollary 8. We have

$$\mathbf{Pr}[Y < k] = \mathbf{Pr}[Y_1 < k \wedge \cdots \wedge Y_n < k] = (\mathbf{Pr}[Y_1 < k])^n$$

The second equality follows from the fact that Y_is are independent. Then we focus on $\mathbf{Pr}[Y_1 < k]$.

$$\begin{aligned}\mathbf{Pr}[Y_1 < k] &= 1 - \mathbf{Pr}[Y_1 \ge k] \\ &\le 1 - \mathbf{Pr}[Y_1 = k] \\ &= 1 - \frac{1}{ek!}\end{aligned}$$

So $\mathbf{Pr}[Y < k] \le (1-\frac{1}{ek!})^n \le e^{-\frac{n}{k!}}$. We now need to prove that there exists a constant c' such that $e^{-\frac{n}{k!}} < n^{-2}$. Then we can obtain $\mathbf{Pr}[X < k] < \frac{e\sqrt{n}}{n^2} < \frac{1}{n}$ by Corollary 8. Note that

$$\begin{aligned} e^{-\frac{n}{k!}} < n^{-2} &\iff \frac{n}{k!} > 2\log n \iff (2\log n)\sqrt{2\pi k}(\frac{k}{e})^k < n \\ &\iff \log 2 + \log\log n + \frac{1}{2}\log 2\pi k + k(\log k - 1) < \log n \end{aligned}$$

Since $k\log k = \frac{c'\log n}{\log\log n}(\log c' + \log\log n - \log\log\log n)$, we just need to let c' be a constant less than 1, and then $e^{-\frac{n}{k!}} < n^{-2}$ holds when n is sufficiently large.

4.2 Coupon Collector's Problem, Re-revisited

Recall the coupon collector's problem: given n coupons, what's the expected number of coupons to draw with replacement before having drawn each coupon at least once? Let X_i be the number of draws to get the i-th distinct coupon while exactly $i-1$ distinct coupons are already in hand, and $X = \sum_{i=0}^{n-1} X_i$.

In lecture 2, we have shown that the expectation of X is $nH_n \approx n\ln n$. In lecture 3, we yield concentration results. By applying Markov Inequality, we get

$$\mathbf{Pr}[X > cnH_n] \le \frac{1}{n}$$

Furthermore, we tighten the concentration with Chebyshev's inequality:

$$\mathbf{Pr}[X > nH_n + cn] \le \frac{\pi^2}{6c}$$

Here, since coupon collector's problem can be thought of as balls-and-bins problem, we can use Poisson approximation to obtain much stronger results:

Theorem 9. $\mathbf{Pr}[X > n\log n + cn] = 1 - e^{-e^{-c}}$ *when n is sufficiently large.*

Proof. Suppose we throw $m = n\log n + cn$ balls. Let Y_i be the Poisson approximation of X_i, and $Y = \sum_{i=1}^n Y_i$. It's clear that $Y_i \sim \mathrm{Pois}(\log n + c)$. So we can get $\mathbf{Pr}[Y_1 = 0] = \frac{e^{-c}}{n}$. Since Y_is are independent and have identical distribution,

$$\begin{aligned} \mathbf{Pr}\left[\bigcap_{i=1}^{n} Y_i \neq 0\right] &= (1 - \mathbf{Pr}[Y_1 = 0])^n \\ &= (1 - \frac{e^{-c}}{n})^n \\ &\approx e^{-e^{-c}} \end{aligned}$$

Our next step is to show the Poisson approximation is accurate, which means the difference between the probability calculated in the Poisson case and in the real-world case is $o(1)$. Note that we can not directly apply Corollary 8 here, because it will multiply $e\sqrt{m}$ to $e^{-e^{-c}}$, making the bound too loose. Instead, we can define a bad event B: $|Y - m|$ is larger than a threshold. By the Chernoff bound for the Poisson distribution[1], we can show that the probability of B is $o(1)$. Now we can assume Y is almost m. Then, we show that the difference between "Y is exactly m" and "Y is almost m" makes an asymptotically negligible difference in the probability that all the coupons are collected in real-world case. Finally, by Theorem 6, on condition that $Y = m$, Y_i and X_i have the same joint distribution. So the Poisson approximation is accurate.

A rigorous proof can be found in Theorem 5.13 of the textbook [2]. □

References

[1] *Chernoff bound — Wikipedia, the free encyclopedia.* https://en.wikipedia.org/wiki/Chernoff_bound, 2020. 6

[2] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*, Cambridge university press, 2017. 6