Lecture 4: Balls and Bins

2024.3.7

Lecturer: 丁虎 *Scribe:* 黄震

Balls and Bins 是一类应用十分广泛的问题，比如：一个班级里有 50 个人，存在两人生日是同一天的概率是多少？将 m 条数据随机映射存储在一个长度为 n 的哈希表中，发生冲突导致的最长链表会有多长？将 m 个任务随机分配给 n 个服务器，单个服务器的最大负载会是多少？这些问题都可以被抽象为向框里面扔球的场景，在本节我们将对这一场景展开讨论。

1 问题描述

Balls and Bins 假设存在 m 个 Balls 和 n 个 Bins，每个 Ball 被随机独立地投到一个 Bin 中。

1.1 期望碰撞次数

如果两个球被投到同一个 Bin，那么我们说这两个球之间发生碰撞。首先，我们来看看投球的过程中期望会发生多少次碰撞。

如果令

$$X_{ij} = \begin{cases} 1 & \text{如果第 } i \text{ 个球和第 } j \text{ 个球落入同一个 Bin} \\ 0 & \text{否则} \end{cases}$$

那么累计碰撞次数就是 $X = \sum_{1\le i<j\le m} X_{ij}$。分析 X_{ij} 和 X 的期望：

$$\begin{aligned} \mathrm{E}[X_{ij}] = \Pr[X_{ij}=1] &= \sum_{l=1}^{n} \Pr[\text{第 } i \text{ 个球和第 } j \text{ 个球同时落入第 } l \text{ 个 Bin}] \\ &= n \cdot \frac{1}{n^2} = \frac{1}{n} \end{aligned}$$

$$\mathrm{E}[X] = \mathrm{E}[\sum_{1\le i<j\le m} X_{ij}] = \sum_{1\le i<j\le m} \mathrm{E}[X_{ij}] = \binom{m}{2}\frac{1}{n}$$

$\mathrm{E}[X]$ 即为期望碰撞次数。注意到，当 $m=\sqrt{2n}$ 时，$\mathrm{E}[X]$ 近似为 1。

1.2 生日悖论

Birthday Problem 假设房间里有 m 个人，每个人的生日均匀随机分布在一年的 $n(=365)$ 天中。当 m 多大时可以以很大的概率（比如 0.9）保证有两个人同一天生日？

根据鸽巢原理，当 $m=n+1$ 时，一定有两个人同一天生日。那么当概率下降时，m 是否会等比例下降？我们接下来进行分析。由前面的事实，我们仅需考虑 $m \le n$ 时：

$$
\begin{aligned}
\Pr[m\text{ 个人的生日都不相同}] &= 1\cdot(1-\frac{1}{n})\cdot(1-\frac{2}{n})\cdots(1-\frac{m-1}{n}) \\
&= \prod_{i=0}^{m-1}(1-\frac{i}{n}) \\
&\le \prod_{i=0}^{m-1} e^{-\frac{i}{n}} \\
&= e^{-\frac{m(m-1)}{2n}}
\end{aligned}
$$

代入 $n=365$ 和 $m=42$，此时所有人生日都不相同的概率小于 0.1，即以 0.9 的概率保证有两人同一天生日。同理，当 $m=60$ 时，即可以超过 99% 的概率保证有两人生日相同，由于这个数学事实十分违反直觉，故称其为**生日悖论（Birthday Paradox）**。由该原理可以引申出密码学中的生日攻击，感兴趣的同学可以参考 Birthday Attack。

1.3 不发生碰撞的概率

生日问题里关注的是 m 取何值时发生碰撞的概率很高，相反地，我们来考虑下如何保证不发生碰撞的概率很高。

$$
\begin{aligned}
\Pr[m\text{ 个球被投进不同的 Bins}] &= \prod_{i=0}^{m-1}(1-\frac{i}{n}) \\
&\ge \prod_{i=0}^{m-1} e^{-\frac{i}{n}-\frac{i^2}{n^2}} \\
&= \exp(-\frac{m(m-1)}{2n}-\frac{m(m-1/2)(m-1)}{3n^2})
\end{aligned}
$$

假设 d 是一个待定的常数，将 $m = d\sqrt{n}$ 代入上式，指数的第一项将起主导作用（第二项是 $o(1)$）。通过挑选合适的 d，我们可以保证最后的概率足够大。

2 最大负载（Max Load）

令 L_i 表示第 i 个 Bin 中含有的球的个数，我们称它为 load。在很多实际场景中，我们会希望 $M = \max_{1\leq i\leq n} L_i$ 不要太大（考虑服务器场景下即希望单个服务器的负载不要过大）。我们下面给出对该目标的一些分析：

Theorem 2.1. 当 $m = \Omega(n\log n)$ 时，

$$\mathrm{E}[M] = \Theta(\frac{m}{n})$$

Proof. 令 X_{ij} 是"第 j 个球落入第 i 个 Bin"这一事件的指示变量，则 $L_i = \sum_{1\leq j\leq m} X_{ij}$。因此 $\mathrm{E}[L_i] = \sum_{1\leq j\leq m} \mathrm{E}[X_i j] = \frac{m}{n}$，结合 M 的定义可知 $\mathrm{E}[M] \geq \frac{m}{n}$。另一方面，根据 Chernoff Bound，我们知道取合适的 δ 可以获得

$$\Pr[L_i > (1+\delta)\frac{m}{n}] \leq \exp(-\frac{\delta^2}{3}\cdot\frac{m}{n}) = \exp(-\frac{\delta^2}{3}\Omega(\log n)) \leq \frac{1}{n}$$

即以至少 $1-\frac{1}{n}$ 的概率，可以保证 $L_i \leq O(\frac{m}{n})$，因此

$$\mathrm{E}[M] \leq (1-\frac{1}{n})\cdot O(\frac{m}{n}) + \frac{1}{n}\cdot n = O(\frac{m}{n})$$

综上，我们可以得到 $\mathrm{E}[M] = \Theta(\frac{m}{n})$。 □

Theorem 2.2. 当 $m = n$ 时，

$$\mathrm{E}[M] = \Theta(\frac{\log n}{\log\log n})$$

Proof. (此处仅证明上界) 直觉上，单个 Bin 中落入的球的个数不会特别大，我们来具体考察一下这个概率

$$\begin{aligned}\Pr[L_i \geq k] &\leq \sum_{1\leq<n_1<\cdots<n_k\leq m} \Pr[\text{第 } n_1,\ldots,n_k \text{ 个球落入第 } i \text{ 个 Bin}] \qquad (1)\\ &= \binom{m}{k}\frac{1}{n^k}\end{aligned}$$

式 (1) 的不等号由 Union Bound 可得。进一步，结合 Stirling 公式可知

$$\binom{m}{k} = \frac{m(m-1)\cdots(m-k+1)}{k!} \le \frac{m^k}{k!} \le m^k \cdot (\frac{e}{k})^k$$

因此, $\Pr[L_i \ge k] \le (\frac{em}{k})^k \frac{1}{n^k} = (\frac{e}{k})^k$。当 k 足够大时, 该概率将会足够小。令 $k = c\frac{\log n}{\log\log n}$, 其中 $c > 3$ 是一个待定的常数。代入得到

$$\begin{aligned}
\Pr[L_i \ge k] &\le (\frac{e\log\log n}{c\log n})^{\frac{c\log n}{\log\log n}} \\
&< (\frac{\log n}{\log\log n})^{-\frac{c\log n}{\log\log n}} \\
&= \exp(-\frac{c\log n}{\log\log n}(\log\log n - \log\log\log n)) \\
&= \exp(-c\log n + c\log n\frac{\log\log\log n}{\log\log n}) \\
&= e^{-c\log n + c\cdot o(\log n)} \\
&= n^{-c+o(1)}
\end{aligned}$$

再结合 Union Bound 可得

$$\Pr[M \ge k] \le \sum_{1\le i\le m} \Pr[L_i \ge k] \le n^{-c+1+o(1)}$$

这意味着取合适的 c，可以保证以至少 $1-\frac{1}{n}$ 的概率，有 $M < k$。因此

$$\mathrm{E}[M] \le (1-\frac{1}{n})\cdot(k-1) + \frac{1}{n}\cdot n = O(\frac{\log n}{\log\log n})$$

□

关于上述定理中下界的证明部分，参考 [1] 的 Lemma 5.12。

3 The Power of Two Choices

在前面我们已经得知，当 $m = n$ 时，最大负载 M 的期望是 $\Theta(\frac{\log n}{\log\log n})$。接下来我们考虑在放置球时引入一个小改动，从而获得更好的理论保证。

Two Choices 每次随机挑选 2 个 Bins，将球放入当前 load 较小的那个 Bin 中。

Lemma 3.1. 令 $X_1, \dots, X_n$ 是一组随机变量，$Y_1, \dots, Y_n$ 是一组 *0-1* 随机变量，并且 Y_i 依赖于 $X_1, \dots, X_i$，如果

$$\Pr[Y_i = 1 \mid X_1, \dots, \boldsymbol{X}_i] \le p$$

那么

$$\Pr[\sum_{i=1}^{n} Y_i > a] \le \Pr[B(n,p) > a]$$

其中 $B(n,p)$ 表示独立试验 n 次，每次成功概率为 p 的二项分布随机变量。

Theorem 3.2. 当 $m = n$ 时，如果采用上述的 *Two Choices* 策略，则

$$\mathrm{E}[M] = \Theta(\log\log n)$$

我们在此处仅考虑其上界，先给出一些直觉上的分析：我们假定 m 个球被依次放入 Bins 中，第 t 个球被放入后的状态称为时刻 t，对应地，$L_i(t)$ 表示时刻 t 时，第 i 个 Bin 内球的个数。令 $\nu_k(t)$ 表示在时刻 t 时，至少含有 k 个球的 Bins 的个数。

如果我们想获得一个含有 $k+1$ 个球的 Bin，那么在球到来时，选择的两个候选 Bins 都必须至少含有 k 个球，这意味着 $\Pr[N_{k+1}(t) \ge 1] \le (\frac{\nu_k(t-1)}{n})^2$。考虑到我们至多只会有 $\frac{n}{4}$ 个 Bins 含有 4 个球，选择两个 Bins 都来自于它们的概率为 $\frac{1}{16}$，从而直觉上会有 $\frac{n}{16}$ 个 Bins 含有 5 个球。同理，会有 $\frac{n}{256} = \frac{n}{2^{2^3}}$ 个 Bins 含有 6 个球，⋯，会有 $\frac{n}{2^{2^{k-3}}}$ 个 Bins 含有 k 个球。因此，最大负载应当是 $O(\log\log n)$ 量级。

接下来，我们给出详细证明：

Proof. 我们尝试构造一组递减的序列 β_k，使得对任意的 k，$\nu_k(n) \le \beta_k$ 以很高的概率成立。如果 $\beta_k < 1$ 时，那么 $M < k$，说明其衰减速度决定了 M 的量级。同时定义事件 $\Phi_k = \{\nu_k(n) \le \beta_k\}$，我们希望构造的 β_k 使得当 Φ_k 成立时，Φ_{k+1} 以很高的概率成立。

假设每个 Bin 都是一个栈，令 $h(t)$ 表示第 t 个球放置的高度，$\mu_k(t)$ 表示时刻 t 时，所有高度至少为 k 的球的个数。考虑某个固定的 k，定义 Y_t 为 “$h(t) \ge k+1$ 且 $\nu_k(t-1) \le \beta_k$” 这一事件的指示变量。令 ω_j 表示第 j 个球选择的 Bin，那么

$$\begin{aligned}
\Pr[Y_t = 1 \mid \omega_1, \dots, \omega_{t-1}] =& \Pr[h(t) \ge k+1 \mid \nu_k(t-1) \le \beta_k, \omega_1, \dots, \omega_{t-1}] \\
& \cdot \Pr[\nu_k(t-1) \le \beta_k \mid \omega_1, \dots, \omega_{t-1}] \\
\le& \Pr[h(t) \ge k+1 \mid \nu_k(t-1) \le \beta_k] \\
\le& (\frac{\beta_k}{n})^2 \overset{\text{def}}{=} p_k
\end{aligned}$$

利用引理3.1可得

$$\Pr[\sum_{t=1}^{n} Y_t > \beta_{k+1}] \leq \Pr[B(n, p_k) > \beta_{k+1}]$$

当 Φ_k 成立时，$\nu_k(t-1) \leq \beta_k$ 一定成立，此时 $\sum_{t=1}^{n} Y_t = \mu_{k+1}(n)$。又因为 $\nu_{k+1}(n) \leq \mu_{k+1}(n)$，可得

$$\begin{aligned}
\Pr[\neg\Phi_{k+1} \mid \Phi_k] &= \Pr[\nu_{k+1}(n) > \beta_{k+1} \mid \Phi_k] \\
&\leq \Pr[\sum_{t=1}^{n} Y_t > \beta_{k+1} \mid \Phi_k] \\
&\leq \frac{\Pr[\sum_{t=1}^{n} Y_t > \beta_{k+1}]}{\Pr[\Phi_k]} \\
&\leq \frac{\Pr[B(n, p_k) > \beta_{k+1}]}{\Pr[\Phi_k]}
\end{aligned}$$

令 $\beta_{k+1} = 2np_k$，考虑 $np_k \geq 6 \ln n$ 的情况，根据 Chernoff Bound 我们有 $\Pr[B(n, p_k) > \beta_{k+1}] \leq e^{-\frac{np_k}{3}} \leq \frac{1}{n^2}$。于是

$$\begin{aligned}
\Pr[\neg\Phi_{k+1}] &= \Pr[\neg\Phi_{k+1} \mid \Phi_k] \cdot \Pr[\Phi_k] + \Pr[\neg\Phi_{k+1} \mid \neg\Phi_k] \cdot \Pr[\neg\Phi_k] \\
&\leq \frac{1}{n^2} + \Pr[\neg\Phi_k] \qquad\qquad (np_k \geq 6 \ln n)
\end{aligned}$$

如果我们令 $\beta_4 = \frac{n}{4}$，由递推关系可得 $\beta_{k+4} = \frac{1}{2} \cdot \frac{n}{2^{2^k}}$。令 $k^* = \min\{k : np_k < 6 \ln n\}$，则 $k^* = O(\log \log n)$。注意到 $\Pr[\neg\Phi_4] = 0$，于是可得

$$\Pr[\neg\Phi_{k^*}] \leq \frac{k^*}{n^2}$$

接下来考虑 $np_k < 6 \ln n$ 即 $k \geq k^*$ 的部分：

$$\begin{aligned}
\Pr[\nu_{k^*+1}(n) > 12 \ln n \mid \Phi_{k^*}] &\leq \Pr[\mu_{k^*+1}(n) > 12 \ln n \mid \Phi_{k^*}] \\
&\leq \frac{\Pr[B(n, p_{k^*}) > 12 \ln n]}{\Pr[\Phi_{k^*}]} \\
&\leq \frac{\Pr[B(n, 6 \ln n / n) > 12 \ln n]}{\Pr[\Phi_{k^*}]} \\
&\leq \frac{1}{n^2} \cdot \frac{1}{\Pr[\Phi_{k^*}]}
\end{aligned}$$

此处最后一个不等号同样由 Chernoff Bound 得来。类似前面的操作，将概率中的条件移除：

$$\Pr[\nu_{k^*+1}(n) > 12\ln n] \le \frac{1}{n^2} + \Pr[\neg\Phi_{k^*}] \le \frac{k^*+1}{n^2}$$

此式子说明至少含有 k^*+1 个球的 Bins 大概率不超过 $12\ln n$ 个。更进一步

$$\begin{aligned}\Pr[\mu_{k^*+2}(n) \ge 2 \mid \nu_{k^*+1}(n) \le 12\ln n] &\le \frac{\Pr[B(n, (\frac{12\ln n}{n})^2) \ge 2]}{\Pr[\nu_{k^*+1}(n) \le 12\ln n]} \\ &\le \frac{\binom{n}{2}(\frac{12\ln n}{n})^4}{\Pr[\nu_{k^*+1}(n) \le 12\ln n]}\end{aligned}$$

继续沿用前面的方式，移除条件：

$$\begin{aligned}\Pr[\mu_{k^*+2}(n) \ge 2] &\le \binom{n}{2}(\frac{12\ln n}{n})^4 + \Pr[\nu_{k^*+1}(n) > 12\ln n] \\ &\le \binom{n}{2}(\frac{12\ln n}{n})^4 + \frac{k^*+1}{n^2} \\ &= o(\frac{1}{n})\end{aligned}$$

因此，$\Pr[\nu_{k^*+3}(n) \ge 1] \le \Pr[\mu_{k^*+2}(n) \ge 2] = o(\frac{1}{n})$。这意味着 $M \ge k^*+3$ 的概率不超过 $o(\frac{1}{n})$，于是

$$\mathrm{E}[M] \le (k^*+2) + n \cdot o(\frac{1}{n}) = O(\log\log n)$$

□

当我们将 Two Choices 策略中的候选 Bins 个数由 2 拓展到 d 时，可以以很大的概率保证 $M \le \frac{\log\log n}{\log d} + O(1)$。

References

[1] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, USA, 2005.