Lecture 2

The Chernoff bound and some applications

In this lecture we see a powerful concentration bound for sums of independent random variables, the Chernoff bound, and some applications of this bound. First let's get back to the streaming algorithm Count-Min Sketch.

2.1 Estimating individual frequencies: Count-Min Sketch

Let $\mathcal{H}$ be a family of 2-wise independent hash functions $h : \{1, \ldots, n\} \to \{1, \ldots, w\}$ for some w. Consider the following algorithm (see Figure 2.1):

1. **Initialization:** Choose $h_1, h_2, \ldots, h_d \in \mathcal{H}$ at random. Initialize $d \cdot w$ counters $c_{i,k} = 0$, where $i \in \{1, \ldots, d\}$ and $k \in \{1, \ldots, w\}$.
2. **Process a_j:** For each $i \in \{1, \ldots, d\}$ add 1 to $c_{i,h_i(a_j)}$.
3. **Output:** For each $x \in \{1, \ldots, n\}$, return $\tilde{f}_x = \min\{c_{i,h_i(x)},\ 1 \leq i \leq d\}$.

The idea is that if the h_i are independent, for each element x it is likely that there is an i such that x is the only element mapped to $h_i(x)$, in which case $c_{i,h_i(x)}$ will return an accurate count of the number of occurrences of x in the stream. We will prove the following theorem.

Theorem 2.1. *By choosing $w = \frac{2}{\epsilon}$ and $d = \log_2(\frac{1}{\delta})$, Count-Min Sketch returns an estimate $\tilde{f}_x$ for f_x that satisfies*

$$f_x \leq \tilde{f}_x \leq f_x + \epsilon m$$

with probability at least $1 - \delta$, using space $O(\frac{1}{\epsilon} \log_2(\frac{1}{\delta}))$.

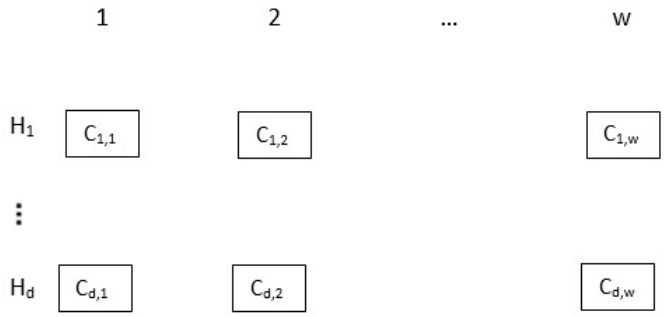

Figure 2.1: $d.w$ counters related to d different functions

Proof. It is clear that $\tilde{f}_x$ is always an overestimate for f_x, proving the first inequality. To show the other direction, fix an x and define random variables $Z_1, \ldots, Z_d$ by

$$Z_i = c_{i,h_i(x)} - f_x \Rightarrow Z_i = \sum_{y \neq x : h_i(x) = h_i(y)} f_y.$$

We also define auxiliary variable $X_{i,y}$ as follows:

$$X_{i,y} = \begin{cases} 1 & \text{if } h_i(y) = h_i(x) \\ 0 & \text{if } h_i(y) \neq h_i(x) \end{cases}$$

Then $Z_i = \sum_{x \neq y} f_y X_{i,y}$. Thus

$$\mathbb{E}[Z_i] = \sum_{y \neq x} f_y \mathbb{E}[X_{i,y}] = \sum_{y \neq x} f_y \, \mathbf{Pr}(h_i(y) = h_i(x)).$$

Due to 2-wise independence of h_i, we know that for any fixed j and $x \neq y$, $\mathbf{Pr}(h_i(y) = h_i(x) = j) = \frac{1}{w^2}$. But there are w possible j, so $\mathbf{Pr}(h_i(y) = h_i(x)) \leq w \frac{1}{w^2} = \frac{1}{w}$, thus

$$\mathbb{E}[Z_i] \leq \sum_{y \neq x} f_y \frac{1}{w} \leq \frac{m}{w}. \tag{2.1}$$

Using Markov's inequality,

$$\mathbb{P}\{Z_i \geq \epsilon m\} \leq \frac{\mathbb{E}[Z_i]}{\epsilon m} \overset{(2.1)}{\leq} \frac{1}{\epsilon w} \leq \frac{1}{2},$$

given our choice of w. Using that the Z_i are independent,

$$\mathbf{Pr}(\tilde{f}_x - f_x \geq \epsilon m) = \mathbf{Pr}(\forall i,\, c_{i,h_i(x)} - f_x \geq \epsilon m) = \mathbf{Pr}(\forall i,\, Z_i \geq \epsilon m) \overset{(*)}{\leq} \left(\frac{1}{2}\right)^{\log_2(d)} = \delta,$$

where $(*)$ follows from independence of the Z_i, and the last equality follows from our choice of d.

□

2.2 The median-of-means trick

Suppose that X is a random variable such that $\mathbb{E}[X] = \mu$ and X has a certain variance. Applying Chebyshev's inequality will give us a bound on the probability that X deviates significantly from its expectation. For instance, X could be the estimators for m_0 or m_2 constructed in the previous lecture, in which case the bound was not so good. A natural way to improve the result is by taking many independent copies of X and averaging them. If we take k copies, the variance will be divided by k, so that by Chebyshev the probability that we deviate from the expectation by a certain amount will also be divided by k. We can do better using the Chernoff bound:

Theorem 2.2. *For any $\varepsilon, \delta > 0$ let*

$$t = C \log \frac{1}{\delta} \qquad \text{and} \qquad k = 3\frac{\mathbf{Var}(X)}{\varepsilon^2 \mathbb{E}[X]^2},$$

where C is some universal constant. Let X_{ij}, for $i \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, k\}$, be independent random variables with the same distribution as X. Let

$$Z = \mathit{median}_{i \in \{1,\ldots,t\}} \Big(\frac{1}{k} \sum_{j=1}^{k} X_{ij} \Big).$$

Then $\mathbb{E}[Z] = \mu$ and $\mathbf{Pr}(|Z - \mu| \geq \varepsilon\mu) \leq \delta$.

Note that the number of copies required to drive the probability of error below δ scales as $\log(1/\delta)$, and not $1/\delta$ as it would if we were to rely only on Chebyshev's inequality.

Proof. Let $Y_i = \frac{1}{k} \sum_j X_{ij}$ for each $i \in \{1, \ldots, t\}$. Using linearity of expectation, $\mathbb{E}[Y_i] = \mu$, and using independence,

$$\mathbf{Var}(Y_i) = \frac{1}{k^2} \sum_j \mathbf{Var}(X_{ij}) = \frac{\mathbf{Var}(X)}{k}.$$

Applying Chebyshev's inequality, for each i,

$$\mathbf{Pr}(|Y_i - \mu| \geq \varepsilon\mu) \leq \frac{\mathbf{Var}(Y_i)}{\varepsilon^2\mu^2} = \frac{\mathbf{Var}(X)}{k\varepsilon^2\mathbb{E}[X]^2} = \frac{1}{3}.$$

For each i let W_i be a random variable that is 1 if $|Y_i - \mu| \geq \varepsilon\mu$. Then by the above bound $\mathbb{E}[W_i] \leq 1/3$, and $|Z - \mu| \geq \varepsilon\mu$ only if $\sum W_i > t/2$. Applying Chebyshev's inequality,

$$\begin{aligned} &\mathbf{Pr}\left(\sum W_i > t/2\right) \mathbf{Pr}\left(|\sum W_i - \mathbf{E}[\sum W_i]| > t/6\right) \\ &\qquad \leq \frac{t\,\mathbf{Var}(W_1)}{(t/6)^2} \\ &\qquad \leq \frac{1}{3}\frac{36}{t}, \end{aligned}$$

so that for the probability to be less than δ it is sufficient to take $t = 12/\delta$. But the theorem claims much better, $t = \log(1/\delta)$! That this t is enough is a consequence of the Chernoff bound, that we will see next. □

2.3 Chernoff Bounds

In the kind of scenario from the previous example, where $Z = X_1 + \cdots + X_n$ is the sum of independent random variables, it is possible to do much better than Markov or Chebyshev. Note that Chebyshev's inequality takes into account information about the variance, and it only requires the X_i to be pairwise independent. The Chernoff bound will do better by looking at all higher-order moments $\mathbb{E}[Z^k]$ simultaneously, and using full independence.

Let's first prove a bound in the simple setting of the coin tossing example. In that case for any $t \geq 0$ we can write

$$\begin{aligned}
\mathbf{Pr}\left(Z \geq (1+\delta)\mu\right) &= \mathbf{Pr}\left(e^{tZ} \geq e^{t(1+\delta)\mu}\right) && (x \mapsto e^x \text{is non-negative increasing}) \\
&\leq \frac{\mathbb{E}[e^{tZ}]}{e^{t(1+\delta)\mu}} && \text{(Markov's inequality)} \\
&= e^{-t(1+\delta)\mu} \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}] && \text{(independence)} \\
&= e^{-(1+\delta)\mu} \prod_{i=1}^{n} \Big(p_i e^t + (1-p_i)1\Big) && \text{(for Boolean } X_i\text{)} \\
&\leq e^{-t(1+\delta)\mu} \prod_{i=1}^{n} e^{p_i(e^t-1))} && \text{(Taylor series: } 1+x \leq e^x\text{)} \\
&= e^{(\mu(e^t-1)-t(1+\delta)\mu)}.
\end{aligned}$$

Now we solve for t to find the best possible bound. If we take the derivative of the exponential term with respect to t and set it to 0, we find that the RHS has a minimum at $\mu e^t - (1+\delta)\mu = 0$, i.e. $t = \ln(1+\delta)$. This gives us the final bound

$$\mathbf{Pr}\left(Z \geq (1+\delta)\mu\right) \leq \Big(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\Big)^\mu.$$

A similar proof can be done for $Z \leq (1-\delta)\mu$. So we've proved the following:

Theorem 2.3. *(Multiplicative Chernoff Bound] For any independent random variables $X_1, \ldots, X_n$ with $X_i \in (0,1]$ and $Z = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}[Z]$:*

$$\mathbf{Pr}\left(Z \geq (1+\delta)\mu\right) \leq \Big(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\Big)^\mu, \qquad \mathbf{Pr}\left(Z \leq (1-\delta)\mu\right) \leq \Big(\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}}\Big)^\mu.$$

Exercise 1. Show that for $\delta \in (0,1]$ the Chernoff bound implies the following weaker but often more convenient form:

$$\mathbf{Pr}\left(Z \geq (1+\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{3}}, \qquad \mathbf{Pr}\left(Z \leq (1-\delta)\mu\right) \leq e^{-\frac{\delta^2\mu}{2}}.$$

(**Note:** This follows from

$$(1+\delta)^{(1+\delta)} \geq e^{\delta+\frac{\delta^2}{3}}, \qquad \text{for } 0 \leq \delta \leq 1,$$

which can be proven by considering better and better approximation via the Taylor series of $\ln(1+\delta)$.)

We can now finish the analysis of the median-of-means trick:

End of proof of Theorem 2.2. Applying the Chernoff bound to the W_i,

$$\mathbf{Pr}\Big(\sum_{i=1}^{t} W_i > \frac{t}{2}\Big) \leq e^{-\frac{(1/2)^2(t/3)}{3}} = e^{-\frac{t}{36}} \leq \delta$$

provided the constant C from the theorem is chosen large enough. □

2.4 Balanced allocation

Suppose there are n servers and n jobs, how might we assign jobs to servers in a balanced way? This problem is equivalent to assigning n balls to n bins in a way that the load of the bin with the most balls is minimized.

Random Allocation

First, suppose that each ball is dropped in a random bin. Let Z_i count the number of balls in bin i: $Z_i = \sum_j X_{ij}$, where $X_{ij} = 1$ if ball j falls into bin i, and $X_{ij} = 0$ otherwise. Then

$$\mathbf{E}[Z_i] = \mathbf{E}\Big[\sum_{j=1}^{n} X_{ij}\Big] = \sum_{j=1}^{n} \mathbf{E}[X_{ij}] = \sum_{j=1}^{n} \frac{1}{n} = 1.$$

Using the Chernoff bound, for any $k \geq 1$

$$\mathbf{Pr}\left(Z_i \geq k\right) = \mathbf{Pr}\left(Z_i \geq k\,\mathbf{E}[Z_i]\right) \leq \Big(\frac{e^{k-1}}{k^k}\Big)^{\mathbf{E}[Z_i]} = \frac{e^{k-1}}{k^k}.$$

Let $k = \frac{3\ln n}{\ln\ln n}$. The reason for this choice will become clear soon. Applying the union bound, the probability that any bin has at least k balls is at most

$$\begin{aligned}
\mathbf{Pr}\Big(\text{any bin} \geq \frac{3\ln n}{\ln\ln n}\Big) &\leq n\Big(\frac{e\ln\ln n}{3\ln n}\Big)^{\frac{3\ln n}{\ln\ln n}} \\
&= n\exp\Big(\frac{3\ln n}{\ln\ln n}\Big(1+\ln\big(\frac{\ln\ln n}{3\ln n}\big)\Big)\Big] \\
&\leq n\exp\Big(\frac{3\ln n}{\ln\ln n}\Big(\ln\ln\ln n - \ln\ln n\Big)\Big] \\
&= n\exp\Big(\frac{3\ln n\cdot\ln\ln\ln n}{\ln\ln n} - 3\ln n\Big).
\end{aligned}$$

For large n, $\ln\ln\ln n \ll \ln\ln n$, and so

$$\mathbf{Pr}\left(\text{any bin} \geq \frac{3\ln n}{\ln\ln n}\right) \leq n\exp(-2\ln n) = \frac{1}{n} \tag{2.2}$$

The motivation for the choice of k is now clear. For $k = \frac{3\ln n}{\ln\ln n}$ and large n, the probability that there is a bin with many balls is small and decreases with n. It is possible to show that this is tight: with high probability there will always exist a bin that contains $\Omega(\ln n/\ln\ln n)$ balls. (**Note:** Show this in exercise.)

The Power of Two Choices

How can we achieve a more balanced allocation? Here is a simple trick we could try to play: for each ball, pick *two* bins at random, and drop the ball in the bin with fewer balls (to make the analysis simpler we'll allow that the two bins happen to be the same, in which case we have no choice).

Let B_i be the random variable that counts the number of bins with at least i balls, after all n balls have been distributed. Let β_i be an upper bound on B_i: $B_i \leq \beta_i$. Suppose that at iteration t, there are T_i bins having at least i balls each. What is the probability that the t-th ball is placed in a bin having at least i balls? For this to happen both bins selected need to have at least i balls. Then

$$\mathbf{Pr}\left(\text{ball } t \text{ placed in bin with} \geq \text{i balls}\right) = \left(\frac{T_i}{n}\right)^2 \leq \frac{B_i^2}{n^2} \leq \frac{\beta_i^2}{n^2}.$$

So the expected number of bins containing at least $i+1$ balls is at most the expected number of successes in a Bernoulli experiment with n trials and probability of success $\frac{\beta_i^2}{n^2}$:

$$\mathbf{E}[B_{i+1}] \leq n \cdot \frac{\beta_i^2}{n^2} = \frac{\beta_i^2}{n}.$$

Applying Markov's inequality, $\mathbf{Pr}(B_{i+1} \geq e\frac{\beta_i^2}{n}) \leq e^{-1}$. This motivates the following sequence of β_i's. Trivially, $B_6 \leq \frac{n}{6} \leq \frac{n}{2e}$. So we set

$$\beta_6 = \frac{n}{2e},\ \beta_7 = \frac{n}{2^2 e},\ \beta_8 = \frac{n}{2^4 e}, \ldots, \beta_i = \frac{n}{2^{2^{i-6}} e}.$$

Since $\mathbf{Pr}\left(B_{i+1} \geq e\frac{\beta_i^2}{n}\right)$ is low, the probability that $B_i \leq \beta_i$ for each i is high. Let E_i be the event that $B_i \leq \beta_i$. Note that $\mathbf{Pr}(E_6) = 1$.

Lemma 2.4. *For each i,* $\mathbf{Pr}(\neg E_{i+1}|E_i) \leq \frac{1}{n^2\,\mathbf{Pr}(E_i)}$.

Proof.

$$\begin{aligned}\mathbf{Pr}(\neg E_{i+1}|E_i) &= \frac{\mathbf{Pr}(\neg E_{i+1} \wedge E_i)}{\mathbf{Pr}(E_i)} \\ &\leq \frac{\mathbf{Pr}\left(\text{Binomial}\big(n, \frac{B_i^2}{n^2}\big) \geq \frac{e\beta_i^2}{n}\right)}{\mathbf{Pr}(E_i)} \\ &= \frac{\mathbf{Pr}\left(\text{Binomial}(n, \frac{B_i^2}{n^2}) \geq e\,\mathbf{E}(\text{Binomial}(n, \frac{B_i^2}{n^2}))\right)}{\mathbf{Pr}(E_i)}.\end{aligned}$$

Here, we used that $\beta_{i+1} = \frac{e\beta_i^2}{n}$. $\neg E_{i+1}$ then occurs if $\text{Binomial}(n, \frac{B_i^2}{n^2}) \geq B_{i+1} > \frac{e\beta_i^2}{n}$. Now, use the Chernoff bound as $\mathbf{Pr}(Z \geq e\,\mathbf{E}(Z)) \leq \exp(-\mathbf{E}(Z))$ to deduce $\mathbf{Pr}(\neg E_{i+1}|E_i) \leq \frac{\exp\left(-\frac{\beta_i^2}{n}\right)}{\mathbf{Pr}(E_i)} \leq \frac{\exp(-2\ln n)}{\mathbf{Pr}(E_i)} \leq \frac{1}{n^2\,\mathbf{Pr}(E_i)}$. □

Lemma 2.5. *For all i s.t. $\beta_i^2 \geq 2n\ln n$, $\mathbf{Pr}(\neg E_{i+1}) \leq \frac{i+1}{n^2}$.*

Proof. Use induction on i. The base case is $\mathbf{Pr}(\neg E_6) = 0 \leq \frac{7}{n^2}$. Next we have

$$\begin{aligned}\mathbf{Pr}(\neg E_{i+1}) &= \mathbf{Pr}(E_i)\,\mathbf{Pr}(\neg E_{i+1}|E_i) + \mathbf{Pr}(\neg E_i)\,\mathbf{Pr}(\neg E_{i+1}|\neg E_i) \\ &\leq \mathbf{Pr}(E_i)\frac{1}{n^2\,\mathbf{Pr}(E_i)} + \frac{i}{n^2}\,\mathbf{Pr}(\neg E_{i+1}|\neg E_i) \\ &\leq \frac{1}{n^2} + \frac{i}{n^2} \\ &\leq \frac{i+1}{n^2},\end{aligned}$$

where $\mathbf{Pr}(\neg E_i) \leq \frac{i}{n^2}$ comes from the induction hypothesis.

□

We have shown that for all $\beta_i^2 \geq 2n\ln n$, the probability that β_i does not bound B_i is low, and decreases like $O\left(\frac{1}{n^2}\right)$. Now, we must consider the other case where $\beta_i^2 < 2n\ln n$. Let i^* be the minimum i for which $\beta_i^2 < 2n\ln n$. Then, $i^* \leq \frac{\ln\ln n}{\ln 2} + O(1)$.

Lemma 2.6. $\mathbf{Pr}(B_{i^*+2} \geq 1) \leq O\big(\frac{\ln(n)^2}{n}\big)$.

Proof. Define $E_{i^*+1} = \{B_{i^*+1} \leq 6\ln n\}$. Then

$$\begin{aligned}\mathbf{Pr}(\neg E_{i^*+1}) &\leq \mathbf{Pr}(B_{i^*+1} \geq 6\ln n|E_{i^*})\,\mathbf{Pr}(E_{i^*}) + \mathbf{Pr}(\neg E_{i^*}) \\ &\leq \frac{\mathbf{Pr}(\text{Binomial}(n, \frac{2\ln n}{n}) \geq 6\ln n)}{\mathbf{Pr}(E_{i^*})} \cdot \mathbf{Pr}(E_{i^*}) + \frac{1}{n} \\ &\leq \frac{1}{n^2} + \frac{1}{n} \\ &= O\Big(\frac{1}{n}\Big).\end{aligned}$$

Thus

$$\begin{aligned}
\mathbf{Pr}(B_{i^*+2} \geq 1) &\leq \mathbf{Pr}(B_{i^*+2} \geq 1 | E_{i^*+1}) \cdot \mathbf{Pr}(E_{i^*+1}) + \mathbf{Pr}(\neg E_{i^*+1}) \\
&\leq \frac{\mathbf{Pr}(\text{Binomial}(n, (6 \ln n/n)^2) \geq 1)}{\mathbf{Pr}(E_{i^*+1})} \cdot \mathbf{Pr}(E_{i^*+1}) + O\Big(\frac{1}{n}\Big) \\
&\leq \Big(\frac{6 \ln n}{n}\Big)^2 \cdot n + O\Big(\frac{1}{n}\Big) \\
&= O\Big(\frac{(\ln n)^2}{n}\Big),
\end{aligned}$$

as desired.

□