Lecture 2

The Chernoff bound and median-of-means amplification

2.1 Derandomization

In the previous lecture we introduced a streaming algorithm for estimating the frequency moment $F_2$. The algorithm used very little space, logarithmic in the length of the stream, provided we could store a random function $h : \{1, \ldots, n\} \to \{\pm 1\}$ “for free” in memory. In general this would require $n$ bits of memory, one for each value of the function. However, observe that our analysis used the fact that $h$ is a random function in a rather weak way: specifically, when computing the expectation and variance of the random variable $Z = c^2$ describing the outcome of the algorithm we used conditions such as $E[Y_i Y_j Y_k Y_\ell] = E[Y_i] E[Y_j] E[Y_k] E[Y_\ell]$ for distinct values $i, j, k, \ell$, where $Y_j = h(j)$ is the random variable that describes the output of the function at a particular point. As it turns out, this requirement is a much weaker requirement than full independence, called 4-wise independence. In particular, it is possible to sample “4-wise independent” functions using much fewer random bits than a uniformly random function. This is the idea behind derandomization: to try to save on the number of random coins needed while keeping the function $h$ “random enough” that the analysis carries over.

2.1.1 $k$-wise independent random variables and hash functions

Definition 2.1. A family of random variables $(X_1, \ldots, X_N)$ is called $k$-wise independent if for every $k$-tuple $(i_1, \ldots, i_k) \in \{1, \ldots, N\}^k$ the random variables $(X_{i_1}, \ldots, X_{i_k})$ are independent.

To see the difference between 2-wise (also called pairwise) independence and full independence, consider for example a triple of random variables $(X_1, X_2, X_3)$ that is uniformly distributed over $\{(1,1,-1), (1,-1,1), (-1,1,1), (-1,-1,-1)\}$. Then $(X_1, X_2, X_3)$ are certainly not independent: the product is always $-1$. But you can check that they are pairwise independent. Note how this lets us save on the randomness: to generate a sample $(X_1, X_2, X_3)$
we only need 2 random bits, instead of 3 for fully independent random variables.

As an immediate consequence of the definition, you can see that if the \( Y_i \) are four-wise independent then the equality \( \mathbb{E}[Y_i Y_j Y_k Y_\ell] = \mathbb{E}[Y_i] \mathbb{E}[Y_j] \mathbb{E}[Y_k] \mathbb{E}[Y_\ell] \) always holds, which is all that was needed for the analysis of our \( F_2 \) algorithm: we only need the values produced by \( h \) to be 4-wise independent, not fully independent. Here is a reformulation of this requirement:

**Definition 2.2.** A family \( \mathcal{H} \) of functions \( h : A \mapsto B \) is called \( k \)-wise independent if for any distinct points \( x_1, ..., x_k \in A \) and \( i_1, ..., i_k \in B \),

\[
\Pr_{h \in \mathcal{H}} (h(x_1) = i_1, \ldots, h(x_k) = i_k) = \frac{1}{|B|^k}.
\]

A 1-wise independent family of hash functions is just a family such that any element \( x \) in the domain is mapped to a random element in the range, when the function is chosen at random. In general, the requirement of \( \mathcal{H} \) being \( k \)-wise independent is the same as requiring that the random variables \( X_i = h(x_i) \), sampled by evaluating a random \( h \leftarrow \mathcal{H} \) at a fixed point \( x_i \), are \( k \)-wise independent random variables.

**Example.** For \( A = B = \{0, 1, ..., p - 1\} \) where \( p \) is a prime number, consider the following family of functions:

\[
\mathcal{H}_2 = \{f_{a,b} : x \mapsto ax + b \mod p, (a,b) \in \{0, \ldots, p - 1\}^2\}.
\]

Then \( \mathcal{H}_2 \) is a family of 2-wise independent hash functions. To check this we only need to evaluate, for any \( x_1 \neq x_2, i_1 \) and \( i_2 \),

\[
\Pr_{a,b} (ax_1 + b = i_1 \land ax_2 + b = i_2) = \Pr_{a,b} \left( a = \frac{i_2 - i_1}{x_2 - x_1}, b = i_2 - x_2a \right) = \frac{1}{p^2},
\]

since \( a \) and \( b \) are chosen independently and uniformly at random.

When we study derandomization it will be important to construct many \( k \)-wise independent random variables using the fewest possible random bits. Here we have \( |A| = p \) random variables, but the number of random bits is only what is required to choose a random \( h \in \mathcal{H} \), so about \( 2 \log p \) bits. Compare this to \( \log p \) bits needed to choose just one random value in \( \{0, \ldots, p - 1\} \!\).!

### 2.1.2 Using pairwise independence

Now we can go back to the \( F_2 \) algorithm and modify it as follows: instead of using a completely random function \( h \) we use a random function \( h : \{1, \ldots, n\} \mapsto \{1, \ldots, n\} \) that is taken from a family \( \mathcal{H} \) of 4-wise independent hash functions. You can easily check that the whole analysis goes through unchanged. But now the number of random bits required to choose \( h \) is only \( O(\log n) \) (you will see an efficient construction of a family of 4-wise
independent hash functions in your homework). To store the function, we simply store the random coins. Every time we need to evaluate \( h(j) \) we read the random coins, recover the function, and evaluate it at the desired value. Thus for our algorithm to be time-efficient (and not only space-efficient) it is important that the evaluation of the function at a particular point can be done quickly, given the “raw” random coins. This is the case for the example of pairwise independent functions we saw earlier.

2.2 The median-of-means trick

In the previous lecture we gave a variance-based analysis of our streaming algorithm for estimating the frequency moment \( F_2 \). Our main tool for the analysis was Chebyshev’s inequality, which is useful whenever we have good control over the variance \( E[(Z - \mu)^2] \) of \( Z \). Using the inequality we were able to show that our algorithm gives an answer such that, first the answer is correct on expectation, and second it falls within a constant multiplicative factor of the right answer with constant probability. Now we’ll see how to “boost” both the accuracy and success probability of any such algorithm, by running it multiple times.

First let’s make a very simple observation. Suppose we execute any randomized procedure \( N \) times, using independent random coins at each time. Let the results be random variables \( X_1, \ldots, X_N \), and set \( Z \) to be the average result, \( Z = (1/N)(X_1 + \cdots + X_N) \). Then

\[
E[Z] = \frac{1}{N}(E[X_1] + \cdots + E[X_N]) = \frac{1}{N}(\mu_1 + \cdots + \mu_N),
\]

and

\[
\text{Var}(Z) = E]\left[\left(\frac{1}{N}(X_1 + \cdots + X_N) - \frac{1}{N}(\mu_1 + \cdots + \mu_N)\right)^2\right]
\]
\[
= \frac{1}{N^2} \sum_i E[(X_i - \mu_i)^2] - \frac{2}{N^2} \sum_{i,j} E[(X_i - \mu_i)(X_j - \mu_j)]
\]
\[
= \frac{1}{N^2} \sum_i \text{Var}(X_i),
\]

where in the last line we used that \( X_i \) and \( X_j \) are independent for \( i \neq j \) to write \( E[X_i X_j] = E[X_i] E[X_j] \). This shows that taking the average of multiple independent runs decreases variance linearly, with a corresponding improvement in accuracy for our algorithm. Moreover, notice how once again the only property we need is really just pairwise independence. Therefore, this method of amplification would work even if we didn’t use completely independent functions \( h \) for each execution of the algorithm, letting us save once again on the random bits.

But suppose now that we can afford full independence, can we get a more efficient amplification of the success probability? We will see how this can be done by using a powerful concentration bound for sums of independent random variables, the Chernoff bound.
Theorem 2.3. For any $\varepsilon, \delta > 0$ let

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

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

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

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

Note that the number of copies required to drive the probability of error below $\delta$ 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,

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

Applying Chebyshev’s inequality, for each $i$,

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

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 $W = \sum W_i > t/2$. Applying Chebyshev’s inequality,

$$\Pr \left( \sum_{i=1}^{n} W_i > \frac{t}{2} \right) \leq \Pr \left( \left| \sum_{i=1}^{n} W_i - \mathbb{E} \left[ \sum_{j=1}^{n} W_j \right] \right| > \frac{t}{6} \right) \leq \frac{t \text{Var}(W_1)}{(t/6)^2} \leq \frac{136}{3 t},$$

since $\text{Var}(W_1) \leq \mathbb{E}[W_1^2] = \mathbb{E}[W_1] \leq 1/3$. To make this bound less than $\delta$ 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. \qed
2.3 Chernoff Bounds

In the kind of scenario from the previous example, where $W = W_1 + \cdots + W_n$ is the sum of independent random variables, it is possible to do much better than Markov or Chebyshev. Chebyshev’s inequality takes into account information about the variance, and it only requires the $W_i$ to be pairwise independent. The Chernoff bound will do better by looking at all higher-order moments $E[W^k]$ simultaneously, and using full independence.

Here is how we do it. For any $t \geq 0$ we can write

$$\Pr(W \geq (1 + \delta)\mu) = \Pr(e^{tW} \geq e^{t(1+\delta)\mu}) \quad (x \mapsto e^x \text{is non-negative increasing})$$

$$\leq \frac{\mathbb{E}[e^{tW}]}{e^{t(1+\delta)\mu}} \quad \text{(Markov’s inequality)}$$

$$= e^{-t(1+\delta)\mu} \prod_{i=1}^{n} \mathbb{E}[e^{tW_i}] \quad \text{(independence)}$$

$$= e^{-t(1+\delta)\mu} \prod_{i=1}^{n} \left(p_i e^t + (1 - p_i)1\right) \quad \text{(for Boolean $W_i$)}$$

$$\leq e^{-t(1+\delta)\mu} \prod_{i=1}^{n} e^{p_i(e^t-1)} \quad \text{(Taylor series: $1 + x \leq e^x$)}$$

$$= e^{(e^t-1)-t(1+\delta)\mu}.$$ 

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

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

A similar proof can be done for $W \leq (1 - \delta)\mu$.

**Exercise 1.** In the fourth step above we used the fact that $W_i \in \{0, 1\}$. Suppose now we only assume that $W_i \in [0, 1]$, with $\mathbb{E}[W_i] = \mu/n$. How does that step need to be updated?

Assuming the result of the exercise, we have proved the following:

**Theorem 2.4.** *(Multiplicative Chernoff Bound)* For any independent random variables $X_1, \ldots, X_n$ with $X_i \in (0, 1]$ and $Z = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}[Z]$:

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

**Exercise 2.** Show that for $\delta \in (0, 1]$ the Chernoff bound implies the following weaker but often more convenient form:

$$\Pr(Z \geq (1 + \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}, \quad \Pr(Z \leq (1 - \delta)\mu) \leq e^{-\frac{\delta^2 \mu}{3}}.$$
We can now finish the analysis of the median-of-means trick:

*End of proof of Theorem 2.3.* Applying the Chernoff bound to the $W_i$,

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

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