CS 787: Advanced Algorithms

A Primer on Discrete Probability

Instructor: Dieter van Melkebeek

These notes provide a brief review of certain concepts in probability, random variables, expectation, and concentration. These concepts are important for the subsequent discussion of randomized algorithms.

1 Probability Theory over Countable Spaces

Definition 1 (Countable Probability Space). *A Countable Probability Space is composed of the tuple* $(\Omega, \Pr)$.

- ***Countable Sample Space*** Ω
  The sample space is a set of outcomes for a particular experiment. "Countable" means that either the size of Ω *is finite or there is a one-to-one and onto correspondence between the the elements of* Ω *and the set of Natural numbers.*

- ***Probability measure*** $\Pr : \Omega \mapsto [0,1]$ ***such that*** $\sum_{\omega \in \Omega} \Pr(\omega) = 1$
  $\Pr$ *maps every element of* Ω *to a real number in the interval [0,1], such that the sum of the probabilities of all the elements of* Ω *is 1.*

Definition 2 (Event). *An Event A is a subset of* Ω*, and* $\Pr(A) = \sum_{\omega \in A} \Pr(\omega)$

Example:
Consider the Sample Space $\Omega = \{$ All the cards in a standard deck $\}$
with, $\Pr(\omega) = \frac{1}{52}$ for each $\omega \in \Omega$ (this is a uniform distribution over the set of all cards).
Some possible events and their probabilities:

- $A_1 = \{$The set of all Aces in the deck$\}$
  $\Pr(A_1) = 4(\frac{1}{52}) = \frac{1}{13}$

- $\mathrm{A}_2 = \{$ The set of all the Hearts in the deck $\}$
  $\Pr(A_2) = 13(\frac{1}{52}) = \frac{1}{4}$

Now, notice that $A_1 \cap A_2 = \{$The event that A_1 and A_2 occur simultaneously$\} = \{$The Ace of Hearts$\}$ and $\Pr(A_1 \cap A_2) = \frac{1}{52} = \Pr(A_1) \cdot \Pr(A_2)$. ⊠

Definition 3 (Independent Events). *Events* A_1 *and* A_2 *are said to be independent iff* $\Pr(A_1 \cap A_2) = \Pr(A_1) \cdot Pr(A_2)$.

In general, k events, $A_1, A_2, \ldots, A_{\mathrm{k}}$ *are mutually independent if* $(\forall I \subseteq [k]) \Pr(\bigcap_{i \in I} A_i) = \prod_{i \in I} \Pr(A_i)$

In simple words, two or more events are said to be independent, if the outcomes of one set of events do not affect the outcome of any of the other events.

Definition 4 (Pairwise Independence)**.** *Events* $A_1, A_2, \ldots, A_\mathrm{k}$ *are pairwise independent if every pair of them is independent.*

Example: Pick 3 bits such that their XOR is equal to zero. Then, they are pairwise independent since any two of them are independent. But they are not mutually independent, since the knowledge of any two of them fixes the other. ⊠

Example:
$\Omega = \{$ Outcome of a Die $\}$, $A_1 = \{$ outcome is even $\}$, and $A_2 = \{$ the opposite side is odd $\}$

Here, A_1 and A_2 are not independent, since the knowledge of the outcome of either of the two fixes the outcome of the other (opposite sides of a die always add up to 7). ⊠

Definition 5 (Conditional Probability)**.** *Given two events* A *and* B*, the* ***conditional probability*** *of* A *given* B *is defined as*

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1}$$

where $\Pr(B) > 0$*.*

Intuitively, the conditional probability gives a measure of dependence of one event on the other.

More specifically, the conditional probability $\Pr(A \mid B)$ is the probability that event A occurs given the fact that B occurred. That is, the probability of A given B is the ratio of elements in B that are also in A. This ratio is illustrated in Figure 1 as the number of elements in the horizontal shaded region divided by the elements in the circle denoting B's region.

Furthermore, from Equation 1, we see that

$$\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B) \tag{2}$$

For example, given the standard 6-sided die from a previous example, we know that $\Pr(A_1) = \frac{1}{2}$. However, $\Pr(A_1|A_2) = 1$ due to the fact that opposite sides of dice add up to 7 and thus, knowing the outcome A_2 yields a different probability of A_1 (That is, A_1 is completely known).

We generalize Equation 2 to probabilities of the intersection of multiple events:

$$\Pr(A_k \cap A_{k-1} \cap \cdots \cap A_1) = \Pr(A_k \mid A_1 \cap A_2 \cap \cdots \cap A_{k-1}) \cdots \Pr(A_3 \mid A_1 \cap A_2) \cdot \Pr(A_2 \mid A_1) \cdot \Pr(A_1) \tag{3}$$

This result is known as the **chain rule**.

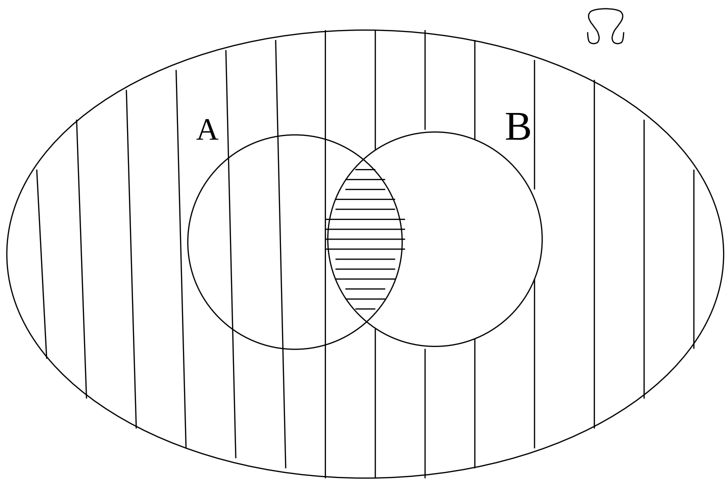

Figure 1: Conditional Probability $\Pr(A \mid B)$

2 Random variables

Definition 6 (Random Variable)**.** *A random variable is any function $f : \Omega \mapsto \mathbb{R}$, where Ω is the sample space.*

Example: The function that determines the running time of a randomized algorithm from the outcomes of the "coin flips" made during its execution is a random variable. ⊠

Example: An indicator variable χ_A for an event A is defined as

$$\chi_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ otherwise} \end{cases}$$

The error indicator, χ_{Err}, associated with the event "the algorithm produced erroneous output", is a random variable. ⊠

To get an idea of the behavior of a random variable f we would like to know what its *expected value* is.

Definition 7 (Expected Value)**.** *The expected value $E[f]$ of a random variable f is defined as*

$$E[f] = \sum_{\omega \in \Omega} f(\omega) \Pr(\omega) \tag{4}$$

Example: $E[\chi_A] = \Pr(A)$ ⊠

Example: Consider an experiment with probability p of success. Perform the experiment until it is successful.

$$
\begin{aligned}
E[\text{number of trials until first success}] &= \sum_{k=1}^{\infty} k(1-p)^{k-1} p \\
&= p \frac{d}{dp}\left(-\sum_{k=0}^{\infty}(1-p)^k\right) \\
&= p \frac{d}{dp}\left(-\frac{1}{p}\right) = \frac{1}{p}
\end{aligned}
$$

The first line of this evaluation states that for the k^{th} trial, the probability of success is p and the probability that we have seen failures in the previous $k-1$ iterations is $(1-p)^{k-1}$. Since we are calculating the expected *number of trials*, we multiply the probability of this event, $p(1-p)^{k-1}$ by the trials seen so far, k.

⊠

Example: One application of this last example is the search for an m-digit prime number. The prime number theorem tells us that the probability that a random number of m digits is prime is asymptotically equal to $\frac{1}{m \ln 2}$. Because we can check whether a number is prime in polynomial time, we can just pick an m digit number at random and run the primality test. By the previous example, the expected number of runs until we find a prime is $\Theta\left(\frac{1}{\frac{1}{m}}\right) = \Theta(m)$. ⊠

Example: Suppose we throw a die and let random variable X_1 be the number on the top of the die, and let random variable X_2 be the number on the bottom of the die. Clearly $E[X_1] = 7/2 = E[X_2]$. Now notice that $E[X_1 + X_2] = 7 = E[X_1] + E[X_2]$. In fact, in general the equality $E[X_1 + X_2] = E[X_1] + E[X_2]$ holds for any random variables X_1, X_2. Can we also use $E[X_1 \cdot X_2] = E[X_1] \cdot E[X_2]$? In general this is not the case, for our example a quick calculation gives $E[X_1 \cdot X_2] = \frac{1\cdot 6+2\cdot 5+3\cdot 4+4\cdot 3+5\cdot 2+6\cdot 1}{6} = 28/3$ and $E[X_1] \cdot E[X_2] = 49/4$. The reason for this is because the two variables are not independent. Let us now look into some properties for random independent variables. ⊠

Definition 8 (Independent Random Variables)**.** *Let $X_1, X_2, \ldots, X_n$ be random variables, we define them to be mutually independent if for all $x_1, x_2, \ldots, x_n$ the events $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ are mutually independent. Pairwise independence of random variables is defined in a similar way.*

We will now derive some properties which will be used often in calculations.

Theorem 1. *Let $X_1, X_2, \ldots, X_n$ be mutually independent random variables. then*

$$
E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i].
$$

Proof. For mutually independent random variables we can use the property that $\Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n] = \Pr[X_1 = x_1] \cdot \Pr[X_2 = x_2] \cdot \ldots \cdot \Pr[X_n = x_n]$. Using simple algebra we can then derive

$$\begin{aligned}
E\left[\prod_{i=1}^{n} X_i\right] &= \sum_{x_1, x_2, \ldots, x_n} \left(\prod_{i=1}^{n} x_i \cdot \Pr[X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n]\right) \\
&= \sum_{x_1, x_2, \ldots, x_n} \left(\prod_{i=1}^{n} x_i \cdot \Pr[X_1 = x_1] \cdot \Pr[X_2 = x_2] \cdot \ldots \cdot \Pr[X_n = x_n]\right) \\
&= \prod_{i=1}^{n} \left(\sum_{x_i} x_i \cdot \Pr[X_i = x_i]\right) \\
&= \prod_{i=1}^{n} E[X_i].
\end{aligned}$$

□

The expectation summarizes information on our probability distribution in one real. Sometimes the information contained in the expected value might not be sufficient. Suppose we have an algorithm which runs for 10 seconds half the time and 90 seconds the other half. A similar algorithm runs exactly 50 seconds every time. The expected running time for both algorithms is 50 seconds but clearly individual runs of the first algorithm deviate more from this expected value. We therefore define a new quantity called the variance which gives a good indication of the spread of a random variable around the expected value.

Definition 9 (Variance)**.** *The variance σ^2 of a random variable X is defined as*

$$\sigma^2[X] = E[(X - E[X])^2]. \tag{5}$$

It is easy to see that

$$\begin{aligned}
\sigma^2[X] &= E[(X - E[X])^2] \\
&= E[X^2 - 2XE[X] + E[X]^2] \\
&= E[X]^2 - 2E[X]E[X] + E[X]^2 \text{ (By linearity of expectation.)} \\
&= E[X^2] - E[X]^2.
\end{aligned}$$

The right hand side can be seen to be nonnegative. This result is also know as the Cauchy-Schwarz inequality. Recall that the expected value of a random variable is additive. The variance is additive, under the restriction that our random variables are pairwise independent:

Theorem 2. *Let $X_1, X_2, \ldots, X_n$ be pairwise independent random variables. Then,*

$$\sigma^2\left[\sum X_i\right] = \sum \sigma^2[X_i].$$

Proof. We will give the proof for two pairwise independent variables.

$$\begin{aligned}\sigma^2[X_1 + X_2] &= E[((X_1 + X_2) - (E[X_1] + E[X_2]))^2] \\ &= E[((X_1 - E[X_1]) + (X_2 - E[X_2]))^2] \\ &= E[((X_1 - E[X_1])^2 + (X_2 - E[X_2])^2 - 2(X_1 - E[X_1])(X_2 - E[X_2])] \\ &= \sigma^2[X_1] + \sigma^2[X_2].\end{aligned}$$

□

Exercise 1. *Extend the proof above for an arbitrary number of variables.*

3 Concentration

When analyzing randomized algorithms we often choose a certain random variable and compute its expected value. What we want to argue then is that within certain probability bounds, the algorithm will stay close to this expected value. The tail inequalities given in this section state that the random variable will be concentrated around its expected value. This gives us the guarantee that the probability that our algorithm behaves very badly is within certain tight bounds.

3.1 Markov's inequality

This inequality gives probability bounds in terms of the expected value.

Theorem 3 (Markov's Inequality)**.** *For any random variable $X \geq 0$, and for all real $t > 0$,*

$$\Pr[X \geq t] \leq \frac{E[X]}{t}. \tag{6}$$

Proof.

$$\begin{aligned}E[X] &= \sum_{x<t} x \cdot \Pr(X = x) + \sum_{x \geq t} x \cdot \Pr[X = x] \\ &\geq \sum_{x \geq t} x \cdot \Pr[X = x] \\ &\geq t \sum_{x \geq t} \Pr[X = x] \\ &= t \cdot \Pr[X \geq t].\end{aligned}$$

Rearranging the terms, we get

$$\Pr[X \geq t] \leq \frac{E[X]}{t}.$$

□

Note that the inequality is only informative when $t \geq E[X]$.
Example: Suppose we take as our random variable X the running time of our algorithm, Markov's inequality would state that the probability that our algorithm runs for twice the expected time is at least 50%. Thus, Markov's Inequality is a tool for proving bounds of concentration. ⊠

Markov's inequality has both an advantage and a disadvantage. One advantage is that since the proof assumes only a nonnegative random variable, the result has great applicability. However, this advantage comes with a less powerful bounding rate. We will see that the next two inequalities give tighter bounds.

3.2 Chebyshev's inequality

This inequality gives bounds in terms of the variance and is easily derived from Markov's inequality.

Theorem 4 (Chebyshev's Inequality). *For any random variable X, and for all real $t > 0$,*

$$\Pr\left[|X - E[X]| \geq t\right] \leq \frac{\sigma^2[X]}{t^2}. \tag{7}$$

Proof.

$$\begin{aligned}\Pr\left[|X - E[X]| \geq t\right] &= \Pr[(X - E[X])^2 \geq t^2] \\ &\leq \frac{E[(X - E[X])^2]}{t^2} \\ &= \frac{\sigma^2[X]}{t^2}.\end{aligned}$$

□

The proof uses Markov's inequality on the random variable $(X - E[X])^2$, which clearly satisfies the nonnegativity constraint. We will now derive a result using Chebyshev's inequality which frequently occurs in algorithm analysis.

Example: We have an experiment which we can run several times over again. Suppose we would like to model some property with an indicator variable. We assign each run a different random indicator variable called X_i for $i = 1...n$ with $E[X_i] = p$. If the runs are pairwise independent then so are the random variables X_i. Applying Chebyshev's inequality we find:

$$\begin{aligned}\Pr\left[\left|\frac{1}{n}\sum X_i - p\right| \geq a\right] &\leq \frac{\sigma^2\left[\frac{1}{n}\sum X_i\right]}{a^2} \\ &= \frac{\sum_{i=1}^{n} \sigma^2[X_i]}{n^2 a^2} \\ &= \frac{\sum_{i=1}^{n} np(1-p)}{n^2 a^2} \\ &\leq \frac{1}{na^2}\end{aligned}$$

The probability that the average of the random variables is off at least a from the expectation is inversely proportional to the number of runs n. ⊠

For some applications even this bound is not tight enough. We therefore state another tail inequality that gives even tighter bounds: Chernoff's inequality.

3.3 Chernoff's inequality

Chernoff's inequality bounds the probability of deviation with an exponentially decreasing function.

Theorem 5 (Chernoff's Inequality). *If $X_1, X_2, \ldots, X_n$ are mutually independent indicator variables, then for any $\delta \geq 0$*

$$\Pr\left[\sum_{i=1}^{n} X_i \geq (1+\delta)\mu\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}, \qquad \textit{where } \mu = E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E\left[X_i\right]. \tag{8}$$

Proof. The left-hand side of Equation 8 can be rewritten as

$$\Pr\left[e^{t\sum X_i} \geq e^{t(1+\delta)\mu}\right], \text{ for any } t > 0. \tag{9}$$

Since $e^{t\sum X_i} \geq 0$ and $e^{t(1+\delta)\mu} > 0$, we can apply Markov's inequality to get

$$\Pr\left[e^{t\sum X_i} \geq e^{t(1+\delta)\mu}\right] \leq \frac{E\left[e^{t\sum X_i}\right]}{e^{t(1+\delta)\mu}}, \textit{ for any } t > 0. \tag{10}$$

The numerator of the right-hand side of Equation 10 can be simplified as

$$E\left[e^{t\sum X_i}\right] = E\left[\prod e^{tX_i}\right]. \tag{11}$$

Since X_i's are mutually independent, so are tX_i's and e^{tX_i}'s for $t > 0$:

$$E\left[\prod e^{tX_i}\right] = \prod E\left[e^{tX_i}\right]. \tag{12}$$

From Equation 11 and 12,

$$E\left[e^{t\sum X_i}\right] = \prod E\left[e^{tX_i}\right]. \tag{13}$$

Let $E[X_i] = p_i$. Then $E\left[e^{tX_i}\right] = (1-p_i)\cdot 1 + p_i \cdot e^t$; therefore,

$$\begin{aligned}
\prod E\left[e^{tX_i}\right] &= \prod\left((1-p_i)\cdot 1 + p_i\cdot e^t\right) \\
&= \prod\left(1 + (e^t - 1)p_i\right) \\
&\leq \prod e^{(e^t-1)p_i}, \text{ since } 1 + x \leq e^x \\
&= e^{(e^t-1)\sum p_i} \\
&= e^{(e^t-1)\mu}, \text{ since } \mu = \sum E[X_i] = \sum p_i.
\end{aligned} \tag{14}$$

Using Equations 13 and 14, Equation 10 simplifies as

$$\Pr\left[e^{t\sum X_i} \geq e^{t(1+\delta)\mu}\right] \leq e^{\left((e^t-1)-t(1+\delta)\right)\mu},\ for\ any\ t > 0. \tag{15}$$

The right-hand side of Equation 15 is minimum when $\left((e^t - 1) - (1+\delta)\right)$ is minimum. Choosing t to minimize this expression,

$$\begin{aligned} \frac{d}{dt}\left(e^t - 1 - t(1+\delta)\right) &= 0 \\ e^t - (1+\delta) &= 0 \\ t &= \ln(1+\delta),\ \text{which is } > 0. \end{aligned} \tag{16}$$

Substituting this value of t in Equation 16 and from 9 we get,

$$\begin{aligned} \Pr\left[\sum_{i=1}^{n} X_i \geq (1+\delta)\mu\right] &\leq e^{(\delta-(1+\delta)\ln(1+\delta))\mu} \\ &= \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mu}. \end{aligned}$$

□

Note: The term $\left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)$ in Chernoff's inequality is < 1, because had we chosen the value of $t = 0$ in step 16 of the proof, the term evaluates to 1, but since we chose a t that further minimizes the term, the value should be < 1.
Example: Given mutually independent random variables $X_1, X_2, \ldots, X_n$ with $E[X_i] = p$,

$$\Pr\left[\frac{1}{n}\sum X_i \geq (1+\delta)p\right] \leq \left(\left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^{p}\right)^{n} = c^n \text{ where } c \text{ is a constant, } c < 1.$$

⊠

For now we want to note several things. First of all, Chernoff's inequality exists in many different variants. We have stated the theorem using indicator variables but you could derive a similar inequality for more general random variables. There also exists an expression bounding $\sum X_i$ from the other side of μ.

Theorem 6. *Let $X_1, X_2, \ldots, X_n$ be mutually independent random indicator variables, for any $\delta \leq 1$, with $\mu = \sum E[X_i]$,*

$$\Pr\left[\sum X_i < (1-\delta)\mu\right] \leq e^{-\frac{\delta}{2}\mu}.$$

3.4 Balls and bins

As an application, consider the process of throwing n balls into m bins uniformly at random. This process models several settings, including chained hashing and distributed load balancing. In the rest of this section we will use the latter terminology, but everything we say also applies to chained hashind and, more generally, to balls and bins.

It is often possible to analyze random variables related to balls and bins from first principles. Instead, we employ the Chernoff bound as an illustration of its applicability. The application can be seen as a first example of a randomized algorithm, namely for distributed load balancing.

Given: n jobs and m processors.
Goal: To distribute the jobs over the processors such that each processor gets about the same number of jobs.

The solution is straight forward: assign roughly $\frac{n}{m}$ jobs to every processor. This can be done by going through the processors and assigning a job to each and cycling back till all the jobs have been distributed. *However*, we introduce the constraint that the decision for assigning each job has to be made locally; i.e., when making the decision to assign a job, we do not know what the current distribution of jobs is. With this constraint, we may not be able to achieve the ideal load of $\frac{n}{m}$ on every processor. The best we know to do in this setting is to choose for each job a processor uniformly at random. This is exactly balls and bins.

We would like to analyze how close we get to the ideal solution. One question we can ask is, what is the probability that a processor gets a relative overload of more than δ?

Analysis

Let $X_{ij} = \chi](i_{th}$ job goes to the j_{th} processor)], where χ denotes the indicator variable of a certain event. The X_{ij}'s are not mutually independent because $X_{ik} = 1$ implies $X_{il} = 0$ for all $k \neq l$. However, for a fixed j, the X_{ij}'s are mutually independent for all i. Let Y_j denote the load on processor j; then

$$\begin{aligned} Y_j &= \sum_{i=1}^{n} X_{ij} \\ E[Y_j] &= \sum_{i=1}^{n} E[X_{ij}] = \sum_{i=1}^{n} \frac{1}{m} = \frac{n}{m}, \end{aligned}$$

which shows that the expected load on each processor matches the ideal solution. Nonetheless, we have yet to prove any bound on concentration around this “good” expected value. We use Chernoff's inequality:

$$\Pr\left[Y_j \geq (1+\delta)\frac{n}{m}\right] \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\frac{n}{m}}. \tag{17}$$

The above bound is for the probability of overload on one processor. The bound on the probability that any processor overloads is bounded by the sum of the probability of overloading for each processor. This is called the **union bound**, and is a useful when you have to take into account the error due to various factors when the factors are not independent.

$$
\begin{aligned}
\Pr\left[(\exists j)\left(Y_j \geq (1+\delta)\frac{n}{m}\right)\right] &\leq \sum_{j=1}^{m} \Pr\left(Y_j \geq (1+\delta)\frac{n}{m}\right) \\
&\leq \sum_{j=1}^{m} \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\frac{n}{m}} \\
&= m\left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\frac{n}{m}}. \qquad (18)
\end{aligned}
$$

We would like the above probability to be $\ll 1$. If we fix one of the parameters, then we can get a relation between the other two parameters in order to make the probability $\ll 1$.

1. For a fixed $\delta > 0$, $\left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)$ is a constant, say c. And we also know that, $c < 1$. Therefore,

$$
\begin{aligned}
mc^{\frac{n}{m}} &\ll 1 \\
c^{\frac{n}{m}} &\ll \frac{1}{m} \\
\frac{n}{m}\log c &\ll \log\frac{1}{m} \\
n\log\frac{1}{c} &\gg m\log m \\
n &= \Omega(m\log m). \qquad (19)
\end{aligned}
$$

   Therefore, we can guarantee that with high probability no processor gets a relative overload of δ or more if the number of jobs is at least a logarithmic factor larger than the number of processors.

2. If $n = \Theta(m)$, then it can be shown that

$$\delta = O\left(\frac{\log n}{\log\log n}\right). \qquad (20)$$

   The proof is left as an exercise. In terms of chained hashing with constant occupancy rate, this means that with high probability the maximum chain length is $O(\frac{\log n}{\log\log n})$.