ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Algorithms and Methods for Distributed Storage Networks

9 Analysis of DHT

Christian Schindelhauer

Albert-Ludwigs-Universität Freiburg
Institut für Informatik
Rechnernetze und Telematik
Wintersemester 2007/08

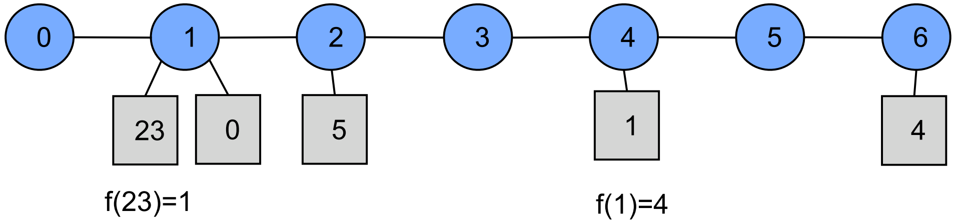
CoNe
Freiburg

IIF
INSTITUT FÜR
INFORMATIK
FREIBURG

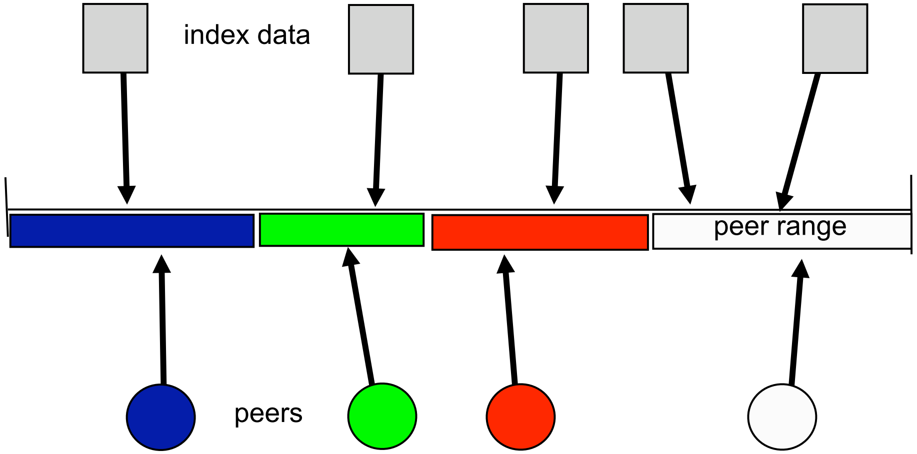
Distributed Hash-Table (DHT)

- **Hash table**
  - does not work efficiently for inserting and deleting
- **Distributed Hash-Table**
  - servers are „hashed“ to a position in an continuos set (e.g. line)
  - data is also „hashed“ to this set
- **Mapping of data to servers**
  - servers are given their own areas depending on the position of the direct neighbors
  - all data in this area is mapped to the corresponding server
- **Literature**
  - *“Consistent Hashing and Random Trees: Distributed Caching Protocols for Relieving Hot Spots on the World Wide Web”,* David Karger, Eric Lehman, Tom Leighton, Mathhew Levine, Daniel Lewin, Rina Panigrahy, STOC 1997

Pure (Poor) Hashing

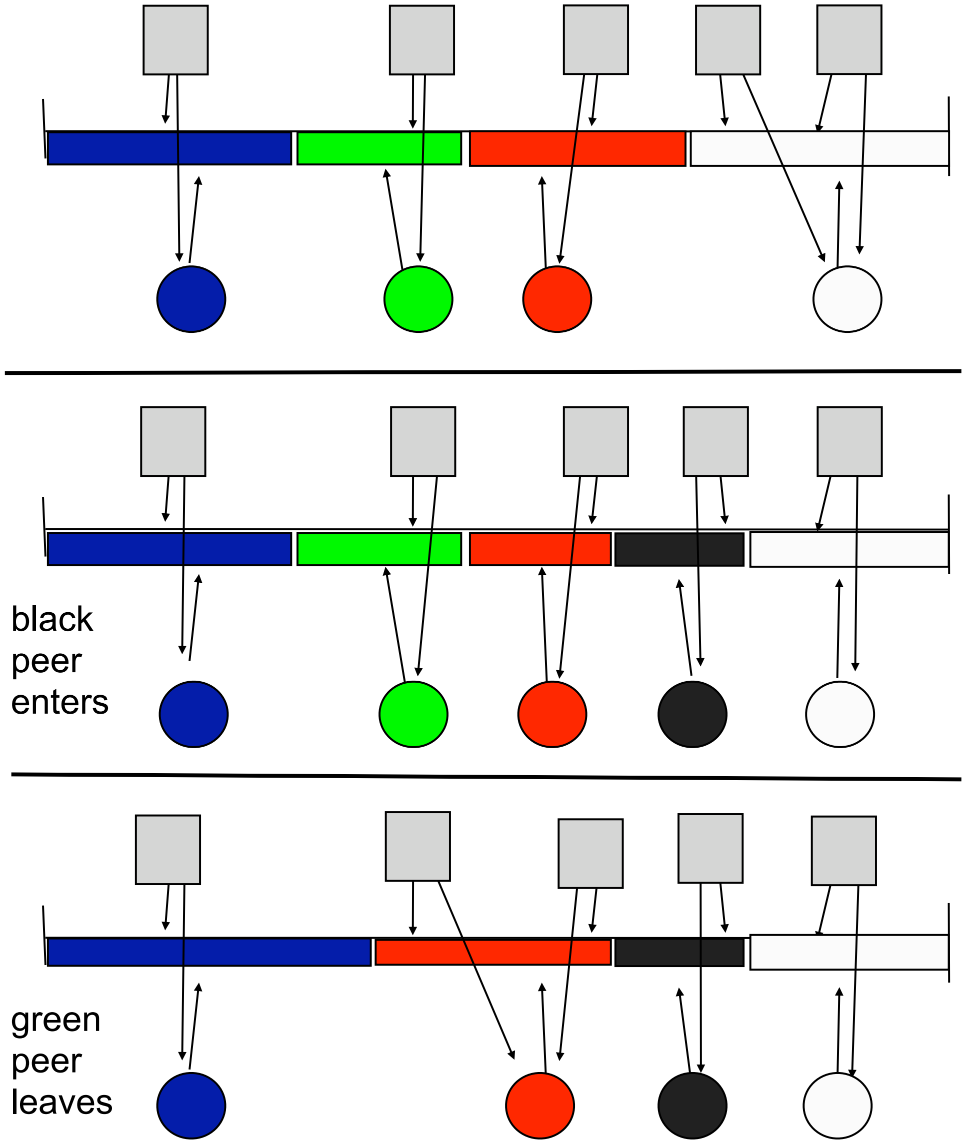


DHT

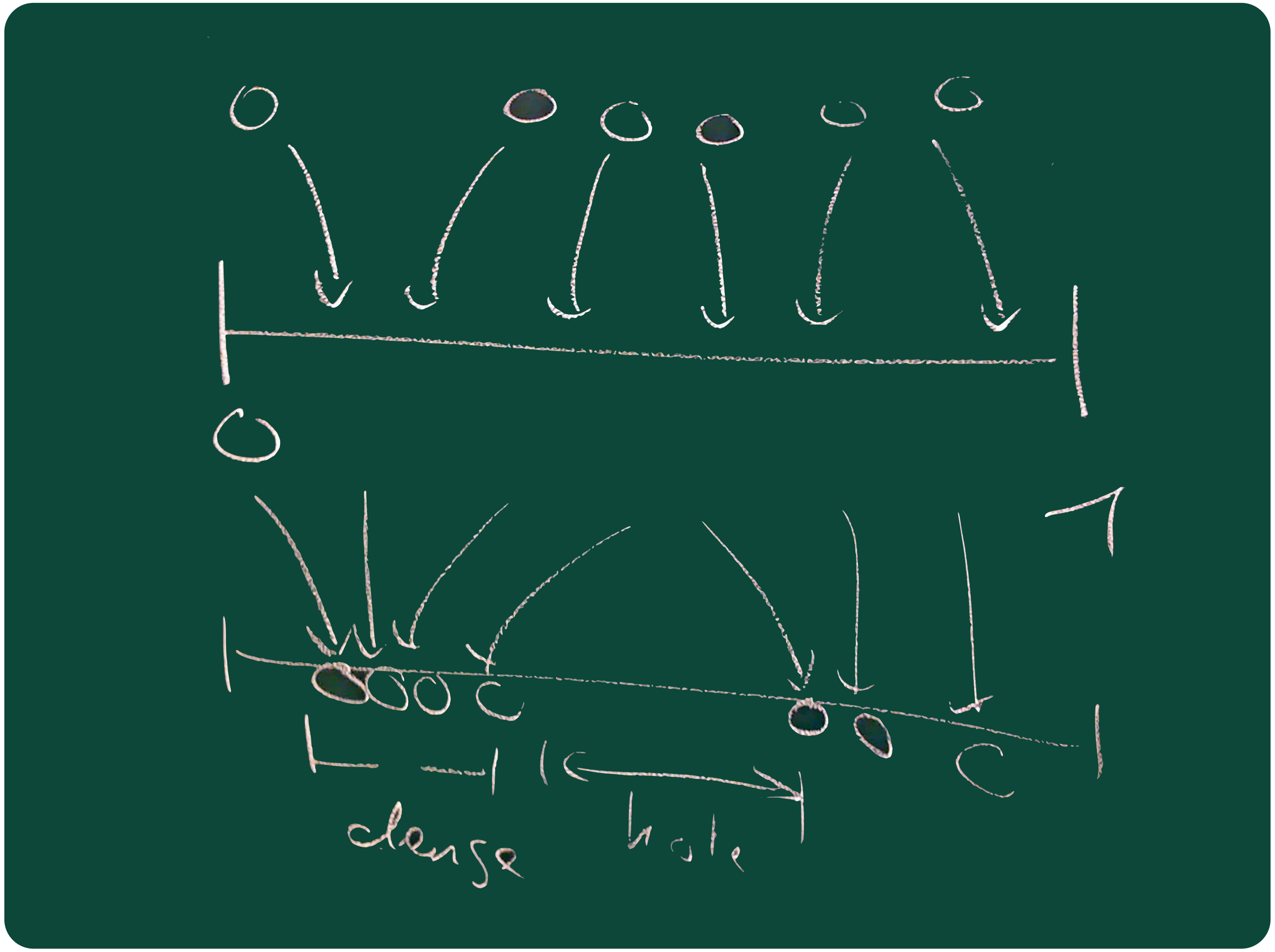


Entering and Leaving a DHT

- **Distributed Hash Table**
  - devices are hashed to to position
  - blocks are hashed according to the ID
- **When a device is added**
  - only blocks from neighbors have to be moved
- **When a device is deleted**
  - blocks are moved only to the neighbors



Holes and Dense Areas



Size of Holes

- **Theorem**
  - If n elements are randomly inserted into an array [0,1[then with constant probability there is a „hole“ of size Ω(log n/n), i.e. an interval without elements.
- **Proof**
  - Consider an interval of size log n / (4n)
  - The chance not to hit such an interval is (1-log n/(4n))
  - The chance that n elements do not hit this interval is

$$\left(1-\frac{\log n}{4n}\right)^{n} = \left(1-\frac{\log n}{4n}\right)^{\frac{4n}{\log n}\frac{\log n}{4}} \geq \left(\frac{1}{4}\right)^{\frac{1}{4}\log n} = \frac{1}{\sqrt{n}}$$

  - The expected number of such intervals is more than 1.
  - Hence the probability for such an interval is at least constant.

Proof of Dense Areas

$$\left(\frac{1}{4}\right)^{\frac{1}{4}\cdot\log n} = 2^{\left(\frac{1}{4}\log n\right)\overbrace{\log_2\frac{1}{4}}^{-2}}$$

$$= 2^{\left(-\frac{1}{2}\right)\cdot\log n}$$

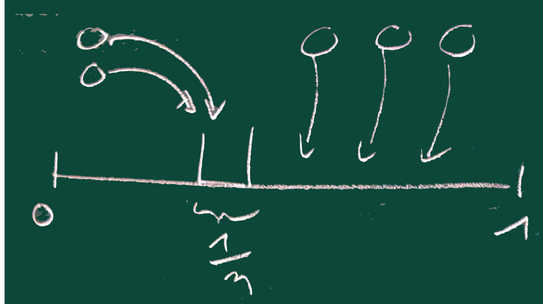
$$= n^{-\frac{1}{2}} = \frac{1}{\sqrt{n}}$$

Expectation: $\frac{4n}{\log n}\cdot\frac{1}{\sqrt{n}} = \frac{4\sqrt{n}}{\log n}$

Dense Spots

- **Theorem**
  - If n elements are randomly inserted into an array [0,1[then with constant probability there is a dense interval of length 1/n with at least Ω(log n/ (log log n)) elements.
- **Proof**
  - The probability to place exactly i elements in to such an interval is $\left(\frac{1}{n}\right)^i \left(1 - \frac{1}{n}\right)^{n-i} \binom{n}{i}$
  - for i = c log n / (log log n) this probability is at least $1/n^k$ for an appropriately chosen c and k<1
  - Then the expected number of intervals is at least 1

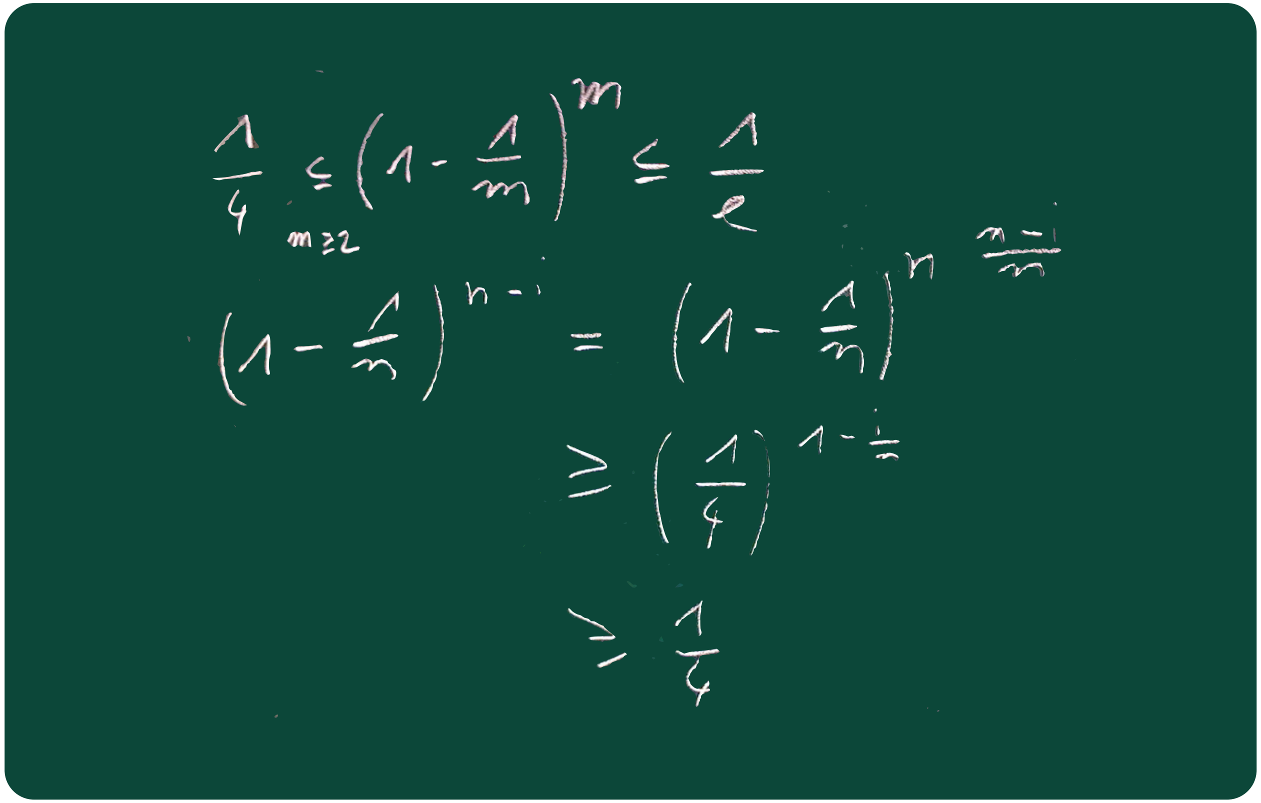
Proof of Dense Areas



$$i = \frac{c \cdot \log n}{\log \log n}$$

$$\Pr\left[\begin{array}{l} i \text{ Balls from } n \text{ Balls} \\ \text{fall into an interval of} \\ \text{size } \frac{1}{n} \end{array}\right] = \left(\frac{1}{n}\right)^i \underbrace{\left(1-\frac{1}{n}\right)^{n-i}}_{\substack{O(1) \\ \leq 1/e}} \underbrace{\binom{n}{i}}_{\geq n^i \cdot \frac{1}{n^k}} \geq \frac{1}{n^k} \quad k<1.$$

Proof of Dense Areas



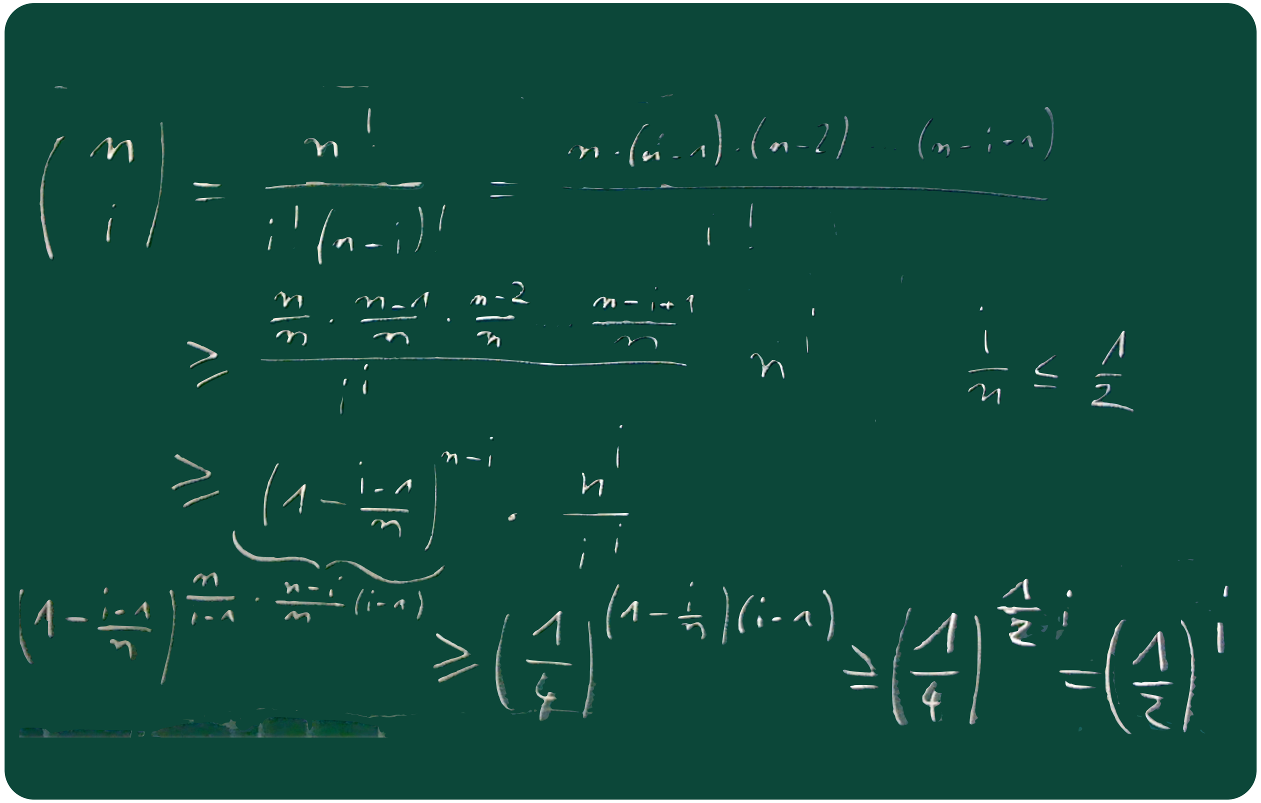
$$\frac{1}{4} \underset{m \geq 2}{\leq} \left(1 - \frac{1}{m}\right)^{m} \leq \frac{1}{e}$$

$$\left(1 - \frac{1}{n}\right)^{n-1} = \left(1 - \frac{1}{n}\right)^{n \frac{n-1}{n}}$$

$$\geq \left(\frac{1}{4}\right)^{1 - \frac{1}{n}}$$

$$\geq \frac{1}{4}$$

Proof of Dense Areas



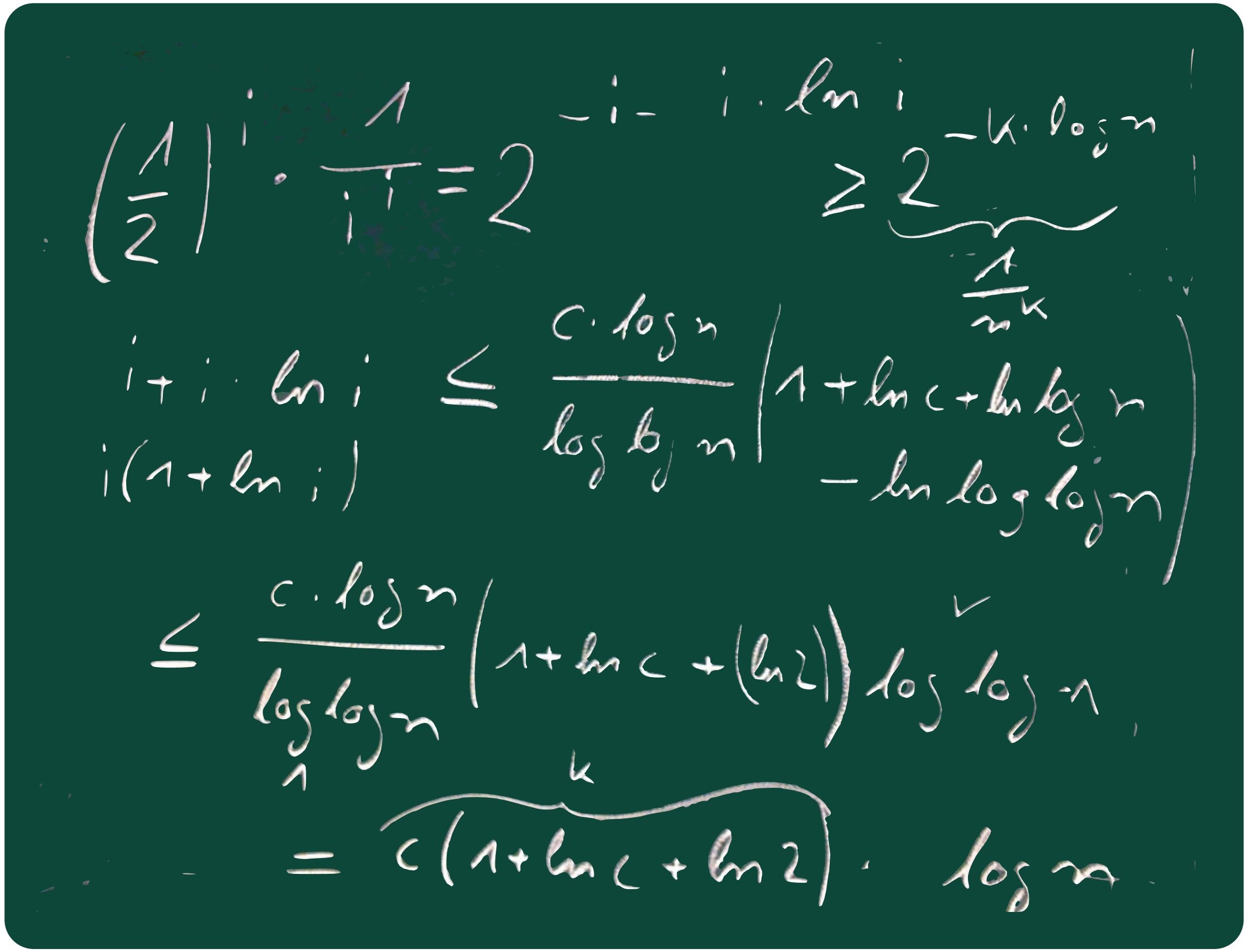
$$\binom{n}{i} = \frac{n!}{i!\,(n-i)!} = \frac{n\cdot(n-1)\cdot(n-2)\,\ldots\,(n-i+1)}{i!}$$

$$\geq \frac{\frac{n}{n}\cdot\frac{n-1}{n}\cdot\frac{n-2}{n}\,\ldots\,\frac{n-i+1}{n}}{i^i}\; n^i \qquad \frac{i}{n} \leq \frac{1}{2}$$

$$\geq \underbrace{\left(1-\frac{i-1}{n}\right)^{n-i}}_{\left(1-\frac{i-1}{n}\right)^{\frac{n}{i-1}\cdot\frac{n-i}{n}(i-1)}} \cdot \frac{n^i}{i^i}$$

$$\geq \left(\frac{1}{4}\right)^{\left(1-\frac{i}{n}\right)(i-1)} \geq \left(\frac{1}{4}\right)^{\frac{1}{2}i} = \left(\frac{1}{2}\right)^i$$

Proof of Dense Areas



$$\left(\frac{1}{2}\right)^{i} \cdot \frac{1}{i!} = 2^{-i - i \cdot \ln i} \geq \underbrace{2^{-k \cdot \log n}}_{\frac{1}{n^k}}$$

$$\begin{aligned} i + i \cdot \ln i &\leq \frac{c \cdot \log n}{\log \log n}\left(1 + \ln c + \ln \log n - \ln \log \log n\right) \\ i(1 + \ln i) & \end{aligned}$$

$$\leq \frac{c \cdot \log n}{\underset{1}{\log \log n}}\left(1 + \ln c + (\ln 2)\right) \overset{\checkmark}{\log \log n}$$

$$= \overbrace{c(1 + \ln c + \ln 2)}^{k} \cdot \log n$$

Averaging Effect

- **Theorem**
  - If Θ(n log n) elements are randomly inserted into an array [0,1[then with high probability in every interval of length 1/n there are Θ(log n) elements.

Excursion

- **Markov-Inequality**
  - For random variable X>0 with E[X] > 0:

$$\mathbf{P}[X \geq k \cdot \mathbf{E}[X]] \leq \frac{1}{k}$$

- **Chebyshev**

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq k] \leq \frac{\mathbf{V}[X]}{k^2}$$

  - for Variance $\mathbf{V}[X] = \mathbf{E}[X^2] - \mathbf{E}[X]^2$
- **Stronger bound: Chernoff**

Chernoff-Bound

- **Theorem Chernoff Bound**
  - Let $x_1,...,x_n$ independent Bernoulli experiments with
    - $P[x_i = 1] = p$
    - $P[x_i = 0] = 1-p$
  - Let $S_n = \sum_{i=1}^{n} x_i$
  - Then for all c>0

$$\mathbf{P}[S_n \geq (1+c) \cdot \mathbf{E}[S_n]] \quad \leq \quad e^{-\frac{1}{3}\min\{c,c^2\}pn}$$

  - For 0≤c≤1

$$\mathbf{P}[S_n \leq (1-c) \cdot \mathbf{E}[S_n]] \quad \leq \quad e^{-\frac{1}{2}c^2 pn}$$

Proof of 1st Chernoff Bound

- **We show** $$\boxed{\mathrm{P}[S_n \ \geq \ (1+c)\mathbf{E}[S_n]] \ \leq \ e^{-\frac{\min\{c,c^2\}}{3}pn}}$$
- **Für t>0:**

$$\mathrm{P}[S_n \geq (1+c)pn] \ = \ \mathrm{P}[e^{tS_n} \geq e^{t(1+c)pn}]$$

$$\frac{1}{k} \ \leq \ e^{-\frac{\min\{c,c^2\}}{3}pn}$$

$$k = e^{t(1+c)pn}/E[e^{t\cdot S_n}]$$

- **Markov yields:** $$\mathbf{P}\left[e^{tS_n} \geq k\mathbf{E}\left[e^{tS_n}\right]\right] \ \leq \ \frac{1}{k}$$
- **To do: Choose t appropriately**

Proof of 1st Chernoff Bound

- **We show** $\frac{1}{k} \leq e^{-\frac{\min\{c,c^2\}}{3}pn}$
- **where** $k = e^{t(1+c)pn} / E[e^{t \cdot S_n}]$

Independence of random variables x_i →

$$\begin{aligned}
\mathbf{E}[e^{tS_n}] &= \mathbf{E}\left[e^{t\sum_{i=1}^{n} x_i}\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{n} e^{tx_i}\right] \\
&= \prod_{i=1}^{n} \mathbf{E}\left[e^{tx_i}\right] \\
&= \prod_{i=1}^{n} (e^0(1-p) + e^t p) \\
&= (1-p+e^t p)^n \\
&= (1+(e^t-1)p)^n
\end{aligned}$$

- **Next we show:**

$$e^{-t(1+c)pn} \cdot (1+p(e^t-1))^n \leq e^{-\frac{\min\{c,c^2\}}{3}pn}$$

Proof of 1st Chernoff Bound

Show:

$$\boxed{e^{-t(1+c)pn} \cdot (1+p(e^t-1))^n \;\leq\; e^{-\frac{\min\{c,c^2\}}{3}pn}}$$

where: $t = \ln(1+c) > 0$

$$\begin{aligned}
e^{-t(1+c)pn} \cdot (1+p(e^t-1))^n &\leq e^{-t(1+c)pn} \cdot e^{pn(e^t-1)} \\
&= e^{-t(1+c)pn+pn(e^t-1)} \\
&= e^{-(1+c)\ln(1+c)pn+cpn} \\
&= e^{(c-(1+c)\ln(1+c))pn}
\end{aligned}$$

Next to show

$$\boxed{(1+c)\ln(1+c) \geq c + \tfrac{1}{3}\min\{c,c^2\}}$$

Proof of 1st Chernoff Bound

To show for c>1:

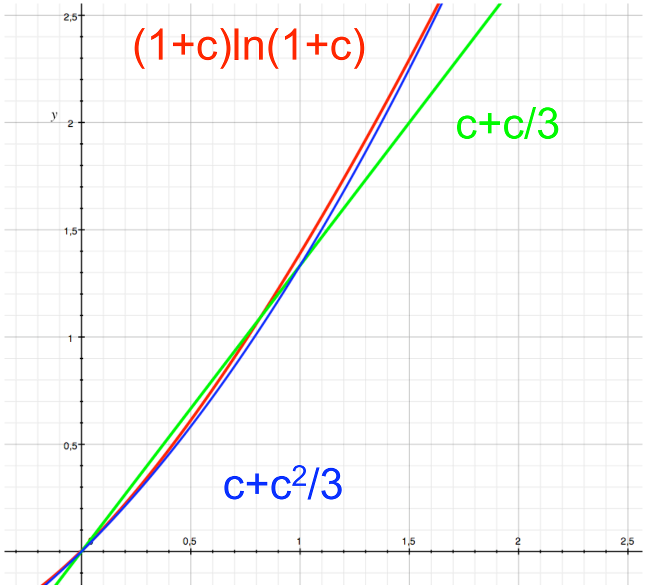
$$\boxed{(1+c)\ln(1+c) \geq c + \frac{1}{3}c}$$

For c=1: 2 ln(2) > 4/3

Derivative:

- left side: ln(1+c)
- right side: 4/3

- **For c>1 the left side is larger than the right side since**
  - ln(1+c)>ln (2) > 4/3
- **Hence the inequality is true for c>0.**



Proof of 1st Chernoff Bound

To show for c< 1: $\boxed{(1+c)\ln(1+c) \geq c + \frac{1}{3}c^2}$

For x>0:

$$\frac{d\ln(1+x)}{dx} = \frac{1}{1+x} = 1 - x + x^2 - x^3 + x^4 - \ldots$$

Hence

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots$$

By multiplication

$$(1+x)\ln(1+x) = x + \left(1 - \frac{1}{2}\right)x^2 - \left(\frac{1}{2} - \frac{1}{3}\right)x^3 + \left(\frac{1}{3} - \frac{1}{4}\right)x^4 - \ldots$$

Substitute (1+c) ln(1+c) which gives for c∈(0,1):

$$(1+c)\ln(1+c) \geq c + \frac{1}{2}c^2 - \frac{1}{6}c^3 \geq c + \frac{1}{3}c^2$$

Chernoff-Bound

- **Theorem Chernoff Bound**
  - Let $x_1,\ldots,x_n$ independent Bernoulli experiments with
    - $P[x_i = 1] = p$
    - $P[x_i = 0] = 1-p$
  - Let $S_n = \sum_{i=1}^{n} x_i$
  - Then for all c>0

$$\mathbf{P}[S_n \geq (1+c) \cdot \mathbf{E}[S_n]] \quad \leq \quad e^{-\frac{1}{3}\min\{c,c^2\}pn}$$

  - For 0≤c≤1

$$\mathbf{P}[S_n \leq (1-c) \cdot \mathbf{E}[S_n]] \quad \leq \quad e^{-\frac{1}{2}c^2pn}$$

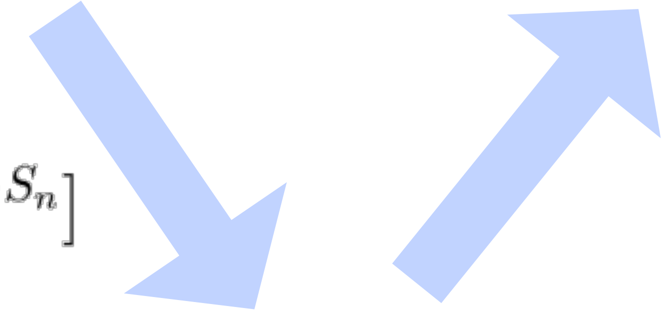
Proof of 2nd Chernoff Bound

- **We show** $\mathrm{P}[S_n \leq (1-c)\mathbf{E}[S_n]] \ \leq \ e^{-\frac{c^2}{2}pn} \ .$
- **For t<0:** $\mathrm{P}[S_n \leq (1-c)pn] \ = \ \mathrm{P}[e^{tS_n} \geq e^{t(1-c)pn}]$

  t<0

  t<0

$$\frac{1}{k} \leq \ e^{-\frac{c^2}{2}pn}$$



$$k = e^{t(1-c)pn} / \mathbf{E}[e^{t \cdot S_n}]$$

- **Markov yields:** $\mathrm{P}\left[e^{tS_n} \geq k\mathbf{E}\left[e^{tS_n}\right]\right] \ \leq \ \frac{1}{k}$

- **To do: Choose t appropriately**

Proof of 2nd Chernoff Bound

- **We show** $\boxed{\frac{1}{k} \leq e^{-\frac{c^2}{2}pn}}$

- **where** $k = e^{t(1-c)pn} / \mathbf{E}[e^{t \cdot S_n}]$

$$\boxed{\begin{aligned}
\mathbf{E}[e^{tS_n}] &= \mathbf{E}\left[e^{t\sum_{i=1}^{n} x_i}\right] \\
&= \mathbf{E}\left[\prod_{i=1}^{n} e^{tx_i}\right] \\
&= \prod_{i=1}^{n} \mathbf{E}\left[e^{tx_i}\right] \\
&= \prod_{i=1}^{n} (e^0(1-p) + e^t p) \\
&= (1 - p + e^t p)^n \\
&= (1 + (e^t - 1)p)^n
\end{aligned}}$$

Independence of random variables x_i $\longrightarrow$

- **Next we show:**

$$\boxed{e^{-t(1-c)pn} \cdot (1 + p(e^t - 1))^n \leq e^{-\frac{c^2}{2}pn}}$$

Proof of 2nd Chernoff Bound

We show

$$\boxed{e^{-t(1-c)pn} \cdot (1 + p(e^t - 1))^n \ \leq \ e^{-\frac{c^2}{2}pn}}$$

where:

$$t = \ln(1-c)$$

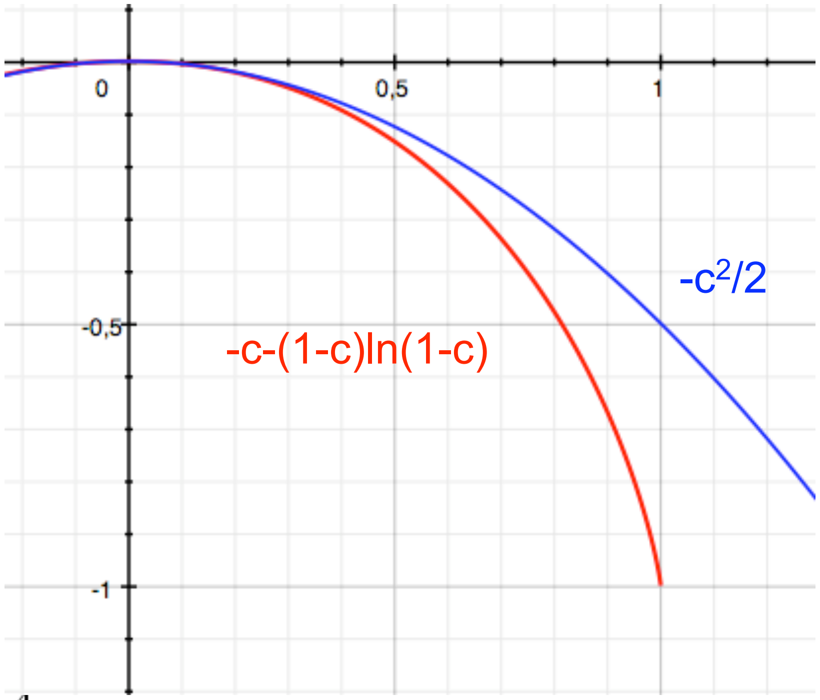
1+x ≤ e^x

$$\begin{aligned} e^{-t(1-c)pn} \cdot (1 + p(e^t - 1))^n &\leq e^{-t(1-c)pn} \cdot e^{pn(e^t-1)} \\ &= e^{-t(1-c)pn+pn(e^t-1)} \\ &= e^{-(1-c)\ln(1-c)pn-cpn} \end{aligned}$$

Next to show

$$\boxed{-c - (1-c)\ln(1-c) \leq -\frac{1}{2}c^2}$$

Proof of 2nd Chernoff Bound



To prove:

$$-c-(1-c)\ln(1-c) \leq -\frac{1}{2}c^2$$

For c=0 we have equality

Derivative of left side: ln(1-c)

Derivative of right side: -c

Now

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \frac{1}{4}x^4 + \ldots$$

This implies

$$\ln(1-c) = -c - \frac{1}{2}c^2 - \frac{1}{3}c^3 - \ldots < -c$$

Proof ctd.

$$-c - (1-c)\left(-c - \frac{1}{2}c^2 - \frac{1}{3}c^3 - \dots\right)$$

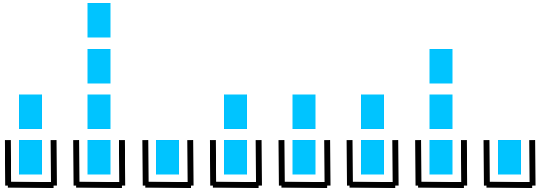
$$-\overset{\checkmark}{c} + \overset{\checkmark}{c} + \frac{1}{2}c^2 + \frac{1}{3}c^3 + \frac{1}{4}c^4 + \frac{1}{5}c^5 \dots$$

$$-c^2 - \frac{1}{2}c^3 - \frac{1}{3}c^4 - \frac{1}{4}c^5 \dots$$

$$= -\frac{1}{2}c^2 - \left(\frac{1}{2} - \frac{1}{3}\right)c^3 - \left(\frac{1}{3} - \frac{1}{4}\right)c^4 - \left(\frac{1}{4} - \frac{1}{5}\right)c^5 \dots$$

$$< -\frac{1}{2}c^2$$

Balls and Bins



Lemma

If m= k n ln n Balls are randomly placed in n bins:

1. Then for all c>k the probability that more than c ln n balls are in a bin is at most O(n$^{-c'}$) for a constant c'>0.
2. Then for all c<k the probability that less than c ln n balls are in a bin is at most O(n$^{-c'}$) for a constant c'>0.

Proof:

Consider a bin and the Bernoulli experiment B(k n ln n,1/n) and expectation: μ = m/n = k ln n

1. Case: c>2k
$$\mathrm{P}[X \geq c \ln n] = \mathrm{P}[X \geq (1+(c/k-1))k \ln n]$$
$$\leq e^{-\frac{1}{3}(c/k-1)k \ln n} \leq n^{-\frac{1}{3}(c-k)}$$
2. Case: k<c<2k
$$\mathrm{P}[X \geq c \ln n] = \mathrm{P}[X \geq (1+(c/k-1))k \ln n]$$
$$\leq e^{-\frac{1}{3}(c/k-1)^2 k \ln n} \leq n^{-\frac{1}{3}(c-k)^2}$$
3. Case: c<k
$$\mathrm{P}[X \leq c \ln n] = \mathrm{P}[X \leq (1-(1-c/k))k \ln n]$$
$$\leq e^{-\frac{1}{2}(1-c/k)^2 k \ln n} \leq n^{-\frac{1}{2}(k-c)^2/k}$$

Concept of High Probability

Lemma

If A(i) holds with **high** probability, i.e. $1-n^{-c}$, then

(A(1) and A(2) and ... and A(n)) with **high** probability,
i.e. $1-n^{-(c-1)}$

Proof:

- For all i: $P[\neg A(i)] \leq n^{-c}$
- Hence: $P[\neg A(1) \text{ or } \neg A(2) \text{ or } \dots \neg A(n)] \leq n \cdot n^{-c}$

$$P[\neg(\neg A(1) \text{ or } \neg A(2) \text{ or } \dots \neg A(n))] \leq 1- n \cdot n^{-c}$$

DeMorgan:

$$P[A(1) \text{ and } A(2) \text{ and } \dots A(n)] \leq 1- n \cdot n^{-c}$$

Principle of Multiple Choice

- **Before inserted check c log n positions**
- **For position p(j) check the distance a(j) between potential left and right neighbor**
- **Insert element at position p(j) in the middle between left and right neighbor, where a(j) was the maximum choice**
- **Lemma**
  - After inserting n elements with high probability only intervals of size 1/(2n), 1/n und 2/n occur.

Proof of Lemma

1. Part: With high probability there is no interval of size larger than 2/n

follows from this Lemma

Lemma*

Let c/n be the largest interval. After inserting 2n/c peers all intervals are smaller than c/(2n) with high probability

From applying this lemma for c=n/2,n/4, ...,4 the first lemma follows.

Proof

- **2nd part: No intervals smaller than 1/(2n) occur**
  - The overall length of intervals of size 1/(2n) before inserting is at most 1/2
  - Such an area is hit with probability at most 1/2
  - The probability to hit this area more than c log n times is at least

$$2^{-c \log n} = n^{-c}$$

  - Then for c>1 such an interval will not further be divided with probability into an interval of size 1/(4m).

ALBERT-LUDWIGS-
UNIVERSITÄT FREIBURG

Algorithms and Methods for Distributed Storage Networks

9 Analysis of DHT

Christian Schindelhauer

Albert-Ludwigs-Universität Freiburg
Institut für Informatik
Rechnernetze und Telematik
Wintersemester 2007/08

CoNe
Freiburg

IIF
INSTITUT FÜR
INFORMATIK
FREIBURG