CS-GY 9223I (CS-UY 3943)

Algorithmic Machine Learning + Data Science

Instructor: Prof. Christopher Musco (cmusco@nyu.edu)

Office Hours: 3-5pm Thursdays (except tomorrow)

Reading Group: TBD. Poll to gauge interest and time.

Webpage: chrismusco.com/9223_2019

NYU Classes for forum

Exams: Midterm - 10/23, in class

Final - 12/18, class time

check for conflicts now!

Course Topic: Algorithmic methods for machine learning + data analysis at scale.

- High throughput / realtime data applications

(think Shazam, Google maps / waze, Amazon product recs, industrial robotics, FinTech, scientific applications)

- More complex models → more training data

(deep neural networks, reinforcement learning, machine translation)

- Data analysis on low compute devices

(smart phones/watches, robots/drones/etc., sensor networks)

Some numbers:

- Google receives ≈ 10,000 Maps queries <u>every second</u>
- NASA collects 6.4 Tb of satallite images <u>every day</u>
  - new "ImageNet" dataset every 3 days
- Large Synoptic Survey Telescope: 15 Tb of images <u>per night</u>
- Broad Institute sequences 24 Tb of genetic data per day

"needle in a haystack problems"

Ushering in new golden age for research in computational methods, using entirely new tools!

Course Objectives:

1 New algorithmic toolkit (randomization, sketching, optimization, spectral methods, etc.)

- lectures
- reading

See cs.princeton.edu/courses/archive/fall18/cos521

2. Learn to apply tools in the wild (industry, academia)

- 4 Problem Sets
- in class work
  - → midway into class
  - → break into groups (auditors included)
  - → pset like problem to solve

3. Theory as an approach to algorithm design.

What we won't cover:

- Software tools or frameworks

  (MapReduce, Tensorflow, Amazon AWS)

  (CS-GY 6513: Big Data)

- Machine learning models

  (neural nets, RL, Bayesian methods, unsupervised learning, function fitting, etc.)

Unit 1: The Power of Randomness

Hashing (+ Load Balancing)

- Work horse of the modern web
- good probability review!

Probability review:

X takes values in set $S \subseteq \mathbb{R}$. Eg. $S = \{1, 2, 3, 4, 5, 6\}$ for dice roll.

Expectation: $\mathbb{E}[X] = \sum_{\omega \in S} \Pr[X = \omega] \cdot \omega$

Continuous r.v.s : $\mathbb{E}[X] = \int_{y \in \mathbb{R}} p(y)\, dy$

Independence Two random events A, B are independent if $Pr(A \mid B) = Pr(A)$

"A given B"

$$Pr(A \mid B) \stackrel{\text{def}}{=} \frac{Pr(A \cap B)}{P(B)}$$

So equivalently, when A and B are independent,

$$\frac{Pr(A \cap B)}{Pr(B)} = Pr(A) \longrightarrow Pr(A \cap B) = Pr(A) \cdot Pr(B)$$

Roll 2 dice. What's the probability the first is odd and the second is < 3?

Independence Given random variables X and Y taking values in S_x and S_y. We say X and Y are independent if for all $w \in S_x$, $z \in S_y$

$[X = w]$ and $[Y = z]$ are independent random events.

Expectation Identities

$\mathbb{E}[aX] = a\,\mathbb{E}[X]$

$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$?

$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$?

$\mathbb{E}[X+Y]$ is true for <u>any</u> X, Y.

$$= \sum_{w \in S_X} \sum_{z \in S_Y} \Pr(X=w, Y=z) \cdot (w+z)$$

$$= \sum_{w} \sum_{z} \Pr(X=w, Y=z) \cdot w + \sum_{w} \sum_{z} \Pr(X=w, Y=z) \cdot z$$

$$= \sum_{w} w \cdot \underbrace{\sum_{z} \Pr(X=w, Y=z)}_{=\Pr(X=w)} + \sum_{z} z \cdot \underbrace{\sum_{w} \Pr(X=w, Y=z)}_{=\Pr(Y=z)}$$

$$= \mathbb{E}[X] + \mathbb{E}[Y]$$

$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$ true for independent r.v.'s.

$$\mathbb{E}[XY] = \sum_{\omega \in S_X} \sum_{z \in S_Y} \Pr(X = \omega, Y = z)\, \omega z$$

$$= \sum_{\omega} \sum_{z} \Pr(X = \omega) \Pr(Y = z)\, \omega \cdot z$$

$$= \sum_{\omega} \left[\omega \cdot \Pr(X = \omega) \cdot \sum_{z} \Pr(Y = z) \cdot z \right]$$

$$= \left[\sum_{\omega} \omega \cdot \Pr(X = \omega) \right] \cdot \left[\sum_{z} \Pr(Y = z) \cdot z \right]$$

$$= \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y] \Rightarrow X, Y$ are "uncorrelated".

Independence $\Rightarrow$ Uncorelated. Uncorelated $\not\Rightarrow$ Independence.

Markov's ^least exciting^ Inequality

For a non-negative random variable X, $\Pr[X \geq a] \leq \mathbb{E}[X]/a$

Equivalent: $\Pr[X \geq c \cdot \mathbb{E}[X]] \leq 1/c$. think $c = 2, 10, \ldots$

"concentration inequality"

Proof: $\mathbb{E}[X] = \sum_{\omega} \Pr[X=\omega] \cdot \omega$

$$= \underbrace{\sum_{\omega < a} \Pr[X=\omega] \cdot \omega}_{\geq 0} + \underbrace{\sum_{\omega \geq a} \Pr[X=\omega] \cdot \omega}$$

$$\geq \sum_{\omega \geq a} \Pr[X=\omega] \cdot a$$

$$= a \cdot \sum_{\omega \geq a} \Pr[X=\omega]$$

$$= a \cdot \Pr[X \geq a]$$

$$\mathbb{E}[X] \geq a \cdot \Pr[X \geq a]$$

Is Markov's Inequality Tight?

Can you prove something better?

In general, Markov's <u>is</u> tight.

$X = 0$ with probability $1 - \frac{t}{a}$

$= a$ with probability t/a.

$\mathbb{E}[X] = 0 \cdot \left(1 - \frac{t}{a}\right) + a \cdot (t/a) = t.$

$\mathbb{P}[X \geq a] = t/a = \mathbb{E}[X]/a.$

Hashing (+ Load Balancing)

- Work horse of the modern web

- good probability review!

(key, value) store

"nyu.edu" 216.165.47.10

3 operations: insert (k_1, v_1)

insert (k_2, v_2)

⋮

insert (k_m, v_m)

delete (k_{10}, v_{10})

⋮

query $(k_j) \to v_j$

query $(k_{10}) \to$ empty

↓

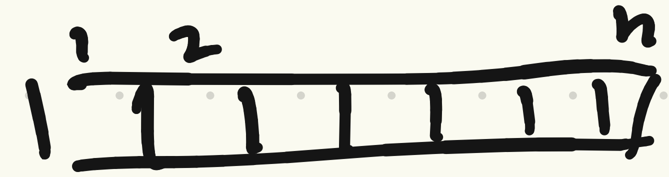
Want

1) Fast queries: $O(1)$ time

2) Small space: $O(m)$ space

Hashing

- Build table T of size n



- Choose **random** function $h : U \to \{1, \ldots, n\}$

universe of possible keys

$k_1, \ldots, k_m \in U$

U	h
u_1	$h(u_1) = 10$
u_2	$h(u_2) = 4$
u_3	$h(u_3) = m$
$\vdots$	$\vdots$
$u_{1000000}$	$h(u_{100}) = 4$

h drawn uniform from $\mathcal{H}$

all possible mappings from $U \to \{1, \ldots, n\}$

"hash family"

Is this possible in practice?

3 Issues!

$(k \cdot r_1 + r_2) \pmod{m}$

Hashing

- for insert(k, v), store v at $T_{h(k)}$
- for delete(k, v), remove v from $T_{h(k)}$
- for query(k), look at $T_{h(k)}$

$m \ll |U|$ so could have $\underbrace{h(k) = h(j) \text{ for } j \neq k.}_{\text{"hash collision"}}$

Store $v(k), v(j)$ in linked list at $T_{h(k)} = T_{h(j)}$

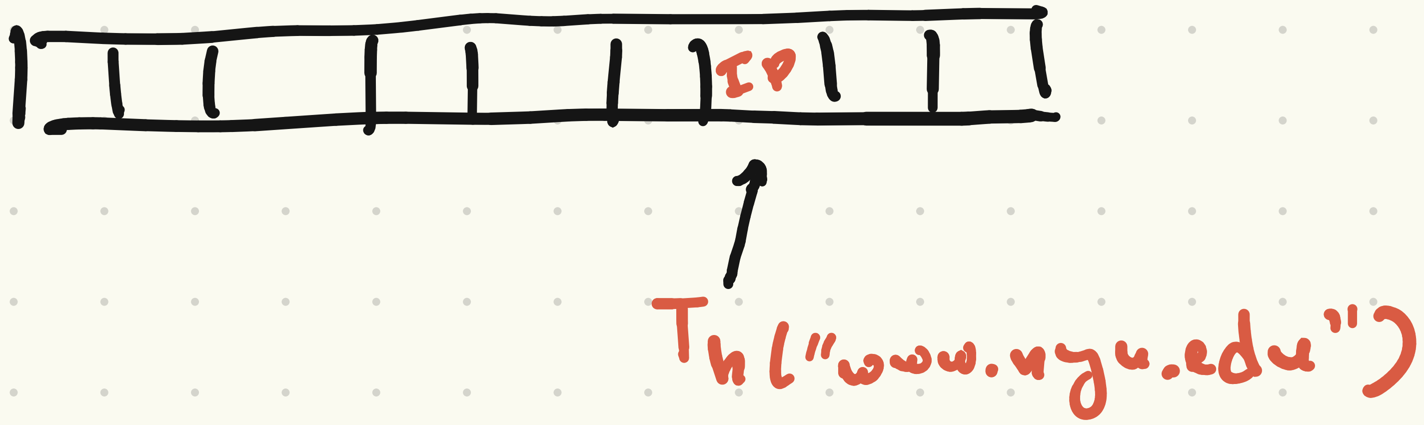
Goal: Hash collisions should be rare!

Lookup time $= O(1)$ if no collision,

otherwise O(length of linked list)

Hashing Applications

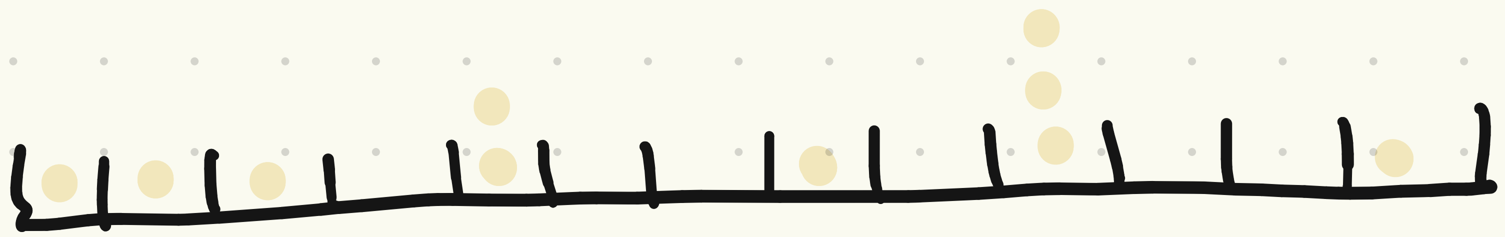
- URL / DNS resolution. What IP address to visit to see www.nyu.edu?



- Web content delivery (distributed hash table)
- Amazon DynamoDB, MongoDB, Cassandra, Google Datastore
  "no-SQL" databases
- Google directions: send "boston to NYC" to $A_{h(\text{"boston to NYC"})}$
  servers $A_1, \dots, A_8$

Goal: Hash collisions should be rare!

How many collisions in <u>expectation</u> when inserting $\underline{m}$ items into table of size $\underline{n}$?



$\underbrace{\text{\# of collisions}}_{C}$ = random variable, with randomness coming from choice of $h \in \mathcal{H}$.

take $k_1, \ldots, k_m$ as fixed.

$$C = \frac{1}{2} \sum_{i=1,\ldots,m} \sum_{j \neq i} \mathbb{1}[h(k_i) = h(k_j)]$$

$\mathbb{1}[\text{true}] = 1$

$\mathbb{1}[\text{false}] = 0$

"indicator function"

How many collisions in <u>expectation</u> when inserting <u>m</u> items into table of size <u>n</u>?

$$\mathbb{E}[C] = \mathbb{E}\left[\frac{1}{2}\sum_i \sum_{j \neq i} \mathbb{1}[h(k_i) = h(k_j)]\right]$$

$$= \frac{1}{2}\sum_i \sum_{j \neq i} \underbrace{\mathbb{E}[\mathbb{1}[h(k_i) = h(k_j)]]}_{= 1/n}$$

$$= \frac{1}{2}\sum_i \sum_{j \neq i} 1/n = \boxed{\frac{m \cdot m-1}{2n}}$$

O(1) time lookups.

Result 1 Collision free hash table with $O(m^2)$ space.

Set $n = 5m^2$. $\mathbb{E}[C] = \frac{m(m-1)}{2n} \leq \frac{1}{10}$. $\underbrace{\Pr[C \geq 1] \leq 1/10}_{\text{Markov's inequality}}$.

Could keep retrying until achieve collision free hash.

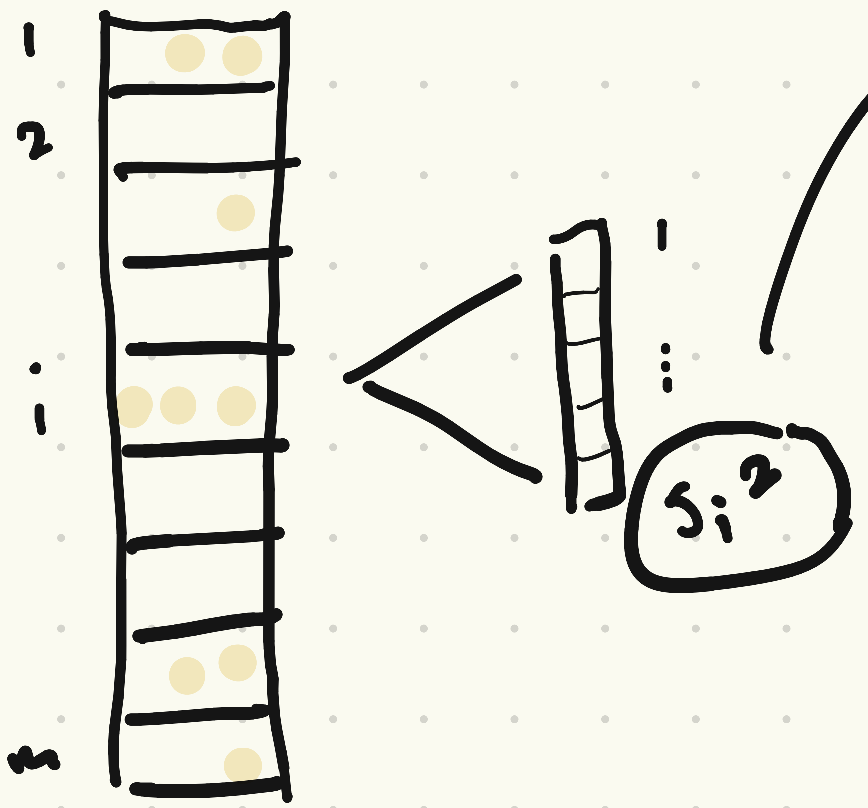
Trials:	1	2	3	4	....	6
Failure Probability:	1/10	1/100	1/1000	1/10000		≤ chance getting struck by lightening.

$O(m^2)$ is a lot of space overhead for m items.

"Birthday Paradox"

Result 2   Collision free hash table with $O(n)$ space.

Key Idea: 2 Layer Hashing



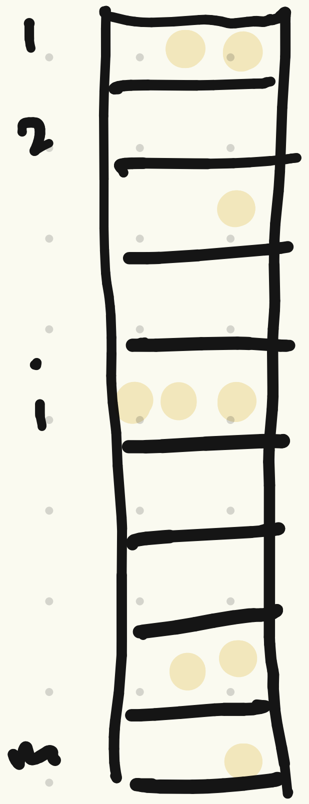
2nd level is collision free → each lookup takes just 2 hash function evaluations.

$$\text{Total space} = m + \sum_{i=1}^{m} s_i^2$$

$$\mathbb{E}\left[m + \sum_{i=1}^{m} s_i^2\right] = m + \underbrace{\mathbb{E}\sum_{i=1}^{m} s_i^2}_{?}$$

items in bin $i = s_i$

$$\mathbb{E}[\text{total space}] = m + \mathbb{E}\left[\sum_{i=1}^{m} s_i^2\right]$$



→ # collisions in bucket $i = s_i(s_i-1)$

total # of collisions

$$= \boxed{C = \frac{1}{2}\sum_{i=1}^{m} s_i(s_i-1)}$$

→ almost s_i^2!

Almost instant lookups with only 3x space overhead!

→ Markov's inequality is actually quite powerful!

$$\mathbb{E}\sum_{i=1}^{m} s_i^2 = \mathbb{E}\left[\sum_{i=1}^{m} s_i(s_i-1) + \sum_{i=1}^{m} s_i\right]$$

$$= 2\,\underbrace{\mathbb{E}[C]}_{} + m \qquad < 2m$$

with m buckets, m items

$$= \frac{m(m-1)}{2m} \le 1/2$$

$$\boxed{\mathbb{E}[\text{total space}] = 3m}$$

- Google directions: send "boston to NYC" to $A_{h(\text{"boston to NYC"})}$

$A_1, A_2, \ldots, A_n$

$\boxed{\text{why?}}$

- Google directions: send "boston to NYC" to $A_{h(\text{"boston to NYC"})}$

$A_1, A_2, \ldots, A_n$

Now we care about the maximum # of elements per bin.

"Load Balancing" $\max[s_1, \ldots, s_n]$

Suppose we hash n items to n servers. "Balls into Bins"

$\to b_1, \ldots, b_n$

What does Markovs give?

$$\mathbb{E}[s_i] = \mathbb{E}\left[\sum_{j=1}^{n} \mathbb{1}[h(b_j) = i]\right] = \sum_{j=1}^{n} \frac{1}{n} = 1.$$

$$\boxed{\Pr[s_i \geq 10] \leq 1/10}$$

1 in 10 servers could be overloaded! No bound on max.

Goal: For any i, $\Pr[s_i \geq B] \leq \frac{1}{10n}$.

Corollary: With probability 9/10 all $s_i \leq B$.

Proof: Union bound: Given rand. on events A, B

$$\underbrace{\Pr[A \text{ or } B \text{ occur}]}_{} \leq \Pr[A] + \Pr[B]$$

$$= \Pr[A + \bar{B}] + \Pr[\bar{A} + B] + \Pr[A + B]$$

$$= \Pr[A + \bar{B}] + \Pr[B] \leq \Pr[A] + \Pr[B]$$

$$\Pr[s_1 \geq B \text{ or } s_2 \geq B \text{ or} \ldots s_n \geq B] \leq \frac{1}{10n} + \frac{1}{10n} + \cdots + \frac{1}{10n} = \frac{1}{10}.$$

Goal: For any i, $\Pr[s_i \geq B] \leq \frac{1}{10n}$.

Markovs: $B = 10n$. $\longrightarrow$ Vacuous bound! Only n items...

Tighter bounds on deviation by considering more information about a random variable.

Variance: For r.v. X, $Var(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$

Equivalent: $Var(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

Proof: $\mathbb{E}[(X - \mathbb{E}X)^2] = \mathbb{E}[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2]$

$$\mathrm{Var}(\alpha X) = \alpha^2 \, \mathrm{var}(X)$$

Linearity of Variance: $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$ if X, Y independent

$$\mathrm{Var}[X_1 + X_2 + \dots + X_m] = \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \dots + \mathrm{Var}[X_m]$$

if $\underbrace{X_i, X_j \text{ independent for all } i,j \in 1, \dots, m}_{\text{"pairwise independence"}}$.

"mutual independence" is stronger. why?

Chebyshev Inequality: Let X be a r.v. with $var(X) = \sigma^2$.

$$\Pr[|X - \mathbb{E}X| \geq k\sigma] \leq 1/k^2$$

upper and lower bound — no assumption that $X > 0$.

Proof. Consider the random variable $Y = (X - \mathbb{E}X)^2$.

Y is nonnegative. By Markov inequality,

$$\Pr[Y \geq k^2 \mathbb{E}[Y]] \leq 1/k^2$$

$$\mathbb{E}[Y] = \mathbb{E}[X - \mathbb{E}X] = var(X) = \sigma^2.$$

$$\Pr[(X - \mathbb{E}X)^2 \geq k^2\sigma^2] \leq 1/k^2$$

Goal: For <u>any</u> i, $\Pr[S_i \geq B] \leq \frac{1}{10n}$

S_i → # of balls to bin i

$$S_i = \sum_{j=1}^{n} \mathbb{1}[h(b_j) = i] \qquad \text{Var}[S_i] = \sum_{j=1}^{n} \text{Var}\, \underbrace{\mathbb{1}[h(b_j) = i]}_{X_j}$$

$$\text{Var}(X_j) = \mathbb{E}[X_j^2] - \mathbb{E}[X_j]^2 = \boxed{\frac{1}{n} - \frac{1}{n^2}}$$

$$X_j = \begin{cases} 1 & \text{w/ prob. } 1/n \\ 0 & \text{otherwise} \end{cases} \qquad X_j^2 = \begin{cases} 1 & \text{w/ prob } 1/n \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}[X_j] = 1/n \qquad \mathbb{E}[X_j^2] = 1/n$$

$$\text{Var}[S_i] = n\left(\frac{1}{n} - \frac{1}{n^2}\right) = 1 - \frac{1}{n} \leq 1.$$

Goal: For any i, $Pr[S_i \geq B] \leq \frac{1}{10n}$.

Chebyshevs: $Pr[|X - \mathbb{E}X| \geq k\sigma] \leq 1/k^2$

$Var[S_i] = 1 - 1/n$.

$\mathbb{E}[S_i] = 1$

Improvable to $O(\log n)$!

Set $k = \sqrt{10n}$.

$Pr[|S_i - 1| \geq \sqrt{10n} \cdot \sqrt{1 - 1/n}] \leq \frac{1}{10n}$

$Pr[S_i - 1 \geq \sqrt{10n} \cdot 1] \leq 1/10]$

With probability 9/10 all bins have $\leq O(\sqrt{n})$ balls.

$$\boxed{Pr[S_i \geq O(\sqrt{n})] \leq 1/10n}$$

$n = 1{,}000{,}000$

max load ≈ 1000.

In class exercise

New York University Tandon School of Engineering
Computer Science and Engineering

CS-GY 9223I: Lecture 1 Coursework

Problem 1: Hash collisions are useful?

Your company is considering paying for a cloud service that provides CAPTCHA-like visual puzzles for verifying that users are human. The company providing the service claims to have a larger database of unique puzzles than any competitors, but you don't trust the salesperson.

(a) The company has provided you with an API end-point which returns puzzles uniformly and independently at random from their database. Using this endpoint, describe a simple randomized estimator for the number of puzzles in the database, n.

(b) The company claims their database has $1,000,000$ unique CAPTCHAs in it. Using your estimator, roughly how many queries do you need to verify their claim with good probability (e.g. $9/10$)? You should need far less than $1,000,000$ queries!

(c) More generally, how many samples are required to estimate the true number of CAPTCHAs, n, in the database up to additive error $\pm\epsilon n$, with good probability?