RECAP

LAST TIME: we bounded the depth of the recursion tree to $8 \log_{4/3} n$ w.h.p., thus

$$\Pr(T_{RQS}(n) > 8n \log_{4/3} n) < 1/n^2$$

MAIN IDEAS:

- A random pivot guarantees a "reasonably good" balance, on average:
  $$\Pr(|S_1|, |S_2| \leq (3/4)|S|) \geq 1/2$$
- Rephrase analysis as deviation from average behavior along fixed path of the recursion tree using a Chernoff Bound
- Extend to any path via Union Bound
- Analysis w.h.p. also implies:
  $$E[T_{RQS}(n)] = O(n \log n)$$

(using Markov's bound)

QUICKSORT : ANALYSIS IN EXPECTATION

We can use Markov's lemma for bounded variables to obtain an upper bound to $E[T] = E[T_{RQS}(n)]$:

$$\forall t > 0 \quad E[T] < t + (b-t)\,\Pr(T > t)$$

where $\Pr(T \le b) = 1$

Clearly we can set $b = n^2$ and $t = 8n\log_{4/3} n$:

Then

$$E[T] < 8n\log_{\frac{4}{3}} n + (\underbrace{n^2}_{< n^2} - 8n\log_{\frac{4}{3}} n)\,\Pr(T > 8n\log_{\frac{4}{3}} n)$$

$$< 8n\log_{\frac{4}{3}} n + \frac{n^2}{n^2} < 8n\log_{\frac{4}{3}} n + 1$$

$$\Rightarrow E[T_{RQS}(n)] = O(n\log n)$$

An almost exact analysis of $E[T_{RQS}(n)]$ yields

$$E[T_{RQS}(n)] \le 2nH(n) =$$

$$\le 2n\ln n + O(n)$$

$$\approx 1.39\, n\log_2 n$$

See online material

PROBABILITY AMPLIFICATION: ROBUSTNESS

Another application of Chernoff Bound concerns the amplification of the correctness probability of a Monte Carlo decision algorithm – giusto a piacere con proprietà base di risposte

Let A_π be two-sided error (most general case) and let $\Pr(i \,\Pi\, A_\pi(i)) \geq \frac{1}{2} + \varepsilon$, $0 < \varepsilon \leq \frac{1}{2}$. Thus on both answers, the algorithm is correct with probability = $\frac{1}{2} + \varepsilon > \frac{1}{2}$

REMARK Observe that a coin flip is sufficient to achieve $\Pr(A_\pi(i) \,\Pi\, i) = \frac{1}{2}$!

Let us see how to design a more robust algorithm via multiple executions of A_π:
We implement a majority protocol:

```
ROBUST_Aπ(i, K)
  count[0] ← count[1] ← 0
  for j ← 1 to 2K-1 do
     d ← Aπ(i)
     count[d] ← count[d] + 1
  if (count[0] > count[1])
     then return 0
     else return 1
```

REMARK: Since $2k-1$ is odd, we have:
$(count[0] > count[1]) \vee (count[1] > count[0])$

Clearly $\boxed{T_{k\text{-}A_\pi}(n) = \Theta(k \cdot T_{A_\pi}(n))}$

Let us study $Pr(i \not\pi R_{A_\pi}(i))$

Let $\boxed{X_1, \dots, X_{2k-1}}$ indicator variables with

$$\boxed{X_i = \begin{cases} 1 & \text{if the } i\text{-th execution of } A_\pi(i) \text{ in } k\text{-}A_\pi(i) \text{ is correct } (i \,\pi\, A_\pi(i)) \\ 0 & \text{otherwise } (i \not\pi A_\pi(i)) \end{cases}}$$

The X_i's are i.i.d. Bernoulli trials with

$$\boxed{Pr(X_i = 1) \geq \frac{1}{2} + \varepsilon}$$

Let $\boxed{X = \sum_{i=1}^{2k-1} X_i}$. The majority protocol $k\text{-}A_\pi(i)$ is incorrect if and only if $X < k$

We want to express $Pr(X < k)$ as $Pr(X < (1-\delta)\mu)$, with $\mu = E[X] \geq (2k-1)(\frac{1}{2}+\varepsilon)$

Let us assume that $\boxed{\mu = (2k-1)(\frac{1}{2}+\varepsilon)}$

(for higher values of μ, $Pr(X < (1-\delta)\mu)$ is even smaller)

We have:

$$\mu = (2k-1)\left(\frac{1}{2}+\varepsilon\right) = k - \frac{1}{2} + 2k\varepsilon - \varepsilon$$

$$\Rightarrow k = \mu + \frac{1}{2} + \varepsilon - 2k\varepsilon$$

Now, we prove that $\boxed{\mu\varepsilon < 2k\varepsilon}$:

$$\mu\varepsilon = \varepsilon(2k-1)\left(\frac{1}{2}+\varepsilon\right) = (2k-1)\left(\frac{\varepsilon}{2}+\varepsilon^2\right)$$

$$< 2k\left(\frac{\varepsilon}{2}+\frac{\varepsilon}{2}\right) = 2k\varepsilon$$

$$\left(\varepsilon^2 \le \frac{\varepsilon}{2} \text{ for } \varepsilon \le \frac{1}{2}\right)$$

Therefore

$$k < \mu + \frac{1}{2} + \varepsilon - \mu\varepsilon \le (1-\varepsilon)\mu + 1$$

($\varepsilon - \mu\varepsilon \le 1$)

We have:

$$Pr(X < k) = Pr(X \le k-1) \le Pr(X < (1-\varepsilon)\mu)$$

$$< e^{-\frac{\varepsilon^2}{2}\mu} \le e^{-\frac{\varepsilon^2}{4}k}$$

(Chernoff 3)

(since $\mu \ge (2k-1)(1/2+\varepsilon) > (2k-1)(1/2) = k - 1/2 \ge k/2$ $\forall k > 0$)

Consider now $n = |i|$. by setting

$$k = \frac{4}{\varepsilon^2} d \ln n, \quad d > 1$$

we obtain

$$Pr(i \neq R\text{-}A_\pi(i)) = Pr(X < k)$$

$$< e^{-\frac{\varepsilon^2}{4}\cdot\frac{4}{\varepsilon^2} d \ln n} < \frac{1}{n^d}$$

Thus $O\left(\frac{\log n}{\varepsilon^2}\right)$ executions of A_π guarantee correctness w.h.p.

REMARK: Correctness amplifies proportionally to ε^{-2}. "Imprecise" algorithms need many iterations to be made robust.

REMARK 2: For one-sided error Monte Carlo algorithms we can amplify correctness probability even when

$$Pr(i \Pi A_\Pi(i)) \ll \frac{1}{2} !$$

(indeed, $Pr(i \Pi A_\Pi(i)) = \varepsilon$ suffices)

Assume that:

$$Pr(i \Pi A_\Pi(i)) \geq \begin{cases} 1 & \text{if } i \Pi 1 \\ \varepsilon & \text{if } i \Pi 0 \end{cases}$$

(in MR(n) this is the case with $\varepsilon = \frac{1}{2}$ and 1 = "prime", 0 = "composite".)

We write

```
REQUEST_A_Π (i, s)
  for i ← 1 to s do
    if A_Π(i) = 0
      then return 0
  return 1
```

Clearly, if $i \Pi 1$, then

$$Pr(\text{REQUEST\_}A_\Pi(i) = 1) = 1$$

(since $\Pr(A_\pi(i) = 1) = 1$)

If $i \neq 0$: $i \notin A_\pi(i)$ for s times

$$\Pr(R_{A_\pi}(i) = 1) \leq (1-\varepsilon)^{\frac{s}{\varepsilon}\varepsilon} < e^{-s\varepsilon}$$

(since $(1-\varepsilon)^{1/\varepsilon} < e^{-1}$)

For $|i| = n$ and $s = d\frac{\ln n}{\varepsilon}$ we obtain:

$$i \neq 0 : \Pr(R_{A_\pi}(i) = 1) < e^{-\frac{d\ln n}{\varepsilon}\cdot\varepsilon} = \frac{1}{n^d}$$

Thus $\Pr(R\_A_\pi \text{ is incorrect}) \leq$

$$\Pr(R_{A_\pi}(i) = 1 \text{ for } i \neq 0) < \frac{1}{n^d}$$

Number of runs: $O\left(\frac{\log n}{\varepsilon}\right)$

↓

matching the lower bound for two-sided error

MONTE CARLO METHOD

Through Chernoff's bounds, sampling may be used to obtain estimates of unknown quantities related to a probability distribution.

EXIT POLLS Typical application: estimating the fraction of votes of a given party in an election

Model: Urn U containing n white or black balls. We want to estimate the fraction αn of white balls without examining all balls in U. We assume that:

1. we have a deterministic lower bound $\alpha_{min} \leq \alpha$ (e.g. we are certain that a given party cannot get $< \alpha_{min} \cdot n$ votes)
2. Our primitive is `b ← RANDOM(U)` (interview random voter)

Given a precision threshold ε, and a confidence δ, we want to show

a randomized estimate $\bar{\alpha}$ of α such that:

relative error

$$Pr\left(\frac{|\bar{\alpha}-\alpha|}{\alpha} > \varepsilon\right) < \delta$$

That is, the probability that the relative error of the estimate $\bar{\alpha}$ over α is greater than ε is less than δ

If this is the case, $\bar{\alpha}$ is said to be an (ε, δ)-approximation of α.

APPROACH: I make a number of calls to RANDOM(U) and return the sample average of the number of white balls

```
APPROX(ε, δ, U)
  k ← f(α_min, ε, δ) { # samples (analysis) }
  X ← 0
  for i ← 1 to k do
      b ← RANDOM(U)
      if (b.color = white)
          then X ← X+1
  return ᾱ = X/k { sample average }
```

Let us set $k : Pr\left(\frac{|\bar{\alpha}-\alpha|}{\alpha} > \varepsilon\right) < \delta$

Consider k indicator variables

$$x_i = \begin{cases} 1 & \text{if } b_i.\text{color} = \text{white} \\ 0 & \text{otherwise} \end{cases}$$

We have $Pr(x_i = 1) = \alpha$ and $X = \sum_{i=1}^{k} x_i$

thus $\mu = E[X] = k\alpha$ and $\bar{\alpha} = X/k$

We have that

$$Pr\left(\frac{|\bar{\alpha} - \alpha|}{\alpha} > \varepsilon\right) = Pr\left(\frac{|X/k - \alpha|}{\alpha} > \varepsilon\right) \quad \begin{matrix} \times k \\ \times k \end{matrix}$$

$$= Pr\left(\frac{|X - \alpha k|}{\alpha k} > \varepsilon\right) = Pr\left(\frac{|X - \mu|}{\mu} > \varepsilon\right)$$

$$= Pr(|X - \mu| > \varepsilon\mu) = Pr(X > (1+\varepsilon)\mu) +$$

$$+ Pr(X < (1-\varepsilon)\mu)$$

"$X - \mu > \varepsilon\mu$" $\cup$ "$\mu - X > \varepsilon\mu$"
disjoint events

$$< e^{-\varepsilon^2\mu/3} + e^{-\varepsilon^2\mu/2} < 2e^{-\varepsilon^2\mu/3}$$

We solve in k. Recalling that $\mu = \alpha k$:

$$2e^{-\varepsilon^2\mu/3} = 2e^{-\varepsilon^2\alpha \cdot k/3} \overset{?}{<} \delta \Rightarrow$$

Since we don't know α, we solve instead

$$2e^{-\varepsilon^2\alpha_{min} k/3} \overset{?}{\leq} \delta$$

$\left(2e^{-\varepsilon^2\alpha k/3} < \right)$ $\varepsilon(\alpha > \alpha_{min})$ ✓

We get

$$e^{-\varepsilon^2 \alpha_{min} k/3} \geq \frac{2}{\delta}$$

$$\Leftrightarrow \varepsilon^2 \alpha_{min} k/3 \geq \ln \frac{2}{\delta}$$

$$\Leftrightarrow k \geq \frac{3}{\alpha_{min} \cdot \varepsilon^2} \ln \frac{2}{\delta}$$

The number of samples is:

- linear in $\ln \frac{1}{\delta}$ (confidence)
- inversely proportional to α_{min} (conservative estimate)
- inversely proportional to ε^2 (precision)

We can generalize the above derivation as follows

THEOREM (MONTE CARLO COUNTING)

Let $X_1, \ldots, X_k$ be i.i.d. Bernoulli trials with $E[X_i] = \alpha$.

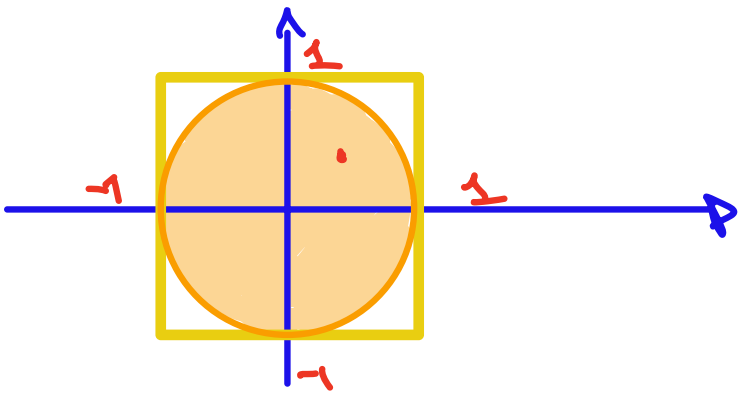
If $k \geq \frac{3}{\alpha \varepsilon^2} \ln(2/\delta)$ then

$$Pr\left(\left| \sum_{i=1}^{k} X_i / k - \alpha \right| / \alpha > \varepsilon \right) < \delta$$

Monte Carlo counting is often used to estimate physical/mathematical quantities that can be related to the probability of a certain observable event

EXAMPLE. Let us use Monte Carlo counting to estimate the value of the irrational constant π to a given accuracy.

Let $(x,y) \in \mathbb{R}^2$ be a random point chosen uniformly in the 2×2 square centered at the origin of $\mathbb{R}^2$:



We know that

$$P\left((x,y) \text{ is in the circle}\right) = P\left(x^2+y^2 \le 1\right)$$

$$= \frac{\text{Area of circle}}{\text{Area of square}} = \frac{\pi}{4}$$

We can generate k points

$$(x_i, y_i) = \text{RANDOM}\left([-1,1] \times [-1,1]\right)$$

and set $Z_i = 1 \iff x_i^2 + y_i^2 \leq 1$

We have that $\alpha = E[Z_i] = \pi/4$

Let $Z = \sum_{i=1}^{k} Z_i$.

Using the Monte Carlo counting theorem we can get an (ε, δ)-approximation to $\pi/4$ by computing the sample mean Z/k, setting

$$k = \frac{3.4}{\pi \varepsilon^2} \ln(2/\delta) \leq \frac{4}{\varepsilon^2} \ln(2/\delta)$$

Thus $4Z$ will be an (ε, δ)-approximation to π

REMARK: The precision of the estimate grows very slowly with the number of tries. To get accuracy $\varepsilon = .01$ (2nd decimal digit!) with confidence .05 (95% correct), we need:

$$\frac{4}{10^{-4}} \cdot \ln\left(\frac{2 \cdot 100}{5}\right) = \frac{4}{10^{-4}} \ln(40) = 148.000 \text{ points!}$$

There are much faster methods for π!

For other physical quantities (e.g. the mass of the proton) Monte Carlo counting is inevitable.

However, faster sampling schemes exist (e.g., Metropolis algorithm, based on Markov Chains)