

## Deviations from Expectation

Sometimes expectation isn't enough. Want to study *deviations*—**probability** and **magnitude** of deviation from expectation.

Example: coupon collection/stable marriage.

- Probability didn't get  $k^{\text{th}}$  coupon after  $r$  rounds is  $(1 - 1/n)^r \leq e^{-r/n}$
- which is  $n^{-\beta}$  for  $r = \beta n \ln n$
- so probability didn't get *some* coupon is at most  $n \cdot n^{-\beta} = n^{1-\beta}$  (using **union bound**)
- we say "time is  $O(n \ln n)$  **with high probability**" because we can make probability  $n^{-\beta}$  for **any** desired  $\beta$  by changing constant that doesn't affect asymptotic claim.
- sometime say "with high probability" when prove it for **some**  $\beta > 1$  even if didn't prove it for all.
- Saying "almost never above  $O(n \ln n)$ " is a much stronger statement than saying " $O(n \ln n)$  on average."

Lower bound:

- After  $n \lg n$  steps,  $1 - 1/n$  chance of getting given coupon
- So chance of getting all  $< (1 - 1/n)^n \approx 1/e$
- So reasonable chance of *not* being faster.
- Can show bound is very tight.

## Tail Bounds—Markov Inequality

At other times, don't want to get down and dirty with problem. So have developed set of bounding techniques that are basically problem independent.

- few assumptions, so applicable almost anywhere
- but for same reason, don't give as tight bounds
- the more you require of problem, the tighter bounds you can prove.

Markov inequality.

- $\Pr[Y \geq t] \leq E[Y]/t$
- $\Pr[Y \geq tE[Y]] \leq 1/t$ .
- Only requires an expectation! So very widely applicable.

Example: coupon collecting  $\Pr[> tn \log n] = O(1/t)$ .

Application:  $ZPP = RP \cap coRP$ .

- If  $RP \cap coRP$ 
  - just run both
  - if neither affirms, run again
  - Each iter has probability 1/2 to affirm
  - So expected iterations 2:
  - So  $ZPP$ .
- If  $ZPP$ 
  - suppose expected time  $T(n)$

- Run for time  $2T(n)$ , then stop and give default answer
- Probability of default answer at most  $1/2$  (Markov)
- So,  $RP$ .
- If flip default answer,  $coRP$

Does this mean can switch Las Vegas to Monte Carlo? Yes.  
 Monte Carlo to Las Vegas? No, unless have checker.

## Chebyshev.

Can make Markov much stronger by generalizing:  $\Pr[h(Y) > t] \leq E[h(Y)]/t$  for **any positive**  $h$ .  
 Better than Markov because uses more info: variance.

- Remind variance, standard deviation.  $\sigma_X^2 = E[(X - \mu_X)^2]$
- For many distributions, this amount of variation is “right amount”
- $E[XY] = E[X]E[Y]$  if independent
- variance of independent variables: sum of variances
- $\Pr[|X - \mu| \geq t\sigma] = \Pr[(X - \mu)^2 \geq t^2\sigma^2] \leq 1/t^2$
- So chebyshev predicts won't stray beyond stdev.
- binomial distribution. variance  $np(1-p)$ . stdev  $\sqrt{n}$ .
- requires (only) a mean and variance. less applicable but more powerful than markov
- Real applications later.

Example: coupon collecting.

- Start with geometric random variable with success prob.  $p$
- mean  $1/p$
- Variance:
  - $E[Y^2] = E[Y^2 | 1 \text{ step}] \cdot p + E[Y^2 | > 1 \text{ step}] \cdot (1-p)$
  - $E[Y^2] = p + E[(Y+1)^2](1-p) = p + (E[Y^2] + 2E[Y] + 1)(1-p)$
  - $Z = (1-p)z + (2-p)/p$
  - $Z = (2-p)/p^2$
  - variance  $(1-p)/p^2 < 1/p^2$

- Coupon collection is sum of independent vars: variance upper bound

$$\sum_n \left( \frac{n}{n-i+1} \right)^2 = n^2 \sum \frac{1}{i^2} = O(n^2)$$

- so  $\Pr[2nH_n] < O(n^2)/(nH_n)^2 + O(1/\ln^2 n)$

## Median Finding

- List  $L$
- median of sample looks like median of whole. neighborhood.
- analysis via Chebyshev bound
- Algorithm
  - choose  $s$  samples *with replacement*
  - take fences before and after sample median
  - keep items between fences. sort.
  - count items on both sides of fences
  - find right element in between
  - works so long as median is between fences
- Analysis
  - claim (i) median within fences and (ii) few items between fences.
  - $s$  Samples  $x_1, \dots, x_s$  in sorted order,  $s = n^{3/4}$ .
  - Sample with replacement to keep analysis clean
  - lemma:  $x_r$  near  $rn/s$  position of original list.
  - To find median element, look at two positions in sorted order of sample:  $a$  at pos.  $s/2 - \sqrt{n}$  and  $b$  at  $s/2 + \sqrt{n}$ 
    - \* Expected number preceding  $n/2$  is  $s/2$ .
    - \* What about variance:
      - each sample indicator has variance  $p(1-p) = 1/4$
      - so  $r$  choices have variance  $< r/4$
      - so  $\sigma = \sqrt{r/4} \leq n^{3/8}/2$
      - chebyshev: prob. less than  $ks/n - \sqrt{n}$  involves deviation by  $n^{1/8}\sigma$  so probability  $O(n^{-1/4})$ .
- Running time:
  - Sorting takes  $O(n^{3/4} \log n) = o(n)$  time
  - Comparing to fences takes  $2n$  time
  - Fail probability  $O(n^{-1/4})$
- Want to improve? repeat!
- Or, repeat till happy: Las Vegas,  $2n + o(n)$  compares in expectation.

Randomized is strictly better:

- Gives important constant factor improvement
- Optimum deterministic:  $\geq (2 + \epsilon)n$
- Optimum randomized:  $\leq (3/2)n + o(n)$

## Pairwise Independence

pseudorandom generators.

- Motivation.
- Idea of randomness as (complexity theoretic) resource like space or time.
- sometime full independence unnecessary
- pairwise independent vars.
- generating over  $Z_p$ .
  - Want random numbers in range  $[1, \dots, p]$
  - pick random  $a, b$
  - $i^{\text{th}}$  random number  $ai + b$
  - Works because invertible over field
- If want over nonprime field, use “slightly larger”  $p$

Application: conserving Random Bits

- Recall Chebyshev inequality
- pairwise sufficient for chebyshev.
- Suppose  $RP$  algorithm using  $n$  bits.
- What do with  $2n$  bits?
- two direct draws: error prob.  $1/4$ .
- pseudorandom generators gives error prob.  $1/t$  for  $t$  trials.
- $\mu = t/2$ .  $\sigma = \sqrt{t}/2$ .
- error if no cert, i.e.  $Y - E[Y] \geq t/2$ , prob.  $1/t$ .