7 Randomization Basics

10 June 2025

Prof. Dr. Sebastian Wild

CS627 (Summer 2025)
Philipps-Universität Marburg

version `2025-06-17 10:01`

Outline

7 Randomization Basics

7.1 Motivation

Computational Lottery?

- If we are faced with solving an NP-hard problem and known smart algorithms are too slow, we likely have to compromise on what “solving” means.
- Classical algorithms are *always* and *exactly* correct.

$\rightsquigarrow$ Here: Let’s compromise on “always”, i. e., allow algorithms to occasionally **fail**!

Computational Lottery?

- If we are faced with solving an NP-hard problem and known smart algorithms are too slow, we likely have to compromise on what "solving" means.
- Classical algorithms are *always* and *exactly* correct.

$\rightsquigarrow$ Here: Let's compromise on "always", i. e., allow algorithms to occasionally **fail**!

ϟ A *deterministic* algorithm A that fails on input x will ***always*** fail for x.

  - $\rightsquigarrow$ What if we require a solution for such an input x? We get **nothing** from A!
  - Must use a form of *nondeterminism*.

Computational Lottery?

- If we are faced with solving an NP-hard problem and known smart algorithms are too slow, we likely have to compromise on what "solving" means.
- Classical algorithms are *always* and *exactly* correct.

⇝ Here: Let's compromise on "always", i. e., allow algorithms to occasionally **fail**!

ϟ A *deterministic* algorithm A that fails on input x will ***always*** fail for x.

  ⇝ What if we require a solution for such an input x? We get **nothing** from A!

  - Must use a form of *nondeterminism*.

- ***Randomization:*** Use *random bits* to guide computation.

⇝ *Instead of always failing on some rare inputs, we rarely fail on any input.*

can make this arbitrarily rare

Why Could Randomization Help?

- Main intuitive reason: (can be) much easier to be 99.999999% correct than 100%
  How can this manifest itself?
  - **Faster and simpler algorithms**
    Random choice can allow to sidestep tricky edge cases
  - We can use **fingerprinting** (a.k.a. checksums) hashing
    Cheap surrogate question, mostly correct, but sometimes wrong.
  - Protect against **adversarial inputs**
    We make our (algorithm's) behavior unpredictable, so it us harder to exploit us.

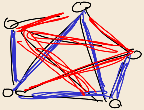
Why Could Randomization Help?

- Main intuitive reason: (can be) much easier to be 99.999999% correct than 100%
  How can this manifest itself?
  - **Faster and simpler algorithms**
    Random choice can allow to sidestep tricky edge cases
  - We can use **fingerprinting** (a.k.a. checksums)
    Cheap surrogate question, mostly correct, but sometimes wrong.
  - Protect against **adversarial inputs**
    We make our (algorithm's) behavior unpredictable, so it us harder to exploit us.

- Also: *probabilistic method* for proofs
  - Goal: Prove existence of discrete object with some property
  - Idea: Design randomized algorithm to find one
  - $\rightsquigarrow$ If algorithm succeeds with prob. > 0, object must exist!

Ramsey theory

complete graph on n vertices



Claim:

$\exists$ monochromatic clique

of size $\geq R(n)$

$R(n) \approx \lg n$

Average-Case Analysis vs. Randomized Algorithms

Average-Case Analysis

- algorithm is **deterministic**
  same input, same computation

Randomized Algorithm (here)

- algorithm is **not** deterministic
  same input, potentially different comp.

Average-Case Analysis vs. Randomized Algorithms

Average-Case Analysis

- algorithm is **deterministic**
  same input, same computation
- input is chosen according to some **probability distribution**

Randomized Algorithm (here)

- algorithm is **not** deterministic
  same input, potentially different comp.
- input is chosen **adversarially** (worst-case inputs)

oblivious adversary
(can't see random bits)

Average-Case Analysis vs. Randomized Algorithms

Average-Case Analysis

- algorithm is **deterministic**
  same input, same computation
- input is chosen according to some **probability distribution**
- cost given as expectation over inputs

Randomized Algorithm (here)

- algorithm is **not** deterministic
  same input, potentially different comp.
- input is chosen **adversarially** (worst-case inputs)
- cost given as expectation over random choices of algorithm

Average-Case Analysis vs. Randomized Algorithms

Average-Case Analysis

- algorithm is **deterministic**
  same input, same computation
- input is chosen according to some **probability distribution**
- cost given as expectation over inputs

Randomized Algorithm (here)

- algorithm is **not** deterministic
  same input, potentially different comp.
- input is chosen **adversarially** (worst-case inputs)
- cost given as expectation over random choices of algorithm

example: sorting by first shuffle

Confusingly enough, the analysis (technique) is often the same!

But: Implications are quite different; randomization is much more versatile and robust.

7.2 Randomized Selection

Separation Example

- Before we introduce randomization more formally, let's see a successful example
- Here, not a "hard" problem, but a showcase where randomization makes something possible that is *provably*

Introductory Example – Quickselect

Selection by Rank

but 0-based & counting dups

- **Given:** array $A[0..n)$ of numbers and number $k \in [0..n)$.
- **Goal:** find element that would be in position k if A was sorted (kth smallest element).
  - $k = \lfloor n/2 \rfloor \rightsquigarrow$ median; $k = \lfloor n/4 \rfloor \rightsquigarrow$ lower quartile
    $k = 0 \rightsquigarrow$ minimum; $k = n - \ell \rightsquigarrow$ ℓth largest

Introductory Example – Quickselect

Selection by Rank

- **Given:** array $A[0..n)$ of numbers and number $k \in [0..n)$.

but 0-based & counting dups

- **Goal:** find element that would be in position k if A was sorted (kth smallest element).
  - $k = \lfloor n/2 \rfloor \rightsquigarrow$ median; $k = \lfloor n/4 \rfloor \rightsquigarrow$ lower quartile
    $k = 0 \rightsquigarrow$ minimum; $k = n - \ell \rightsquigarrow$ ℓth largest

```
procedure quickselect(A[0..n), k):
    l := 0;  r := n
    while r − l > 1
        b := random pivot from A[l..r)
        j := partition(A[l..r), b)
        if j ≥ k then r := j − 1
        if j ≤ k then l := j + 1
    return A[k]
```

- simple algorithm: determine rank of random element, recurse

over random choices

$\rightsquigarrow$ $O(n)$ time in expectation

- worst case: $\Theta(n^2)$
- $O(n)$ also possible deterministically, but algorithms is more involved

median of medians

A closer look at Selection

While all within $\Theta(n)$, we do get a strict separation for selecting the median.

Theorem 7.1 (Bent & John (1985))

Any **deterministic** comparison-based algorithm for finding the median of n elements uses at least $\mathbf{2n} - o(n)$ comparisons in the worst case. ◀

Proof omitted.

A closer look at Selection

While all within $\Theta(n)$, we do get a strict separation for selecting the median.

Theorem 7.1 (Bent & John (1985))

Any **deterministic** comparison-based algorithm for finding the median of n elements uses at least $\mathbf{2n} - o(n)$ comparisons in the worst case. ◀

Proof omitted.

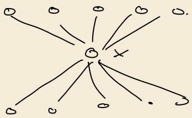
The following weaker result is easier to see:

Theorem 7.2 (Blum et al. (1973))

Any deterministic comparison-based algorithm for finding the median of n elements uses at least $\underbrace{n-1}_{\checkmark} + \underbrace{(n-1)/2}_{\checkmark} \sim 1.5n$ comparisons in the worst case. ◀

Proof: Two types of comparisons

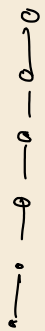
(1) certificate comparisons $n-1$



always necessary for correct algorithm

$n = 11$

sorting



n elements

$n-1$ comps

A Median Adversary

Proof (Theorem 7.2):

(2) "nonessential" comparisons

(not part of certificate)

in particular, comparisons between L and S

m = true median

$L = \{x : x > m\}$

$S = \{x : x < m\}$

$(|S| = |L|)$

Given a deterministic algorithm A, we (the adversary) try to answer comparison queries by A in the least useful way (for A)

Here: maintain elements in 3 sets, S, L and U (undecided)

initially all in U

query "$x \overset{?}{<} y$"

if x and y <u>not</u> in same set, answer $S < U < L$

$\left.\begin{array}{l} x, y \in S \\ x, y \in L \end{array}\right\}$ arbitrary answer

$x, y \in U$ $\quad x < y$, put x to S, y into L

$\Rightarrow$ created <u>one</u> non-essential cmp for A

remove 2 elements from U

$\Rightarrow \geq \frac{n-1}{2}$ non-essential comparisons $\square$

Randomized Selection

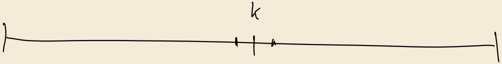
- Can prove: Randomized Quickselect uses in expectation $\sim (2\ln 2 + 2)n \approx 3.39n$ comparisons to find the median
- But we can do better!

Randomized Selection

- Can prove: Randomized Quickselect uses in expectation $\sim (2\ln 2 + 2)n \approx 3.39n$ comparisons to find the median
- But we can do better!

```
procedure floydRivest(A[ℓ..r], k):
    n := r − ℓ
    if n < n_0 return quickselect(A, k)
    s := (1/2) n^{2/3}  // all numbers to be rounded
    sd := (1/2) √(ln(n) s(n − s)/n)
    S[0..s) := random sample from A
    k̂ := s k/n
    p := floydRivest(S, k̂ − sd)
    q := floydRivest(S, k̂ + sd)
    (i, j) := partition A around p_0 and p_1
    if i == k return A[i]
    if j == k return A[j]
    if k < i return floydRivest(A[ℓ..i], k)
    if k > j return floydRivest(A[j..r], k)
    return floydRivest(A[i..j], k)
```

- Variant of Quickselect with huge sample
- Analysis sketch:
  - partition costs $1.5n$ comparisons



Randomized Selection

- Can prove: Randomized Quickselect uses in expectation $\sim (2\ln 2 + 2)n \approx 3.39n$ comparisons to find the median
- But we can do better!

```
procedure floydRivest(A[ℓ..r), k):
    n := r − ℓ
    if n < n_0 return quickselect(A, k)
    s := (1/2) n^(2/3)   // all numbers to be rounded
    sd := (1/2) √(ln(n) s(n − s)/n)
    S[0..s) := random sample from A
    k̂ := s k/n
    p := floydRivest(S, k̂ − sd)
    q := floydRivest(S, k̂ + sd)
    (i, j) := partition A around p_0 and p_1
    if i == k return A[i]
    if j == k return A[j]
    if k < i return floydRivest(A[ℓ..i), k)
    if k > j return floydRivest(A[j..r), k)
    return floydRivest(A[i..j), k)
```

- Variant of Quickselect with huge sample
- Analysis sketch:
  - partition costs $1.5n$ comparisons
  - Everything on sample has cost $o(n)$
  - by the choice of parameters, with prob $1 - o(1)$:
    (a) $i < k < j$ after partition
    (b) $j - i = o(n)$

  $\rightsquigarrow$ all recursive calls expected $o(n)$ cost

Randomized Selection

- Can prove: Randomized Quickselect uses in expectation $\sim (2\ln 2 + 2)n \approx 3.39n$ comparisons to find the median
- But we can do better!

```
procedure floydRivest(A[ℓ..r], k):
    n := r − ℓ
    if n < n_0 return quickselect(A, k)
    s := ½ n^{2/3} // all numbers to be rounded
    sd := ½ √(ln(n)s(n − s)/n)
    S[0..s) := random sample from A
    k̂ := s k/n
    p := floydRivest(S, k̂ − sd)
    q := floydRivest(S, k̂ + sd)
    (i, j) := partition A around p_0 and p_1
    if i == k return A[i]
    if j == k return A[j]
    if k < i return floydRivest(A[ℓ..i), k)
    if k > j return floydRivest(A[j..r), k)
    return floydRivest(A[i..j), k)
```

- Variant of Quickselect with huge sample
- Analysis sketch:
  - partition costs $1.5n$ comparisons
  - Everything on sample has cost $o(n)$
  - by the choice of parameters, with prob $1 - o(1)$:
    (a) $i < k < j$ after partition
    (b) $j - i = o(n)$
  - ⇝ all recursive calls expected $o(n)$ cost

⇝ Randomized median selection with $1.5n + o(n)$ comparisons

⇝ Separation from deterministic case!

Power of Randomness

- Selection by Rank shows two aspects of randomization:
  - A simpler algorithm by avoiding edge cases (like an initial order giving bad pivots)
  - Protection against adversarial inputs
    (inputs constructed with knowledge about the algorithm)

    Here randomization provably more powerful than any thinkable deterministic algorithm!

    constant factor for #cmps

Power of Randomness

- Selection by Rank shows two aspects of randomization:
  - A simpler algorithm by avoiding edge cases (like an initial order giving bad pivots)
  - Protection against adversarial inputs
    (inputs constructed with knowledge about the algorithm)

    Here randomization provably more powerful than any thinkable deterministic algorithm!

    constant factor for #cmps

- What can we gain for (NP-)hard problems?
- But first, let's define things properly.

7.3 Recap of Probability Theory

Probability Theory

- We will quickly revisit some key terms from probability theory
  - Single place to look up notation etc.
- Much will focus on discrete probability, but some continuous tools useful, too

Probability Spaces

Discrete probability space $(\Omega, \mathbb{P})$:

- $\Omega = \{\omega_1, \omega_2, \ldots\}$ a (finite or) *countable* set
- $\mathbb{P} : 2^\Omega \longrightarrow [0, 1]$ a discrete probability measure, i. e.,
  - $\mathbb{P}[\Omega] = 1$
  - $\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}[\omega]$ $\rightsquigarrow$ $\mathbb{P}$ determined by $w_i = \mathbb{P}[\omega_i]$.

fair die

Ω = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}

$\mathbb{P}$[⚀] $= \frac{1}{6}$

Probability Spaces

Discrete probability space $(\Omega, \mathbb{P})$:

- $\Omega = \{\omega_1, \omega_2, \ldots\}$ a (finite or) *countable* set
- $\mathbb{P} : 2^\Omega \to [0,1]$ a discrete probability measure, i. e.,
  - $\mathbb{P}[\Omega] = 1$
  - $\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}[\omega]$ $\rightsquigarrow$ $\mathbb{P}$ determined by $w_i = \mathbb{P}[\omega_i]$.

General probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- Ω is a set of points (the universe)
- $\mathcal{F} \subseteq 2^\Omega$ is a σ-algebra, i. e., (discrete case: $\mathcal{F} = 2^\Omega$; $\Omega = \mathbb{R}$: Borel σ-algebra $\mathcal{B}$ generated by (a, b))
  - $\emptyset \in \mathcal{F}$
  - closed under complementation: $A \in \mathcal{F} \implies \overline{A} = \Omega \setminus A \in \mathcal{F}$
  - closed under *countable* union: $A_1, A_2, \ldots \in \mathcal{F} \implies \bigcup_{i=1}^\infty A_i \in \mathcal{F}$
- $\mathbb{P} : \mathcal{F} \to [0,1]$ is a probability measure, i. e., ($\Omega = \mathbb{R}$ $\rightsquigarrow$ Lebesgue measure $\lambda((a,b)) = b - a$)
  - $\mathbb{P}[\Omega] = 1$
  - If $A_1, A_2, \ldots \in \mathcal{F}$ are pairwise *disjoint* then $\mathbb{P}\left[\bigcup_{i=1}^\infty A_i\right] = \sum_{i=1}^\infty \mathbb{P}[A_i]$

Events

something we can assign a probability to

$A \in \mathcal{F}$ is called an *event* of $(\Omega, \mathcal{F}, \mathbb{P})$; also a *measurable set*.

Events

something we can assign a probability to

$A \in \mathcal{F}$ is called an *event* of $(\Omega, \mathcal{F}, \mathbb{P})$; also a *measurable set*.

Basic properties

- $\mathbb{P}\left[\,\overline{A}\,\right] = 1 - \mathbb{P}[A]$ counter-probability $\quad(\overline{A} = \Omega \setminus A)$
- $\mathbb{P}\left[\bigcup A_i\right] \le \sum_i \mathbb{P}[A]$ the *union bound* (a.k.a. Boole's inequality a.k.a. σ-subadditivity)

Events

something we can assign a probability to

$A \in \mathcal{F}$ is called an *event* of $(\Omega, \mathcal{F}, \mathbb{P})$; also a *measurable set*.

Basic properties

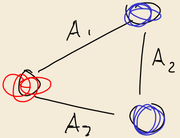
- $\mathbb{P}\left[\,\overline{A}\,\right] = 1 - \mathbb{P}[A]$ counter-probability $\quad (\overline{A} = \Omega \setminus A)$
- $\mathbb{P}\left[\bigcup A_i\right] \le \sum_i \mathbb{P}[A]$ the *union bound* (a.k.a. Boole's inequality a.k.a. σ-subadditivity)
- $\{A_1, \ldots, A_k\}$ *(mutually) independent* $\iff \mathbb{P}\left[\bigcap_i A_i\right] = \prod_i \mathbb{P}[A_i]$

  An infinite set of events is mutually independent if every finite subset is so.

  k-wise independence means that only all size-k subsets are independent.

pairwise indp. $\quad \forall i \neq j \in [k] \quad \mathbb{P}[A_i \cap A_j] = \mathbb{P}[A_i] \cdot \mathbb{P}[A_j]$

$\not\Rightarrow$ mutual independence



A_i = edge has 2 colors at endpoints

$\{A_1, A_2, A_3\}$ 2-wise indp. but not mut. indp.

Events

something we can assign a probability to

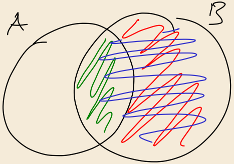
$A \in \mathcal{F}$ is called an *event* of $(\Omega, \mathcal{F}, \mathbb{P})$; also a *measurable set*.

Basic properties

- $\mathbb{P}\left[\overline{A}\right] = 1 - \mathbb{P}[A]$ counter-probability $\quad (\overline{A} = \Omega \setminus A)$
- $\mathbb{P}\left[\bigcup A_i\right] \le \sum_i \mathbb{P}[A]$ the *union bound* (a.k.a. Boole's inequality a.k.a. σ-subadditivity)
- $\{A_1, \ldots, A_k\}$ *(mutually) independent* $\iff \mathbb{P}\left[\bigcap_i A_i\right] = \prod_i \mathbb{P}[A_i]$

  An infinite set of events is mutually independent if every finite subset is so.

  k-wise independence means that only all size-k subsets are independent.



- *conditional probability* for A given B: $\mathbb{P}[A \mid B] = \underline{\mathbb{P}[A \cap B]}/\underline{\mathbb{P}[B]}$
  generally undefined if $\mathbb{P}[B] = 0$
- *law of total probability*: If $\Omega = B_1 \dot{\cup} B_2 \dot{\cup} \cdots$ is a partition of Ω, we have

$$\mathbb{P}[A] = \sum_{\substack{i \\ \mathbb{P}[B_i] \neq 0}} \mathbb{P}[A \mid B_i] \cdot \mathbb{P}[B_i].$$

Random Variables

Random variables (r.v.) $X : \Omega \to \mathcal{X}$; often $\mathcal{X} = \mathbb{R}$ (in general spaces: only *measurable* functions)

fair die

$\Omega = \{$⚀, ⚁, ⚂, ⚃, ⚄, ⚅$\}$

$X : \Omega \to \{1, \ldots, 6\}$

$\mathbb{P}[$⚀$] = \frac{1}{6}$

$Y : \Omega \to \{\text{even}, \text{odd}\}$

Random Variables

Random variables (r.v.) $X : \Omega \to \mathcal{X}$; often $\mathcal{X} = \mathbb{R}$ (in general spaces: only *measurable* functions)

Basic properties and conventions:

- event $\{X = x\}$ is defined as $\{\omega \in \Omega : X(\omega) = x\}$.
- For event A define the indicator r.v. $\mathbb{1}_A$ via $\mathbb{1}_A(\omega) = [\omega \in A]$ $= \begin{cases} 1 & \omega \in A \\ 0 & \text{else} \end{cases}$

Random Variables

Random variables (r.v.) $X : \Omega \to \mathcal{X}$; often $\mathcal{X} = \mathbb{R}$ (in general spaces: only *measurable* functions)

Basic properties and conventions:

- event $\{X = x\}$ is defined as $\{\omega \in \Omega : X(\omega) = x\}$.
- For event A define the indicator r.v. $\mathbb{1}_A$ via $\mathbb{1}_A(\omega) = [\omega \in A]$
- $F_X(x) = \mathbb{P}[X \le x]$ is the *cumulative distribution function (CDF)*.
- X is *discrete* if $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$ is countable. $\mathcal{X} = \mathbb{N}$
- for discrete r.v. X define $f_X(n) = \mathbb{P}[X = n]$ the *probability mass function (PMF)*.
- If F_X is everywhere differentiable, X is *continuous*.
  Then $f_X = F'_X$ is its *probability density function*.

Random Variables

Random variables (r.v.) $X : \Omega \to \mathcal{X}$; often $\mathcal{X} = \mathbb{R}$ (in general spaces: only *measurable* functions)

Basic properties and conventions:

- event $\{X = x\}$ is defined as $\{\omega \in \Omega : X(\omega) = x\}$.
- For event A define the indicator r.v. $\mathbb{1}_A$ via $\mathbb{1}_A(\omega) = [\omega \in A]$
- $F_X(x) = \mathbb{P}[X \le x]$ is the *cumulative distribution function (CDF)*.
- X is *discrete* if $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$ is countable.
- for discrete r.v. X define $f_X(n) = \mathbb{P}[X = n]$ the *probability mass function (PMF)*.
- If F_X is everywhere differentiable, X is *continuous*.
  Then $f_X = F'_X$ is its *probability density function*.

$X \stackrel{\mathcal{D}}{=} Y + Z$

Equality in distribution:

- We write $X \stackrel{\mathcal{D}}{=} Y$ if $F_X = F_Y$

Independent Random Variables

Independence:

- Consider *vector* $\boldsymbol{X} = (X_1, \ldots, X_k)$ as single function from Ω to $\mathbb{R}^k$.
  CDF/PMF/PDF of $\boldsymbol{X}$ is called *joint CDF/PMF/PDF* of $X_1, \ldots, X_k$.

- r.v.s *independent* $\iff$ joint PMF/PDF *factors*:
  X and Y independent $\iff$ $\mathbb{P}[X = x \wedge Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$ for all x, y.

  (Naturally follows from independent events)

Independent Random Variables

Independence:

- Consider *vector* $\boldsymbol{X} = (X_1, \ldots, X_k)$ as single function from Ω to $\mathbb{R}^k$.
  CDF/PMF/PDF of $\boldsymbol{X}$ is called *joint CDF/PMF/PDF* of $X_1, \ldots, X_k$.
- r.v.s *independent* $\iff$ joint PMF/PDF *factors*:
  X and Y independent $\iff$ $\mathbb{P}[X = x \wedge Y = y] = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$ for all x, y.

  (Naturally follows from independent events)

i.i.d. sequences

- We often talk about sequences of random variables $X_1, X_2, \ldots$
- a sequence of *i.i.d.* r.v. $X_1, X_2, \ldots$ mutually (*independent and identically distributed*)
  has $X_i \stackrel{\mathcal{D}}{=} X_1$ and $\{X_i\}_{i \geq 1}$ are mutually independent
  - typical example: sequence of coin tosses (with same coin)

Expected Values

Expectation of an $\mathcal{X}$-valued r.v. X, written $\mathbb{E}[X]$, is given by

- $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot f_X(x)$ for *discrete* X with PMF f_X,
- $\mathbb{E}[X] = \int_{x \in \mathcal{X}} x \cdot f_X(x)\, dx$ for *continuous* X with PDF f_X.
- undefined if sum does not converge / integral does not exist.

Expected Values

Expectation of an $\mathcal{X}$-valued r.v. X, written $\mathbb{E}[X]$, is given by

- $\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot f_X(x)$ for *discrete* X with PMF f_X,
- $\mathbb{E}[X] = \int_{x \in \mathcal{X}} x \cdot f_X(x)\, dx$ for *continuous* X with PDF f_X.
- undefined if sum does not converge / integral does not exist.

Properties:

- *linearity*: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ (X, Y r.v. and a, b constants)
  even if X and Y are not independent
  only for *finite* sums / linear combinations!
- X and Y *independent* $\implies \mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

Conditional Expectation

Similar to conditional *probability*, we can define conditional *expectations*.

- *conditional expectation* on event $\mathbb{E}[X \mid A] \;=\; \sum_x \overset{x}{\vee}\, \mathbb{P}[X = x \mid A]$ for *discrete* X.
  for general A, continuous definition problematic

- *conditional expectation* on $\{Y = y\}$, written $\mathbb{E}[X \mid Y = y]$.
  - for *discrete* X and Y

$$\mathbb{E}[X \mid Y = y] \;=\; \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}\big[X = x \,\big|\, \{Y = y\}\big]$$

Conditional Expectation

Similar to conditional *probability*, we can define conditional *expectations*.

- *conditional expectation* on event $\mathbb{E}[X \mid A] = \sum_x \mathbb{P}[X = x \mid A]$ for *discrete* X.
  for general A, continuous definition problematic
- *conditional expectation* on $\{Y = y\}$, written $\mathbb{E}[X \mid Y = y]$.
  - for *discrete* X and Y
    $$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}\big[X = x \,\big|\, \{Y = y\}\big]$$
  - for *continuous* X and Y, use the joint density $f_{(X,Y)}$ and define the *marginal density* of Y as $f_Y(y) = \int_{x \in \mathcal{X}} f(x, y)\, dx$. Then
    $$\mathbb{E}[X \mid Y = y] = \int_{\mathcal{X}} x \cdot f_{X|Y}(x, y)\, dx \qquad \text{with} \qquad f_{X|Y}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$

Conditional Expectation

Similar to conditional *probability*, we can define conditional *expectations*.

- *conditional expectation* on event $\mathbb{E}[X \mid A] = \sum_x \mathbb{P}[X = x \mid A]$ for *discrete* X.
  for general A, continuous definition problematic
- *conditional expectation* on $\{Y = y\}$, written $\mathbb{E}[X \mid Y = y]$.
  - for *discrete* X and Y
    $$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \cdot \mathbb{P}\big[X = x \,\big|\, \{Y = y\}\big]$$
  - for *continuous* X and Y, use the joint density $f_{(X,Y)}$ and define the *marginal density* of Y as $f_Y(y) = \int_{x \in \mathcal{X}} f(x, y)\, dx$. Then
    $$\mathbb{E}[X \mid Y = y] = \int_{\mathcal{X}} x \cdot f_{X|Y}(x, y)\, dx \qquad \text{with} \qquad f_{X|Y}(x, y) = \frac{f_{(X,Y)}(x, y)}{f_Y(y)}$$
- With $g(y) := \mathbb{E}[X \mid Y = y]$ we obtain a *new r.v.* $\underbrace{\mathbb{E}[X \mid Y]} = g(Y)$.
- *law of total expectation*: $\mathbb{E}[X] = \mathbb{E}_Y\big[\mathbb{E}_X[X \mid Y]\big]$.

Famous Distributions

discrete

- *Bernoulli r.v.* $X \overset{\mathcal{D}}{=} \underbrace{\mathrm{B}(p)}$ $\rightsquigarrow$ $\mathbb{P}[X = 1] = p$, $\mathbb{P}[X = 0] = 1 - p$
- *Binomial r.v.* $Y \overset{\mathcal{D}}{=} \mathrm{Bin}(n, p)$ $\rightsquigarrow$ $Y = X_1 + \cdots + X_n$ for $X_1, \ldots, X_n$ i.i.d. $X_i \overset{\mathcal{D}}{=} \mathrm{B}(p)$
- discrete *uniform r.v.* $X \overset{\mathcal{D}}{=} \mathcal{U}([0..n))$ $\rightsquigarrow$ $\mathbb{P}[X = i] = \frac{1}{n}$ for $i \in [0..n)$ (else 0) $die \overset{\mathcal{D}}{=} \mathcal{U}([1..6])$
- *Geometric r.v.* $X \overset{\mathcal{D}}{=} \mathrm{Geo}(p)$ $\rightsquigarrow$ $\mathbb{P}[X = k] = (1 - p)^{k-1} p$ for $k \in \mathbb{N}_{\geq 1}$

continuous

- continuous uniform $X \overset{\mathcal{D}}{=} \mathcal{U}([0, 1))$ $\rightsquigarrow$ $f_X(x) = 1$ for $x \in [0, 1)$ (else 0)

(of course there are many more)

7.4 Probabilistic Turing Machines

Model of Computation

Definition 7.3 (Probabilistic Turing Machine)

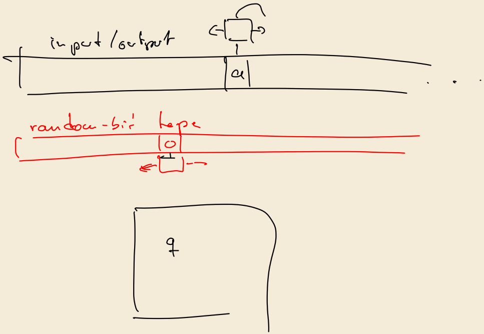
A *probabilistic Turing Machine* (PTM) $M = (Q, \Sigma, \Gamma, \delta, q_0, \Box, q_{\text{halt}})$ is a deterministic TM with an additional read-only tape, filled with random bits.

The *transition function* δ takes as input

- the current state q
- the current tape symbol a
- the current *random-tape symbol* $r \in \{0, 1\}$

and outputs

- the next state q'
- the new tape symbol b
- the tape-head movement $d \in \{L, R, N\}$
- the *random-tape head movement* $d_r \in \{L, R, N\}$



Model of Computation

Definition 7.3 (Probabilistic Turing Machine)

A *probabilistic Turing Machine* (PTM) $M = (Q, \Sigma, \Gamma, \delta, q_0, \Box, q_{\text{halt}})$ is a deterministic TM with an additional read-only tape, filled with random bits.
The *transition function* δ takes as input

- the current state q
- the current tape symbol a
- the current *random-tape symbol* $r \in \{0, 1\}$

and outputs

- the next state q'
- the new tape symbol b
- the tape-head movement $d \in \{L, R, N\}$
- the *random-tape head movement* $d_r \in \{L, R, N\}$

Intended semantics: random tape filled with i.i.d. $\mathrm{B}(\frac{1}{2})$ r.v.

Randomized Computation

- *Configuration* of PTM: $(\alpha q \beta, \rho q \sigma)$
  $\alpha q \beta$ normal TM config
  $\rho\sigma$ content of random tape, with head on first bit of σ

- *computation relation* $\vdash$ similar to TM
  content of random tape unchanged, heads can move independently

- *function computed* by PTM M:
  for input x and **fixed random bits** ρ, computation is deterministic:
  $M(x, \rho) = y$ if $(q_0 x, q_0 \rho) \vdash^\star (q_{\text{halt}} y, \rho' q_{\text{halt}} \rho'')$

Randomized Computation

- *Configuration* of PTM: $(\alpha q \beta, \rho q \sigma)$
  $\alpha q \beta$ normal TM config
  $\rho \sigma$ content of random tape, with head on first bit of σ
- *computation relation* $\vdash$ similar to TM
  content of random tape unchanged, heads can move independently
- *function computed* by PTM M:
  for input x and **fixed random bits** ρ, computation is deterministic:
  $M(x, \rho) \;=\; y$ if $(q_0 x, q_0 \rho) \;\vdash^\star\; (q_{\text{halt}} y, \rho' q_{\text{halt}} \rho'')$

⇝ ***Randomized computation of PTM:***
random variable $M(x, B_0 B_1 B_2 \ldots)$ where $\underline{B_0, B_1, B_2, \ldots \text{ are i.i.d. } \mathrm{B}(\tfrac{1}{2}) \text{ distributed}}$

⇝ Write $\mathbb{P}[M(x) = y] \;=\; \sum_b \mathbb{P}[B_0 B_1 \ldots = b] \cdot [M(x, b) = y]$

- Hope: PTM M so that correct output computed with high probability

Warmup: Rejection Sampling

We assume only random *bits*. How to simulate, say, a fair (6-sided) die?

Warmup: Rejection Sampling

We assume only random *bits*. How to simulate, say, a fair (6-sided) die?

---

procedure rollDie():
    do
        Draw 3 random bits b_2, b_1, b_0
        // Interpret as binary representation of a number in $[0..7]$
        $n = \sum_{i=0}^{2} 2^i b_i$
    while $(n = 0 \vee n = 7)$
    return n

Warmup: Rejection Sampling

We assume only random *bits*. How to simulate, say, a fair (6-sided) die?

---

procedure rollDie():
    do
        Draw 3 random bits b_2, b_1, b_0
        // Interpret as binary representation of a number in $[0..7]$
        $n = \sum_{i=0}^{2} 2^i b_i$
    while $(n = 0 \vee n = 7)$
    return n

Correctness: Every output $1, \ldots, 6$ equally likely by construction.

Warmup: Rejection Sampling

We assume only random *bits*. How to simulate, say, a fair (6-sided) die?

```
procedure rollDie():
    do
        Draw 3 random bits b_2, b_1, b_0
        // Interpret as binary representation of a number in [0..7]
        n = ∑_{i=0}^{2} 2^i b_i
    while (n = 0 ∨ n = 7)
    return n
```

Correctness: Every output $1, \ldots, 6$ equally likely by construction.

\# iterations $\overset{\mathcal{D}}{=} \mathrm{Geo}\left(\frac{3}{4}\right)$

Termination: *Infinite* runs possible!

$\mathbb{E}[\mathrm{Geo}(p)] = \frac{1}{p}$

***Expected* Running Time:** Leave loop with probability $\frac{6}{8} = \frac{3}{4}$ in each iteration

$\rightsquigarrow$ in expectation, only $\dfrac{4}{3} = \displaystyle\sum_{i \geq 1} i \cdot \left(\frac{1}{4}\right)^{i-1} \frac{3}{4}$ repetitions.

rollDie is a correct and practically efficient algorithm.

What can go wrong?

What can go wrong in a randomized computation?

- Computation could run into a deterministic infinite loop (as for deterministic TM)
  - ϟ don't ever terminate, no output
  - ⇝ Clearly don't want that (just as before)

    (annoyingly undecidable to check . . . also just as before)

What can go wrong?

What can go wrong in a randomized computation?

- Computation could run into a deterministic infinite loop (as for deterministic TM)
  - ϟ don't ever terminate, no output
  - $\leadsto$ Clearly don't want that (just as before)

    (annoyingly undecidable to check . . . also just as before)

- Computation could repeatedly have branches that keep looping (as for rollDie)
  - $\leadsto$ For every t, there is a probability $p > 0$ to run for more than t time steps
  - This is a new option that deterministic TMs didn't have
    . . . but nondeterministic TMs did, and we just defined running time to be ∞ there!

    So, is that a problem? Or is it not??

Random Termination

Key question: *What is the probability space for the running time of the PTM simulating rollDie?*

- Note: this could indeed be a problem.
  - $\{0,1\}^\star$ (the set of **finite** bitstrings) is countably infinite (=discrete)
  - But the set of ***infinite strings*** (ω-*language*) is not!
    $\{0,1\}^\omega = \{b_0 b_1 \ldots : b_i \in \{0,1\}\} = \{b : b : \mathbb{N}_0 \to \{0,1\}\}$ surjectively maps to $[0,1) \subset \mathbb{R}$

    $b \mapsto 0.b_0 b_1 b_2 \ldots$

Random Termination

Key question: *What is the probability space for the running time of the PTM simulating rollDie?*

- Note: this could indeed be a problem.
  - $\{0,1\}^\star$ (the set of **finite** bitstrings) is countably infinite (=discrete)
  - But the set of ***infinite strings*** (ω-*language*) is not!
    $\{0,1\}^\omega = \{b_0 b_1 \ldots : b_i \in \{0,1\}\} = \{b : b : \mathbb{N}_0 \to \{0,1\}\}$ surjectively maps to $[0,1) \subset \mathbb{R}$
    $b \mapsto 0.b_0 b_1 b_2 \ldots$
- Config $(\alpha q \beta, \rho q \boldsymbol{\sigma})$ for PTM needs $\sigma \in \{0,1\}^{\boldsymbol{\omega}}$ in general

Random Termination

Key question: *What is the probability space for the running time of the PTM simulating rollDie?*

- Note: this could indeed be a problem.
  - $\{0,1\}^\star$ (the set of **finite** bitstrings) is countably infinite (=discrete)
  - But the set of ***infinite strings*** (ω-*language*) is not!
    $\{0,1\}^\omega = \{b_0 b_1 \ldots : b_i \in \{0,1\}\} = \{b : b : \mathbb{N}_0 \to \{0,1\}\}$ surjectively maps to $[0,1) \subset \mathbb{R}$

    $b \mapsto 0.b_0 b_1 b_2 \ldots$
- Config $(\alpha q \beta, \rho q \boldsymbol{\sigma})$ for PTM needs $\sigma \in \{0,1\}^{\boldsymbol{\omega}}$ in general

- Define the random variable $Time_M(x) \in \mathbb{N}_0 \cup \{\infty\}$ on the *Bernoulli probability space*
  - generators: $\{\pi_x : x \in \{0,1\}^\star\}$ where $\pi_x = \{xw : w \in \{0,1\}^\omega\} \subseteq \{0,1\}^\omega$
  - Bernoulli σ-algebra: smallest $\mathcal{F}$ containing all $\{\pi_x\}_x$ that is closed under countable union and complement
  - $\mathbb{P}[\pi_x] = 2^{-|x|}$

rollDie terminates iff $\rho \in T \subseteq \{0,1\}^\omega$ $\quad\quad$ $\mathbb{P}[T]$

Random Termination

Key question: *What is the probability space for the running time of the PTM simulating rollDie?*

- Note: this could indeed be a problem.
  - $\{0,1\}^\star$ (the set of **finite** bitstrings) is countably infinite (=discrete)
  - But the set of ***infinite strings*** (ω*-language*) is not!
    $\{0,1\}^\omega = \{b_0 b_1 \ldots : b_i \in \{0,1\}\} = \{b : b : \mathbb{N}_0 \to \{0,1\}\}$ surjectively maps to $[0,1) \subset \mathbb{R}$
    $b \mapsto 0.b_0 b_1 b_2 \ldots$
- Config $(\alpha q \beta, \rho q \boldsymbol{\sigma})$ for PTM needs $\sigma \in \{0,1\}^{\boldsymbol{\omega}}$ in general

- Define the random variable $Time_M(x) \in \mathbb{N}_0 \cup \{\infty\}$ on the *Bernoulli probability space*
  - generators: $\{\pi_x : x \in \{0,1\}^\star\}$ where $\pi_x = \{xw : w \in \{0,1\}^\omega\} \subseteq \{0,1\}^\omega$
  - Bernoulli σ-algebra: smallest $\mathcal{F}$ containing all $\{\pi_x\}_x$ that is closed under countable union and complement
  - $\mathbb{P}[\pi_x] = 2^{-|x|}$

$\rightsquigarrow$ expectations over $\rho \in \{0,1\}^\omega$, the infinite initial random-bit tape input are well-defined

(Expected) Time

Definition 7.4 (PTM running time)

For a PTM M, we define $\mathit{time}_M(x)$ as for nondeterministic TMs as the supremum of time steps over all computations.

worst case

$\mathit{time}_{\mathrm{rollDie}}() = \infty$

Moreover, we define the *expected time* as

$$\mathbb{E}\text{-}\mathit{time}_M(x) \;=\; \mathbb{E}[\mathit{time}_M(x)] \;=\; \mathbb{E}_{\rho}\Big[\inf\{t \in \mathbb{N}_0 : (q_0 x, q_0 \underset{\text{random}}{\underline{\rho}}) \vdash^{t} (q_{\mathrm{halt}} y, \rho' q_{\mathrm{halt}} \rho'')\Big]$$

Similarly

$$\mathbb{E}\text{-}\mathit{Time}_M(n) \;=\; \sup\big\{\mathbb{E}\text{-}\mathit{time}_M(x) : x \in \Sigma^n\big\}$$

◀

(Expected) Time

Definition 7.4 (PTM running time)

For a PTM M, we define $time_M(x)$ as for nondeterministic TMs as the supremum of time steps over all computations.
Moreover, we define the *expected time* as

$$\mathbb{E}\text{-}time_M(x) \;=\; \mathbb{E}[time_M(x)] \;=\; \mathbb{E}_\rho\Big[\inf\{t \in \mathbb{N}_0 : (q_0 x, q_0 \rho) \vdash^t (q_{\text{halt}} y, \rho' q_{\text{halt}} \rho'')\Big]$$

Similarly

$$\mathbb{E}\text{-}Time_M(n) \;=\; \sup\big\{\mathbb{E}\text{-}time_M(x) : x \in \Sigma^n\big\}$$

◀

- We can of course also study full distribution of $time_M(x)$
- Useful property of expected time:
  $\mathbb{E}\text{-}time_M(x) < \infty \quad$ iff $\quad \mathbb{P}[time_M(x) = \infty] = 0$

A New Complexity Measure: Random Bits

Definition 7.5 (Random-bit complexity)

For a PTM M computing with input alphabet Σ, the *random-bit cost* for an input $x \in \Sigma^\star$ is denote by

$$random_M(x) \;=\; \sup\left\{|\rho'| : (xq_0, q_0\rho) \vdash^\star (\alpha q \beta, \rho' q \rho'') \vdash^\star (q_{\text{halt}} y, \rho' q_{\text{halt}} \rho'')\right\}$$

and similarly

$$Random_M(n) \;=\; \sup\left\{random_M(x) : x \in \Sigma^n\right\}.$$

Further, the *expected random-bit cost* are defined as
$\mathbb{E}\text{-}random_M(x) = \mathbb{E}_\rho[random_M(x)]$ and
$\mathbb{E}\text{-}Random_M(n) = \sup\left\{\mathbb{E}\text{-}random_M(x) : x \in \Sigma^n\right\}$ ◀

Randomization vs. Nondeterminism

- Superficially similar concepts
- Key difference: meaning of number of computations of TM
  - nondeterministic TM: accept if **some (single)** accepting computation is possible
  - randomized TM: accept if **most** possible computations are accepting

$\rightsquigarrow$ nondeterminism = purely theoretical construction (overly powerful yardstick)

- randomization = widely applied efficient design technique

7.5 Classification of Randomized Algorithms

Las Vegas

Consider here the general problem to compute some *function* $f : \Sigma^\star \to \Gamma^\star$.

$\rightsquigarrow$ Covers *decision problems* $L \subseteq \Sigma^\star$ by setting $\Gamma = \{0, 1\}$ and $f(w) = \begin{cases} 1 & w \in L \\ 0 & w \notin L \end{cases}$

Las Vegas

Consider here the general problem to compute some *function* $f : \Sigma^\star \to \Gamma^\star$.

$\rightsquigarrow$ Covers *decision problems* $L \subseteq \Sigma^\star$ by setting $\Gamma = \{0, 1\}$ and $f(w) = \begin{cases} 1 & w \in L \\ 0 & w \notin L \end{cases}$

Definition 7.6 (Las Vegas Algorithm)

A randomized algorithm A is a *Las-Vegas (LV) algorithm* for a problem $f : \Sigma^\star \to \Gamma^\star$ if for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[time_A(x) < \infty\big] \;=\; 1$ (*terminate* almost surely)

(b) $A(x) \;\in\; \{f(x), \boxed{?}\}$ (answer always *correct or "don't know"*)

(c) $\mathbb{P}\big[A(x) = f(x)\big] \;\geq\; \frac{1}{2}$ (*correct half the time*)

Don't Know vs. Won't Terminate

Theorem 7.7 (Don't know don't needed)
Every Las Vegas algorithm A for $f : \Sigma^\star \to \Gamma^\star$ can be transformed into a randomized algorithm B for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[B(x) = f(x)\big] \;=\; 1$ (always correct) (can never output $\boxed{?}$, not "just" almost surely)

(b) $\mathbb{E}\text{-}time_B(x) \;\le\; 2 \cdot time_A(x)$

Don't Know vs. Won't Terminate

Theorem 7.7 (Don't know don't needed)
Every Las Vegas algorithm A for $f : \Sigma^\star \to \Gamma^\star$ can be transformed into a randomized algorithm B for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[B(x) = f(x)\big] \;=\; 1$ (always correct)

(b) $\mathbb{E}\text{-}\mathit{time}_B(x) \;\le\; 2 \cdot \mathit{time}_A(x)$ ◀

Proof:
See exercises. ■

Don't Know vs. Won't Terminate

Theorem 7.7 (Don't know don't needed)

Every Las Vegas algorithm A for $f : \Sigma^\star \to \Gamma^\star$ can be transformed into a randomized algorithm B for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[B(x) = f(x)\big] \;=\; 1$ (always correct)

(b) $\mathbb{E}\text{-}time_B(x) \;\leq\; 2 \cdot time_A(x)$ ◀

Proof:
See exercises. ■

Theorem 7.8 (Termination Enforcible)

Every randomized algorithm B for $f : \Sigma^\star \to \Gamma^\star$ with $\mathbb{P}[B(x) = f(x)] = 1$ can be transformed into a Las Vegas algorithm A for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[A(x) = f(x)\big] \;\geq\; \frac{1}{2}$

(b) $time_A(x) \;\leq\; 2 \cdot \mathbb{E}\text{-}time_B(x)$ (always terminates). ◀

Don't Know vs. Won't Terminate

Theorem 7.7 (Don't know don't needed)

Every Las Vegas algorithm A for $f : \Sigma^\star \to \Gamma^\star$ can be transformed into a randomized algorithm B for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[B(x) = f(x)\big] \;=\; 1$ (always correct)

(b) $\mathbb{E}\text{-}\mathit{time}_B(x) \;\le\; 2 \cdot \mathit{time}_A(x)$ ◀

Proof:
See exercises. ■

Theorem 7.8 (Termination Enforcible)

Every randomized algorithm B for $f : \Sigma^\star \to \Gamma^\star$ with $\mathbb{P}[B(x) = f(x)] = 1$ can be transformed into a Las Vegas algorithm A for f so that for all $x \in \Sigma^\star$ holds

(a) $\mathbb{P}\big[A(x) = f(x)\big] \;\ge\; \frac{1}{2}$

(b) $\mathit{time}_A(x) \;\le\; 2 \cdot \mathbb{E}\text{-}\mathit{time}_B(x)$ (always terminates). ◀

Proof:
See exercises. ■

Las Vegas Variants

$\rightsquigarrow$ *Can trade expected time bound for worst-case bound by allowing "don't know" and* ***vice versa****!*

Both types are called commonly LV algorithms; where helpful, we distinguish:

(A) *Always-Decisive Las Vegas algorithms* (output of Theorem 7.7)

(B) *Always-Terminating Las Vegas algorithms* (output of Theorem 7.8)

Las Vegas Examples

rollDie by rejection sampling is Las Vegas of unbounded worst-case type.

Easy to transform into Las Vegas according to Definition 7.6:

```
procedure rollDieLasVegas:
    Draw 3 random bits b_2, b_1, b_0
    n = ∑_{i=0}^{2} 2^i b_i   // Interpret as binary representation of a number in [0 : 7]
    if (n = 0 ∨ n = 7)
        return ?
    else
        return n
```

Other famous examples: *(randomized)* *Quicksort* and *Quickselect*

- always correct *and*
- $time(n) = O(n^2) < \infty$
- much better average:
  - $\mathbb{E}\text{-}time_{QSort}(n) = \Theta(n \log n)$
  - $\mathbb{E}\text{-}time_{QSelect}(n) = \Theta(n)$

To Err is Algorithmic

Sometimes sensible to allow *wrong / imprecise* answers ...
but random should not mean *arbitrary*!

To Err is Algorithmic

Sometimes sensible to allow *wrong / imprecise* answers ...
but random should not mean *arbitrary*!

Definition 7.9 (Monte Carlo Algorithm)

A randomized algorithm A is a *Monte Carlo algorithm* for $f : \Sigma^\star \to \Gamma^\star$

- with *bounded error* if $\quad \exists \varepsilon > 0 \, \forall x \in \Sigma^\star \; : \; \mathbb{P}[A(x) = f(x)] \ge \frac{1}{2} + \varepsilon.$
- with ***un**bounded error* if $\quad \forall x \in \Sigma^\star \; : \; \mathbb{P}[A(x) = f(x)] > \frac{1}{2}.$

To Err is Algorithmic

Sometimes sensible to allow *wrong / imprecise* answers ...
but random should not mean *arbitrary*!

Definition 7.9 (Monte Carlo Algorithm)

A randomized algorithm A is a *Monte Carlo algorithm* for $f : \Sigma^\star \to \Gamma^\star$

- with *bounded error* if $\quad \exists \varepsilon > 0 \, \forall x \in \Sigma^\star \;:\; \mathbb{P}[A(x) = f(x)] \ge \frac{1}{2} + \varepsilon.$
- with ***un**bounded error* if $\quad \forall x \in \Sigma^\star \;:\; \mathbb{P}[A(x) = f(x)] > \frac{1}{2}.$

Seems like a minuscule difference? We will see it is vital!

7.6 Tail Inequalities and Concentration Bounds

How To Summarize A Random Variable?

- running time of randomized algorithm is a **random variable**
- For two randomized algorithms A and B, we'd like to decide which is better
  Whether A is faster than B is also random.

$\rightsquigarrow$ need a way to **compare** random variables!

How To Summarize A Random Variable?

- running time of randomized algorithm is a **random variable**
- For two randomized algorithms A and B, we'd like to decide which is better
  Whether A is faster than B is also random.

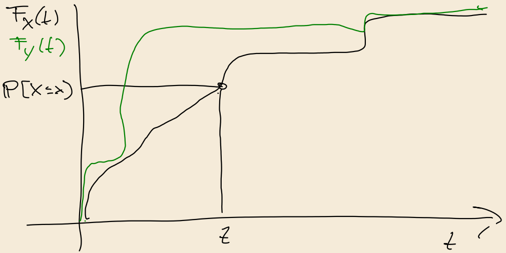
⇝ need a way to **compare** random variables!

- One option: *stochastic dominance*
  - If $\forall t : \mathbb{P}[\overset{X}{time_A(x)} \geq t] \geq \mathbb{P}[\overset{Y}{time_B(x)} \geq t]$,
    we say $time_A(x)$ (weakly) stochastically dominates $time_B(x)$ (on input x)
  - Would rather use B then!

$\mathbb{P}[X \geq t]$ $\quad X \succeq Y$

$\overbrace{1 - F_X(t)} \geq 1 - F_Y(t)$

$F_X(t) \leq F_Y(t)$



How To Summarize A Random Variable?

- running time of randomized algorithm is a **random variable**
- For two randomized algorithms A and B, we'd like to decide which is better
  Whether A is faster than B is also random.

$\rightsquigarrow$ need a way to **compare** random variables!

- One option: *stochastic dominance*
  - If $\forall t : \mathbb{P}[time_A(x) \ge t] \ge \mathbb{P}[time_B(x) \ge t]$,
    we say $time_A(x)$ (weakly) stochastically dominates $time_B(x)$ (on input x)
  - Would rather use B then!

  👎 dominance rarely true for real algorithms

  👎 no prediction of running time / comparison with explicit bound

How To Summarize A Random Variable?

- running time of randomized algorithm is a **random variable**
- For two randomized algorithms A and B, we'd like to decide which is better
  Whether A is faster than B is also random.

$\rightsquigarrow$ need a way to **compare** random variables!

- One option: *stochastic dominance*
  - If $\forall t : \mathbb{P}[time_A(x) \ge t] \ge \mathbb{P}[time_B(x) \ge t]$,
    we say $time_A(x)$ (weakly) stochastically dominates $time_B(x)$ (on input x)
  - Would rather use B then!

  👎 dominance rarely true for real algorithms

  👎 no prediction of running time / comparison with explicit bound

$\rightsquigarrow$ look at **expected value** $\mathbb{E}$-$time(x)$ (randomized version of average case)

👍 simple (one number); reflects typical case

👎 not always reliable / representative

When Expectation Isn't Enough

- Two hypothetical algorithms:
  - A takes 1 step in half the cases and 3 steps otherwise
  - B takes 1 step in 99% of cases and **101** steps otherwise
  - $\rightsquigarrow$ both have $\mathbb{E}\text{-}time(x) = 2$
  - probably want A ... certainly would want to be able to distinguish them!

When Expectation Isn't Enough

- Two hypothetical algorithms:
  - A takes 1 step in half the cases and 3 steps otherwise
  - B takes 1 step in 99% of cases and **101** steps otherwise
  - ⇝ both have $\mathbb{E}\text{-}time(x) = 2$
  - probably want A … certainly would want to be able to distinguish them!

- **Goal:** Strengthen algorithms so $time(x)$ rarely far from $\mathbb{E}\text{-}time(x)$
  - formally: bound probability that X (far) exceeds $\mathbb{E}[X]$
    - ⇝ *concentration bounds* a.k.a. *tail inequalities*
  - 👍 can then compare these typical times again
  - 👍 also obtain more reliable algorithms
  - ⇝ Let's establish some tools for that!

With High Probability

⚠ not always the same definition

Definition 7.10 (With high probability)

We say

- an event $X = X(n)$ happens *with high probability (w.h.p.)* when $\forall c : \mathbb{P}[X(n)] = 1 \pm O(n^{-c})$ as $n \to \infty$.
- a random variable $X = X(n)$ is *in $O(f(n))$ with high probability (w.h.p.)* when $\forall c\ \exists d : \mathbb{P}[X \le df(n)] = 1 \pm O(n^{-c})$ as $n \to \infty$.
  (This means, the constant in $O(f(n))$ may depend on c.)

With High Probability

Definition 7.10 (With high probability)

We say

- an event $X = X(n)$ happens *with high probability (w.h.p.)* when $\forall c \;:\; \mathbb{P}[X(n)] \;=\; 1 \pm O(n^{-c})$ as $n \to \infty$.
- a random variable $X = X(n)$ is *in $O(f(n))$ with high probability (w.h.p.)* when $\forall c\ \exists d \;:\; \mathbb{P}[X \le df(n)] \;=\; 1 \pm O(n^{-c})$ as $n \to \infty$.
  (This means, the constant in $O(f(n))$ may depend on c.) ◀

- Very strong notion: failure probability smaller than any polynomial

$\rightsquigarrow$ If A succeeds w.h.p. then also polynomially many repetitions of A succeed w.h.p.

Lemma 7.11 (Repetitions w.h.p.)

Suppose A is an algorithm that w.h.p. does not fail.
In n^d *independent* repetitions of A on inputs of size n, w.h.p. *no* repetition fails. ◀

With High Probability [2]

- an event $X = X(n)$ happens *with high probability (w.h.p.)* when $\forall c : \mathbb{P}[X(n)] = 1 \pm O(n^{-c})$ as $n \to \infty$.

Proof (Lemma 7.11):

Let c from the definition of w.h.p. be given.

$F_i \equiv$ ith A fails $\qquad B = \bigcup_{i=1}^{n^d} A_i \qquad$ claim: $\overline{B}$ whp

$\rightsquigarrow$ events that happen with high probability can be combined

With High Probability [2]

Proof (Lemma 7.11):

Let c from the definition of w.h.p. be given.

The event F_i that the ith run of A fails happens with probability $O(n^{-(c+d)})$ by definition.

$\rightsquigarrow$ events that happen with high probability can be combined

With High Probability [2]

Proof (Lemma 7.11):

Let c from the definition of w.h.p. be given.
The event F_i that the ith run of A fails happens with probability $O(n^{-(c+d)})$ by definition.

Then $\mathbb{P}[\cup_{i=1}^{n^d} F_i]$

$\rightsquigarrow$ events that happen with high probability can be combined

With High Probability [2]

Proof (Lemma 7.11):

Let c from the definition of w.h.p. be given.

The event F_i that the ith run of A fails happens with probability $O(n^{-(c+d)})$ by definition.

$$\text{Then } \mathbb{P}[\cup_{i=1}^{n^d} F_i] \underbrace{\leq}_{\text{union bound}} \sum_{i=1}^{n^d} \mathbb{P}[F_i]$$

$\rightsquigarrow$ events that happen with high probability can be combined

With High Probability [2]

Proof (Lemma 7.11):

Let c from the definition of w.h.p. be given.

The event F_i that the ith run of A fails happens with probability $O(n^{-(c+d)})$ by definition.

Then $\mathbb{P}[\cup_{i=1}^{n^d} F_i] \le \sum_{i=1}^{n^d} \mathbb{P}[F_i] = n^d \cdot O(n^{-(c+d)})$

$\rightsquigarrow$ events that happen with high probability can be combined

With High Probability [2]

Proof (Lemma 7.11):
Let c from the definition of w.h.p. be given.
The event F_i that the ith run of A fails happens with probability $O(n^{-(c+d)})$ by definition.

Then $\mathbb{P}[\cup_{i=1}^{n^d} F_i] \;\le\; \displaystyle\sum_{i=1}^{n^d} \mathbb{P}[F_i] \;=\; n^d \cdot O(n^{-(c+d)}) \;=\; O(n^{-c}).$ ∎

$\rightsquigarrow$ events that happen with high probability can be combined

Concentration I – Markov's Inequality

Theorem 7.12 (Markov's Inequality)

Let $X \in \mathbb{R}_{\geq 0}$ be a r.v. that assumes only *weakly positive* values. Then holds

$$\forall a > 0 \ : \ \mathbb{P}[X \geq a] \ \leq \ \frac{\mathbb{E}[X]}{a}$$

Proof: Let $a > 0$ given, define $I := \mathbb{1}\{X \geq a\} = [X \geq a] = \begin{cases} 1 & X \geq a \\ 0 & X < a \end{cases}$

$I \leq \dfrac{X}{a}$

if $X \geq a \rightsquigarrow I = 1$ and $\frac{X}{a} \geq 1$ ✓

if $X < a \rightsquigarrow I = 0$ and $X \geq 0 \rightsquigarrow \frac{X}{a} \geq 0$ ✓

take $\mathbb{E}$

$$\frac{\mathbb{E}[X]}{a} \geq \mathbb{E}[I] = \mathbb{P}[X \geq a]$$

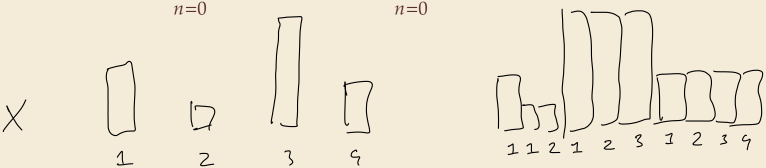
Since $X \geq 0$ implies $\mathbb{E}[X] \geq 0$, nicer equivalent form: $\forall a > 0 \ : \ \mathbb{P}\big[X \geq a\mathbb{E}[X]\big] \ \leq \ \dfrac{1}{a}$

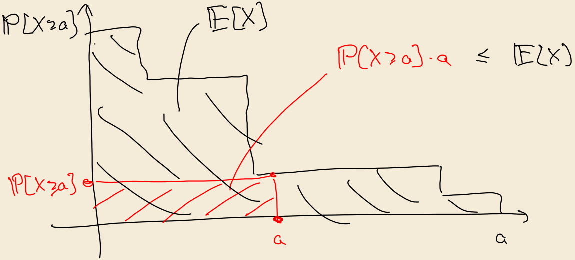
Mnemonic: With probability at least $\frac{1}{a}$, won't exceed expectation by a factor a.

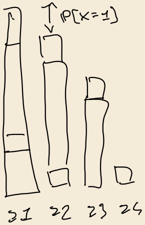
Markov's Inequality Visually

If $X \in \mathbb{N}_0$, we can visualize Markov's Inequality nicely.

Recall: $X \in \mathbb{N}_0$ implies $\mathbb{E}[X] = \sum_{n=0}^{\infty} n \cdot \mathbb{P}[X = n] = \sum_{n=0}^{\infty} \mathbb{P}[X \geq n]$.







Moments

- Markov's Inequality is tight (for some r.v.), but not usually a strong concentration result.
- but we can apply it to $f(X)$ for any (positive) *function* of X!

Moments

- Markov's Inequality is tight (for some r.v.), but not usually a strong concentration result.
- but we can apply it to $f(X)$ for any (positive) *function* of X!

Towards this, we consider moments of r.v.:

Definition 7.13 (Moments, variance, standard deviation)
For a random variable $X \in \mathbb{R}$,

- $\mathbb{E}[X^k]$ is the *kth moment* of X. $\mathbb{E}[X]$ 1st moment
- $\mathbb{E}\left[\left|X - \mathbb{E}[X]\right|^k\right]$ is the *kth centered moment* of X.
- The *variance* of X is the second centered moment: $\mathrm{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$
- The *standard deviation* of X is $\sigma[X] = \sqrt{\mathrm{Var}[X]}$. ◀

Note: None of the moments is guaranteed to exist!

Concentration II – Chebychev's Inequality

Using second moments, we obtain a stronger concentration inequality.

Theorem 7.14 (Chebychev's Inequality)

Let X be a random variable. We have

$$\forall a > 0 \quad : \quad \mathbb{P}\Big[\big|X - \mathbb{E}[X]\big| \geq a\Big] \;\leq\; \frac{\mathrm{Var}[X]}{a^2}.$$

Proof: Apply Markov to $Y = (X - \mathbb{E}[X])^2$ $\qquad \mathbb{P}[Y \geq b] \leq \frac{\mathbb{E}[Y]}{b}$

$$\mathbb{P}[Y \geq b] = \mathbb{P}\left[(X-\mathbb{E}[X])^2 \geq b\right] = \mathbb{P}\Big[\big|X-\mathbb{E}[X]\big| \geq \underbrace{\sqrt{b}}_{= a}\Big] \leq \frac{\mathbb{E}[Y]}{b} = \frac{\mathrm{Var}[X]}{a^2}$$

take $\sqrt{\ }$ on both sides

Convergence in Probability

Corollary 7.15 (Chebychev Concentration)

Let $X_1, X_2, \ldots$ be a sequence of random variables and assume

- $\mathbb{E}[X_n]$ and $\operatorname{Var}[X_n]$ exist for all n and
- $\boxed{\sigma[X_n] = o(\mathbb{E}[X_n])}$ as $n \to \infty$.

Then holds

$$\forall \varepsilon > 0 \ : \ \mathbb{P}\left[\left|\frac{X_n}{\mathbb{E}[X_n]} - 1\right| \geq \varepsilon\right] \to 0 \qquad (n \to \infty).$$

This means $\dfrac{X_n}{\mathbb{E}[X_n]}$ *converges in probability* to 1.

Convergence in Probability

Corollary 7.15 (Chebychev Concentration)

Let $X_1, X_2, \ldots$ be a sequence of random variables and assume

- $\mathbb{E}[X_n]$ and $\mathrm{Var}[X_n]$ exist for all n and
- $\sigma[X_n] = o(\mathbb{E}[X_n])$ as $n \to \infty$.

Then holds

$$\forall \varepsilon > 0 \quad : \quad \mathbb{P}\left[\left|\frac{X_n}{\mathbb{E}[X_n]} - 1\right| \ge \varepsilon\right] \;\to\; 0 \qquad (n \to \infty).$$ ◀

This means $\dfrac{X_n}{\mathbb{E}[X_n]}$ *converges in probability* to 1.

We're getting there, but this is still a far cry from our "with high probability"!
(superpolynomial convergence rate in above limit)

Concentration III – Chernoff Bounds

Stronger concentration inequalities require more assumptions on distribution of X.
A classical one is that X consists of many **small and independent** parts.

Concentration III – Chernoff Bounds

Stronger concentration inequalities require more assumptions on distribution of X.
A classical one is that X consists of many **small and independent** parts.

Theorem 7.16 (Chernoff Bound for Bernoulli trials)

Let $X_1, \ldots, X_n \in \{0,1\}$ be *(mutually) independent* with $X_i \stackrel{\mathcal{D}}{=} \mathrm{B}(p_i)$.
Define $X = X_1 + \cdots + X_n$ and $\underset{\overset{\parallel}{\mathbb{E}[X]}}{\mu} = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = p_1 + \cdots + p_n$. Then holds

$\delta = 1$ $\quad \frac{e}{2^2} = \frac{e}{4} < \frac{?}{4}$

$$(1) \qquad \forall \delta > 0 \;:\; \mathbb{P}[X \geq (1+\delta)\mu] \;<\; \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

$$(2) \qquad \forall \delta \in (0,1] \;:\; \mathbb{P}[X \geq (1+\delta)\mu] \;\leq\; \exp(-\mu\delta^2/3)$$

$(1) \Rightarrow (2)$

Proof: Again Markov ... Let $t > 0$ $\quad Y = e^{t \cdot X} \geq 0$ $\qquad Y_i = e^{tX_i}$

$Y_1, \ldots, Y_n$ mut. indep.

$$\mathbb{P}[X \geq (1+\delta)\mu] \overset{e^{t\,\cdot}\text{ on both sides}}{\stackrel{!}{=}} \mathbb{P}[Y \geq e^{t(1+\delta)\mu}] \underset{\text{Markov}}{\leq} \frac{\mathbb{E}[Y]}{e^{t(1+\delta)\mu}}$$

$$\mathbb{E}[Y] = \mathbb{E}[e^{t \cdot X}] = \mathbb{E}\left[e^{t \sum X_i}\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{t \cdot X_i}\right] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}]$$

Concentration III – Chernoff Bounds [2]

Proof (cont.):

$$\mathbb{E}[e^{tX_i}] = 1 \cdot \mathbb{P}[X_i = 0] + e^t\, \mathbb{P}[X_i = 1]$$

$$= (1-p_i) + e^t p_i$$

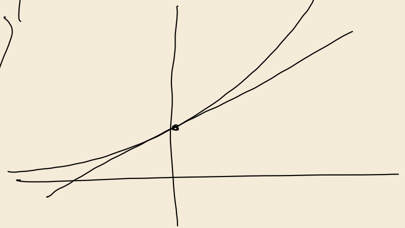
$$\mathbb{E}[Y] = \prod_{i=1}^{n} \Big(\underbrace{1 + p_i(e^t - 1)}_{\leq\, e^{p_i(e^t-1)}} \Big)$$

$$\leq \exp\Big(\sum_{i=1}^{n} p_i (e^t - 1) \Big)$$

$$= \exp\Big((e^t - 1) \underbrace{\sum p_i}_{=\mu} \Big) \qquad (*)$$

convex function lies above all its tangents

$$1 + y \leq e^y$$



$$\mathbb{E}[X \geq (1+\delta)\mu] \leq \frac{\mathbb{E}[Y]}{e^{t(1+\delta)\mu}} \qquad \boxed{t > 0}$$

$$\underset{(*)}{\leq} \frac{e^{\mu(e^t-1)}}{e^{\mu t(1+\delta)}} = \frac{\left(e^{e^t-1}\right)^\mu}{\left(e^{t(1+\delta)}\right)^\mu}$$

with $t = \ln(1+\delta) > 0$

$$= \frac{e^{\mu(1+\delta-1)}}{(1+\delta)^{\mu(1+\delta)}} = \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mu}$$

$\square$

Chernoff Bound for Binomial Distribution

The algorithmically most widely used special case has identical coin flips.

Corollary 7.17 (Chernoff Bound for Binomial Distribution)

Let $X \stackrel{\mathcal{D}}{=} \mathrm{Bin}(n, p)$. Then

$$\forall \delta \geq 0 \quad : \quad \mathbb{P}\left[\left|\frac{X}{n} - p\right| \geq \delta\right] \;\leq\; 2\exp(-2\delta^2 n) \qquad \blacktriangleleft$$

$\rightsquigarrow$ $\boxed{\mathrm{Bin}(n, p) \;\in\; np \pm n^{0.501} \text{ w.h.p.}}$