

INFO 6205

Program Structure and Algorithms

Nik Bear Brown
Probability

Topics

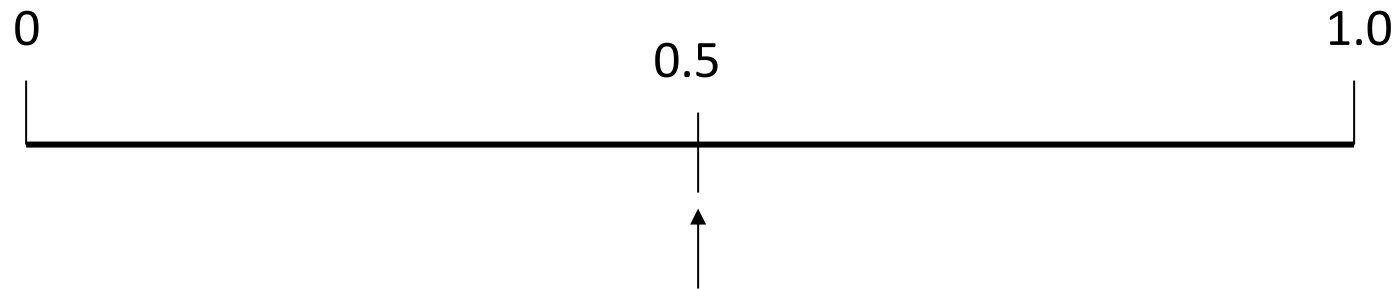
- Probability
- Probability distributions
- Chernoff Bounds

Probability and Statistics



Probability

Probability is the measure of the likelihood that an event will occur.



The occurrence of the event is just as likely as it is unlikely

Probability

- *Probability* is a measure of the likelihood of a random phenomenon or chance behavior. Probability describes the long-term proportion with which a certain outcome will occur in situations with short-term uncertainty.
- Probability is expressed in numbers between 0 and 1. Probability = 0 means the event never happens; probability = 1 means it always happens.
- The total probability of all possible event always sums to 1.

Probability

The probability of an event equals the number of times it happens divided by the number of opportunities.

These numbers can be determined by experiment or by knowledge of the system.

For instance, rolling a die (singular of dice). The chance of rolling a 2 is $1/6$, because there is a 2 on one face and a total of 6 faces. So, assuming the die is balanced, a 2 will come up 1 time in 6.

It is also possible to determine probability by experiment: if the die were unbalanced (loaded = cheating), you could roll it hundreds or thousands of times to get the actual probability of getting a 2. For a fair die, the experimentally determined number should be quite close to $1/6$, especially with many rolls.

Assigning Probabilities

- *Classical approach*: make certain assumptions (such as equally likely, independence) about situation.
- *Relative frequency*: assigning probabilities based on experimentation or historical data.
- *Subjective approach*: Assigning^{0.5} probabilities based on the assignor's judgment. [Bayesian]

Sample Space

- Coin Toss = {head, tail}
- Two coins $S = \{HH, HT, TH, TT\}$
- Inspecting a part = {good, bad}
- Rolling a die $S = \{1, 2, 3, 4, 5, 6\}$

Probability

1. The probability of any event E , $P(E)$, must be between 0 and 1 inclusive. That is,

$$0 \leq P(E) \leq 1.$$

2. If an event is **impossible**, the probability of the event is 0.
3. If an event is a **certainty**, the probability of the event is 1.
4. If $S = \{e_1, e_2, \dots, e_n\}$, then

$$P(e_1) + P(e_2) + \dots + P(e_n) = 1.$$

Unions and Intersections

AND Rule of Probability

The probability of 2 independent events both happening is the product of their individual probabilities.

Called the AND rule because “this event happens AND that event happens”.

For example, what is the probability of rolling a 2 on one die and a 2 on a second die? For each event, the probability is $1/6$, so the probability of both happening is $1/6 \times 1/6 = 1/36$.

Note that the events have to be independent.

Probability - The OR Rule of Probability

- The probability that either one of 2 different events will occur is the sum of their separate probabilities.
- For example, the chance of rolling either a 2 or a 3 on a die is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Joint Probability

- For events A and B, **joint probability** $\Pr(AB)$ stands for the probability that both events happen.
- Example: $A=\{HH\}$, $B=\{HT, TH\}$, what is the joint probability $\Pr(AB)$?

Independence

- Two events *A and B are independent* in case
$$\Pr(AB) = \Pr(A)\Pr(B)$$
- A set of events $\{A_i\}$ is independent in case

$$\Pr(\bigcap_i A_i) = \prod_i \Pr(A_i)$$

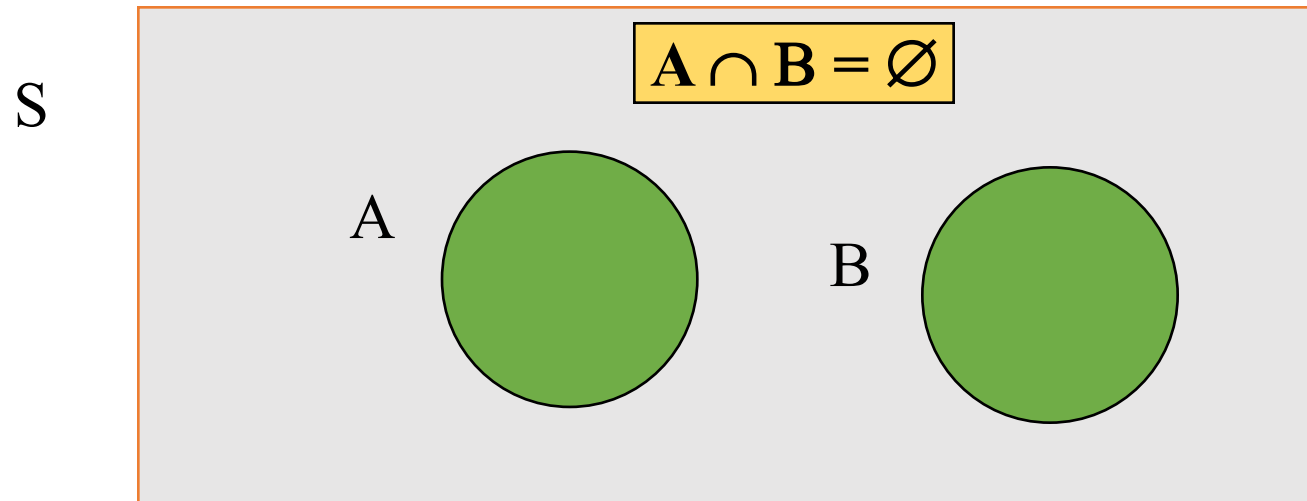
Conditional Probability

- If A and B are events with $\Pr(A) > 0$, the *conditional probability of B given A* is

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

Mutually Exclusive Events

- The OR Rule of Probability
- Mutually exclusive events-no outcomes from S in common



Probability - The OR Rule of Probability

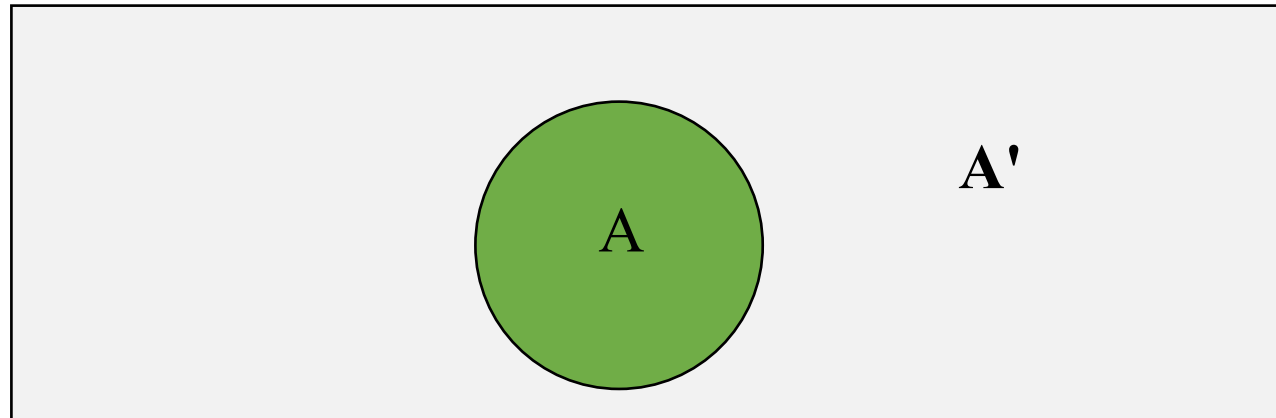
- The probability that either one of 2 different events will occur is the sum of their separate probabilities.
- For example, the chance of rolling either a 2 or a 3 on a die is $\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$.

Probability - NOT Rule

- The chance of an event not happening is 1 minus the chance of it happening.
- For example, the chance of not getting a 2 on a die is $1 - 1/6 = 5/6$.
- This rule can be very useful. Sometimes complicated problems are greatly simplified by examining them backwards.

$$P(A') = 1 - P(A)$$

For an event A , A' is the **complement of A** ; A' is everything in S that is not in A .



Probability

- if A and B are mutually exclusive events:

$$P(A \text{ or } B) = P(A) + P(B)$$

ex., die roll: $P(1 \text{ or } 6) = 1/6 + 1/6 = .33$

- possibility set:

sum of all possible outcomes

$\sim A$ = anything other than A

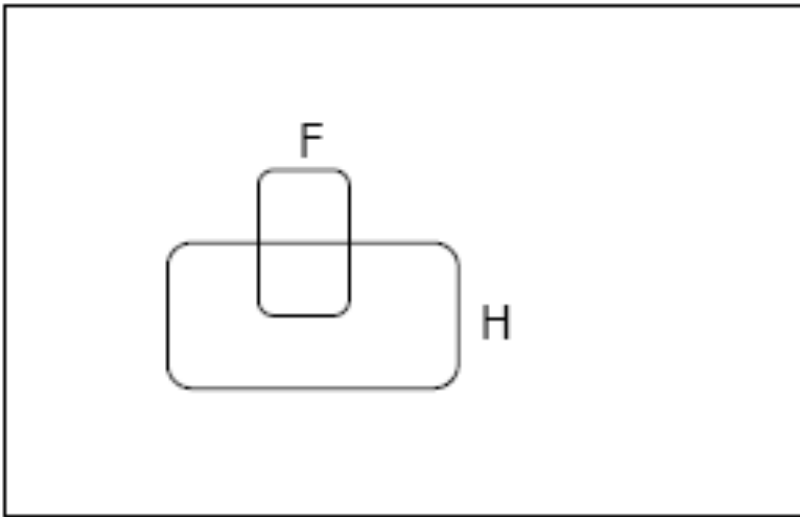
$$P(A \text{ or } \sim A) = P(A) + P(\sim A) = 1$$

Probability

- one event has no influence on the outcome of another event
- if events A & B are independent
then $P(A \& B) = P(A) * P(B)$
- if $P(A \& B) = P(A) * P(B)$
then events A & B are independent
- coin flipping
if $P(H) = P(T) = .5$ then
 $P(HTHTH) = P(HHHHH) =$
 $.5 * .5 * .5 * .5 * .5 = .5^5 = .03$

Conditional Probability

$P(F | H)$ = Fraction of worlds in which H is true that also have F true



$$p(f | h) = \frac{p(F \cap H)}{p(H)}$$

The Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Where do these axioms come from? Were they “discovered”?
Answers coming up later.

Theorems from the Axioms

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(\text{not } A) = P(\sim A) = 1 - P(A)$$

Another important theorem

- $0 \leq P(A) \leq 1$, $P(\text{True}) = 1$, $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(A) = P(A \wedge B) + P(A \wedge \sim B)$$

Multivalued Random Variables

- Suppose A can take on more than 2 values
- A is a *random variable with arity k* if it can take on exactly one value out of $\{v_1, v_2, \dots, v_k\}$
- Thus...

$$P(A = v_i \wedge A = v_j) = 0 \text{ if } i \neq j$$

$$P(A = v_1 \vee A = v_2 \vee \dots \vee A = v_k) = 1$$

A note about independence

- Assume A and B are Boolean Random Variables. Then

“A and B are independent”

if and only if

$$P(A | B) = P(A)$$

- “A and B are independent” is often notated as

$$A \perp B$$

Independence Theorems

- Assume $P(A|B) = P(A)$
- Then $P(A \wedge B) =$

$$= P(A) P(B)$$

- Assume $P(A|B) = P(A)$
- Then $P(B|A) =$

$$= P(B)$$

Independence Theorems

- Assume $P(A|B) = P(A)$
- Then $P(\sim A|B) =$

$$= P(\sim A)$$

- Assume $P(A|B) = P(A)$
- Then $P(A|\sim B) =$

$$= P(A)$$

Multivalued Independence

For multivalued Random Variables A and B,

$$A \perp B$$

if and only if

$$\forall u, v : P(A = u \mid B = v) = P(A = u)$$

from which you can then prove things like...

$$\forall u, v : P(A = u \wedge B = v) = P(A = u)P(B = v)$$

$$\forall u, v : P(B = v \mid A = u) = P(B = v)$$

Expectation

$$E[X] = \sum_{j=0}^{\infty} j \Pr[X = j]$$

- Expectation. Given a discrete random variables X , its expectation $E[X]$ is defined by:

$$E[X] = \sum_{j=0}^{\infty} j \cdot \Pr[X = j] = \sum_{j=0}^{\infty} j (1-p)^{j-1} p = \frac{p}{1-p} \sum_{j=0}^{\infty} j (1-p)^j = \frac{p}{1-p} \cdot \frac{1-p}{p^2} = \frac{1}{p}$$

- Waiting for a first success. Coin is heads with probability p and tails with probability $1-p$. How many independent flips X until first heads?

Expectation: Two Properties

Useful property. If X is a 0/1 random variable, $E[X] = \Pr[X = 1]$.

$$E[X] = \sum_{j=0}^{\infty} j \cdot \Pr[X = j] = \sum_{j=0}^1 j \cdot \Pr[X = j] = \Pr[X = 1]$$

Pf.

not necessarily independent



Linearity of expectation. Given two random variables X and Y defined over the same probability space, $E[X + Y] = E[X] + E[Y]$.

- **Decouples** a complex calculation into simpler pieces.

Guessing Cards

- Game. Shuffle a deck of n cards; turn them over one at a time; try to guess each card.
- Memoryless guessing. No psychic abilities; can't even remember what's been turned over already. Guess a card from full deck uniformly at random.
- Claim. The expected number of correct guesses is 1.
- Pf. (surprisingly effortless using linearity of expectation)
 - Let $X_i = 1$ if i^{th} prediction is correct and 0 otherwise.
 - Let $X =$ number of correct guesses $= X_1 + \dots + X_n$.
 - $E[X_i] = \Pr[X_i = 1] = 1/n$.
 - $E[X] = E[X_1] + \dots + E[X_n] = 1/n + \dots + 1/n = 1$. ■

Guessing Cards

- Game. Shuffle a deck of n cards; turn them over one at a time; try to guess each card.
- Guessing with memory. Guess a card uniformly at random from cards not yet seen.
- Claim. The expected number of correct guesses is $\Theta(\log n)$.
- Pf.
 - Let $X_i = 1$ if i^{th} prediction is correct and 0 otherwise.
 - Let $X =$ number of correct guesses $= X_1 + \dots + X_n$.
 - $E[X_i] = \Pr[X_i = 1] = 1 / (n - i - 1)$.
 - $E[X] = E[X_1] + \dots + E[X_n] = 1/n + \dots + 1/2 + 1/1 = H(n)$. ■

linearity of expectation

$$\ln(n+1) < H(n) < 1 + \ln n$$

Coupon Collector

- Coupon collector. Each box of cereal contains a coupon. There are n different types of coupons. Assuming all boxes are equally likely to contain each coupon, how many boxes before you have ≥ 1 coupon of each type?

- Claim. The expected number of steps is $\Theta(n \log n)$.

- Pf.
$$E[X] = \sum_{j=0}^{n-1} E[X_j] = \sum_{j=0}^{n-1} \frac{n}{n-j} = n \sum_{i=1}^n \frac{1}{i} = nH(n)$$
 - Phase j = time between j and $j+1$ distinct coupons.
 - Let X_j = number of steps you spend in phase j .
Proof of Success: High \Rightarrow expected waiting time = $n/(n-j)$
 - Let X = number of steps in total = $X_0 + X_1 + \dots + X_{n-1}$.

13.4 MAX 3-SAT

Maximum 3-Satisfiability

exactly 3 distinct literals per clause

- MAX-3SAT. Given 3-SAT formula, find a truth assignment that satisfies as many clauses as possible.

$$C_1 = x_2 \vee x_3 \vee x_4$$

$$C_2 = x_2 \vee x_3 \vee \overline{x_4}$$

$$C_3 = \overline{x_1} \vee x_2 \vee x_4$$

$$C_4 = \overline{x_1} \vee \overline{x_2} \vee x_3$$

$$C_5 = x_1 \vee \overline{x_2} \vee \overline{x_4}$$

- Remark. NP-hard search problem.
- Simple idea. Flip a coin, and set each variable true with probability $\frac{1}{2}$, independently for each variable.

Maximum 3-Satisfiability: Analysis

- Claim. Given a 3-SAT formula with k clauses, the expected number of clauses satisfied by a random assignment is $7k/8$.

- Pf. Consider random variable

$$E[Z] = \sum_{j=1}^k E[Z_j]$$

linearity of expectation

- Let Z = weight of clauses satisfied by assignment Z_j .

$$= \frac{7}{8} k$$

The Probabilistic Method

- Corollary. For any instance of 3-SAT , **there exists** a truth assignment that satisfies at least a $7/8$ fraction of all clauses.
- Pf. Random variable is at least its expectation some of the time. ■
- Probabilistic method. We showed the existence of a non-obvious property of 3-SAT by showing that a random construction produces it with positive probability!

Maximum 3-Satisfiability: Analysis

- Q. Can we turn this idea into a $7/8$ -approximation algorithm? In general, a random variable can almost always be below its mean.
- Lemma. The probability that a random assignment satisfies $\geq 7k/8$ clauses is at least $1/(8k)$.
- Pf. Let p_j be probability that exactly j clauses are satisfied; let p be probability that $\geq 7k/8$ clauses are satisfied.

$$\begin{aligned}\frac{7}{8}k &= E[Z] = \sum_{j \geq 0} j p_j \\ &= \sum_{j < 7k/8} j p_j + \sum_{j \geq 7k/8} j p_j \\ &\leq \left(\frac{7k}{8} - \frac{1}{8}\right) \sum_{j < 7k/8} p_j + k \sum_{j \geq 7k/8} p_j \\ &\leq \left(\frac{7}{8}k - \frac{1}{8}\right) \cdot 1 + k p\end{aligned}$$

- Rearranging terms yields $p \geq 1 / (8k)$. ■

Maximum 3-Satisfiability: Analysis

- Johnson's algorithm. Repeatedly generate random truth assignments until one of them satisfies $\geq 7k/8$ clauses.
- Theorem. Johnson's algorithm is a $7/8$ -approximation algorithm.
- Pf. By previous lemma, each iteration succeeds with probability at least $1/(8k)$. By the waiting-time bound, the expected number of trials to find the satisfying assignment is at most $8k$. ■

Maximum Satisfiability

- Extensions.
 - Allow one, two, or more literals per clause.
 - Find max **weighted** set of satisfied clauses.
- Theorem. **[Asano-Williamson 2000]** There exists a 0.784-approximation algorithm for MAX-SAT.
- Theorem. **[Karloff-Zwick 1997, Zwick+computer 2002]** There exists a 7/8-approximation algorithm for version of MAX-3SAT where each clause has **at most** 3 literals.
- Theorem. **[Håstad 1997]** Unless $P = NP$, no ρ -approximation algorithm for MAX-3SAT (and hence MAX-SAT) for any $\rho > 7/8$.
very unlikely to improve over simple randomized algorithm for MAX-3SAT

Monte Carlo vs. Las Vegas Algorithms

- Monte Carlo algorithm. Guaranteed to run in poly-time, likely to find correct answer.
- Ex: Contraction algorithm for global min cut.

- Las Vegas algorithm. Guaranteed to find correct answer, likely to run in poly-time.
- Ex: Randomized quicksort, Johnson's MAX-3SAT algorithm.

stop algorithm after a certain point



- Remark. Can always convert a Las Vegas algorithm into Monte Carlo, but no known method to convert the other way.

Random variables

Random variables assign a real number to each outcome:

$$\begin{aligned} X : \Omega &\rightarrow \mathbb{R} \\ \omega &\rightarrow X(\omega) \end{aligned}$$

Random variables can be:

Discrete: if it takes at most countably many values (integers).

Continuous: if it can take any real number.

Random variables

Distribution of a random variable $F(x) = F_X(x) = P(X \leq x)$

(i) $F(x) \rightarrow 0$ when $x \rightarrow -\infty$

(ii) $F(x) \rightarrow 1$ when $x \rightarrow +\infty$

(iii) $F(x)$ is nondecreasing.

$$x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$$

(iv) $F(x)$ is right-continuous.

$$F(x) \rightarrow F(x_0) \quad \text{when} \quad \begin{array}{l} x \rightarrow x_0 \\ x > x_0 \end{array}$$

Random variables

- For a random variable, we define
 - Probability function
 - Density function,
- depending on whether it is discrete or continuous

Random variables

Probability function

$$p(x) = p_X(x) = P(X = x)$$

verifies

$$(i) \ p(x) \geq 0$$

$$(ii) \ \sum_x p(x) = 1$$

Random variables

Probability density function

$$f(x)$$

verifies

$$(i) \quad f(x) \geq 0$$

$$(ii) \quad \int_{-\infty}^{+\infty} f(x)dx = 1$$

We have

$$F(x) = \int_{-\infty}^x f(t)dt \quad \text{and} \quad f(x) = F'(x).$$

Random variables

F completely determines the distribution of a random variable.

$$P(a < X \leq b) = F(b) - F(a) = \begin{cases} \sum_{a < x \leq b} p(x) \\ \int_a^b f(t) dt \end{cases}$$

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
 - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with arity k if it can take on exactly one value out of $\{x_1, \dots, x_k\}$
 - E.g. the possible values that X can take on are 0, 1, 2, ..., 100

Probability of Discrete RV

- Probability mass function (pmf): $P(X = x_i)$
- Easy facts about pmf
 - $\sum_i P(X = x_i) = 1$
 - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
 - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
 - $P(X = x_1 \cup X = x_2 \cup \dots \cup X = x_k) = 1$

Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x .

Probability of Continuous RV

Properties of pdf

$$f(x) \geq 0, \forall x$$

$$\int_{-\infty}^{+\infty} f(x) = 1$$

Actual probability can be obtained by taking the integral of pdf

E.g. the probability of X being between 0 and 1 is

$$P(0 \leq X \leq 1) = \int_0^1 f(x) dx$$

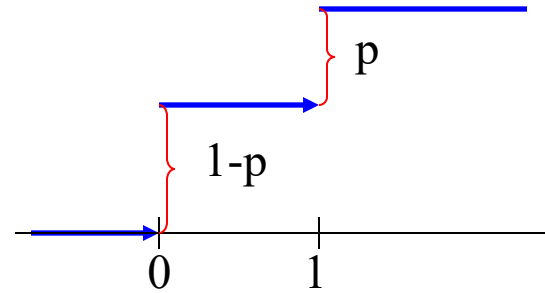
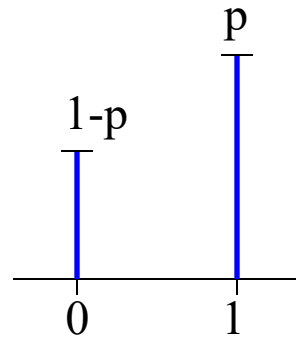
Random variables

Bernoulli

$$X \equiv B(1, p)$$

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p$$



Random variables

Binomial

Successes in n independent Bernoulli trials with success probability p

$$X \equiv B(n, p)$$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, 2, \dots, n$$

$$\text{with } \binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Random variables

Geometric

Time of first success in a sequence of independent Bernoulli trials with success probability p

$$X \equiv G(p)$$

$$P(X = x) = (1 - p)^{x-1} \cdot p \quad x = 1, 2, 3, \dots$$

Random variables

Poisson

X expresses the number of “rare events”

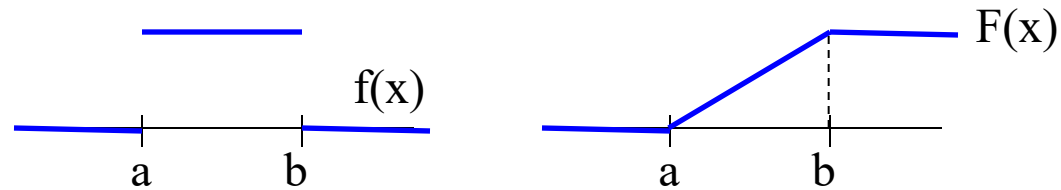
$$X \equiv P(\lambda), \quad \lambda > 0$$

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x = 0, 1, 2, \dots$$

Random variables

Uniform

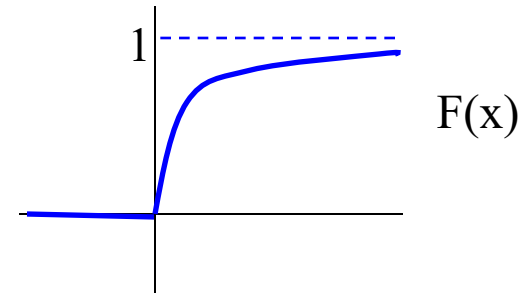
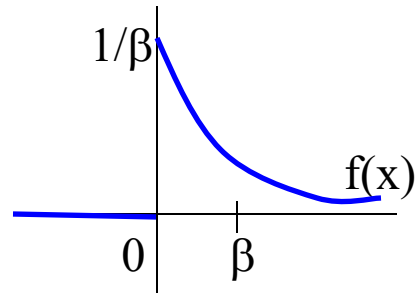
$$X \equiv U[a, b] \quad f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a < x < b \\ 0 & \text{otherwise} \end{cases}$$
$$F(x) = \begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } a \leq x < b \\ 1 & \text{for } x \geq b \end{cases}$$



Random variables

Exponential

$$X \equiv \exp(\beta) \quad f(x) = \begin{cases} \frac{1}{\beta} e^{\frac{-x}{\beta}} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$
$$F(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 - e^{\frac{-x}{\beta}} & \text{for } x \geq 0 \end{cases}$$



Random variables

Normal

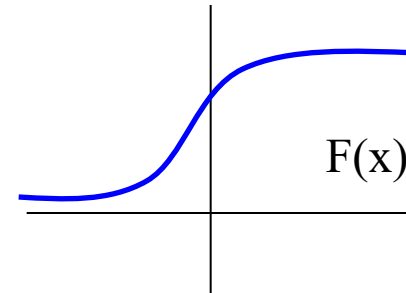
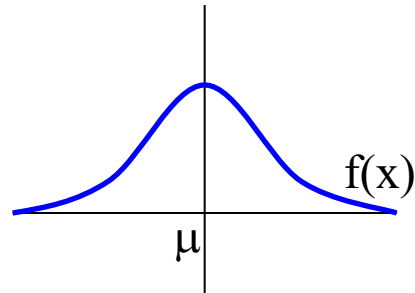
$$X \equiv N(\mu, \sigma^2)$$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$x \in \mathbb{R}$$

$$\mu \in \mathbb{R}$$

$$\sigma^2 \geq 0$$



Random variables

Properties of normal distribution

$$(i) \quad \frac{X - \mu}{\sigma} \equiv N(0,1) \quad \text{standard normal}$$

$$(ii) \quad Z \equiv N(0,1) \Rightarrow \sigma Z + \mu \equiv N(\mu, \sigma^2)$$

$$(iii) \quad \begin{aligned} &X_i \equiv N(\mu_i, \sigma_i^2) \\ &\Rightarrow \sum_i X_i \equiv N\left(\sum_i \mu_i, \sum_i \sigma_i^2\right) \quad \text{independent } i=1,2,\dots,n \end{aligned}$$

Random variables

Two random variables X and Y are independent if and only if:

$$p(x, y) = p_X(x)p_Y(y)$$

$$f(x, y) = f_X(x)f_Y(y),$$

for all values x and y .

Random variables

Discrete variables

$$p(x | y) = P(X = x | Y = y) = \frac{p(x, y)}{p(y)}$$

Continuous variables

$$f(x | y) = \frac{f(x, y)}{f(y)}$$

If X and Y are independent:

$$p(x | y) = p(x)$$

$$f(x | y) = f(x)$$

Random variables

$$EX = \mu_X = \sum_x xp(x)$$

$$EX = \mu_X = \int xf(x)dx$$

Properties:

$$(i) \quad E \sum_i \alpha_i X_i = \sum_i \alpha_i EX_i \quad i = 1, \dots, n$$

(ii) If $X_i, i = 1, \dots, n$ are independent then:

$$E \prod_i X_i = \prod_i EX_i$$

Random variables

Moment of order k

$$EX^k = \sum_x x^k p(x)$$

$$EX^k = \int x^k f(x) dx$$

Random variables

Variance

Given X with $\mu = EX$

$$VX = \sigma_X^2 = E(X - \mu)^2$$

$$\sigma_X = \sqrt{VX} = (E(X - \mu)^2)^{1/2}$$

standard deviation

Permutations

A B C D E

- How many ways can we choose 2 letters from the above 5, without replacement, when the order in which we choose the letters is important?
- $\underline{5} \times \underline{4} = 20$

Permutations (cont.)

$$\underline{5} \times \underline{4} = 20 = \frac{5!}{(5-2)!} = \frac{5!}{3!} = 5 \times 4$$

$$\textit{Notation} : {}_5P_2 = \frac{5!}{(5-2)!} = 20$$

Combinations

A B C D E

- How many ways can we choose 2 letters from the above 5, without replacement, when the order in which we choose the letters is not important?
- $\underline{5} \times \underline{4} = 20$ when order important
- Divide by 2: $(5 \times 4)/2 = 10$ ways
- N choose K (5 choose 2)

Bounding Numbers of Combinations

$$\binom{n}{k} = \frac{n!}{k! (n-k)!}$$

= number of (unordered)
combinations of n objects
taken k at a time

- N choose K

$$\binom{n}{k} \sim \frac{n^k e^{-\frac{k^2}{2n} - \frac{k^3}{6n^2}}}{k!} (1 - o(1)) \quad \text{for } k = o\left(n^{\frac{3}{4}}\right)$$

Combinations (cont.)

$$\binom{5}{2} = {}_5C_2 = \frac{5!}{(5-2)!2!} = \frac{5!}{3!2!} = \frac{5 \times 4}{1 \times 2} = \frac{20}{2} = 10$$

$$\binom{n}{r} = {}_nC_r = \frac{n!}{(n-r)!r!}$$

Tail Bounds

- In the analysis of randomized algorithms, we need to know how much does an algorithms run-time/cost deviate from its expected run-time/cost.
- That is we need to find an upper bound on $\Pr[X \text{ deviates from } E[X] \text{ a lot}]$. This we refer to as the tail bound on X .

Markov and Chebychev Inequalities

- Markov Inequality (uses only mean)

$$\text{Prob} (A \geq x) \leq \frac{\mu}{x}$$

- Chebychev Inequality (uses mean and variance)

$$\text{Prob} (|A - \mu| \geq \Delta) \leq \frac{\sigma^2}{\Delta^2}$$

Markov and Chebychev Inequalities

- Example, if B is a Binomial with parameters n,p

$$\text{Then Prob } (B \geq x) \leq \frac{np}{x}$$

$$\text{Prob } (|B - np| \geq \Delta) \leq \frac{np(1-p)}{\Delta^2}$$

Chernoff bounds

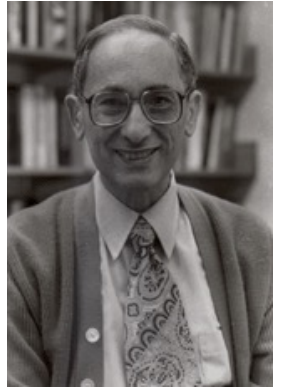
The Chernoff bound for a random variable X is obtained as follows: for any $t > 0$,

$$\Pr[X \geq a] = \Pr[e^{tX} \geq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

Similarly, for any $t < 0$,

$$\Pr[X \leq a] = \Pr[e^{tX} \leq e^{ta}] \leq E[e^{tX}] / e^{ta}$$

The value of t that minimizes $E[e^{tX}] / e^{ta}$ gives the best possible bounds.



Chernoff Bound of Random Variable A

- Uses **all** moments
- Uses **moment generating function**

$$\begin{aligned}\text{Prob } (A \geq x) &\leq e^{-sx} M_A(s) \text{ for } s \geq 0 \\ &= e^{\gamma(s) - sx} \text{ where } \gamma(s) = \ln (M_A(s)) \\ &\leq e^{\gamma(s) - s \gamma'(s)}\end{aligned}$$

By setting $x = \gamma'(s)$
1st derivative minimizes bounds

Chernoff Bound of Discrete Random Variable A

$$\text{Prob}(A \geq x) \leq z^{-x} G_A(z) \quad \text{for } z \geq 1$$

- Choose $z = z_0$ to **minimize** above bound
- Need Probability Generating function

$$G_A(z) = \sum_{x \geq 0} z^x f_A(x) = E(z^A)$$

Chernoff Bounds for Binomial B with parameters n,p

- Above mean $x \geq \mu$

Prob ($B \geq x$)

$$\leq \left(\frac{n-\mu}{n-x} \right)^{n-x} \left(\frac{\mu}{x} \right)^x$$

$$\leq e^{x-\mu} \left(\frac{\mu}{x} \right)^x \text{ since } \left(1 - \frac{1}{x} \right)^x < e^{-1}$$

$$\leq e^{-x - \mu} \text{ for } x \geq \mu e^2$$

Chernoff Bounds for Binomial B with parameters n,p

- Below mean $x \leq \mu$

Prob ($B \leq x$)

$$\leq \left(\frac{n-\mu}{n-x} \right)^{n-x} \left(\frac{\mu}{x} \right)^x$$

Load Balancing

- Load balancing. System in which m jobs arrive in a stream and need to be processed immediately on n identical processors. Find an assignment that balances the workload across processors.
- Centralized controller. Assign jobs in round-robin manner. Each processor receives at most $\lceil m/n \rceil$ jobs.
- Decentralized controller. Assign jobs to processors uniformly at random. How likely is it that some processor is assigned "too many" jobs?

Load Balancing

- Analysis.

- Let X_i = number of jobs assigned to processor i .
- Let $Y_{ij} = 1$ if job j assigned to processor i , and 0 otherwise.
- We have $E[Y_{ij}] = 1/n$
- Thus, $X_i = \sum_j Y_{ij}$, and $\mu = E[X_i] = 1$.
- Applying Chernoff bounds with $\delta = c - 1$ yields

$$\Pr[X_i > c] < \frac{e^{c-1}}{c^c}$$

- Let $\gamma(n)$ be number x such that $\frac{e^{c-1}}{c^c} = \frac{1}{n^x}$, and choose $c = e \gamma(n)$.

$$\Pr[X_i > c] < \frac{e^{c-1}}{c^c} < \left(\frac{e}{c}\right)^c = \left(\frac{1}{\gamma(n)}\right)^{e\gamma(n)} < \left(\frac{1}{\gamma(n)}\right)^{2\gamma(n)} = \frac{1}{n^2}$$

- Union bound \Rightarrow with probability $\geq 1 - 1/n$ no processor receives more than $e \gamma(n) = \Theta(\log n / \log \log n)$ jobs.



Fact: this bound is asymptotically tight: with high probability, some processor receives $\Theta(\log n / \log \log n)$ jobs.

Load Balancing: Many Jobs

- Theorem. Suppose the number of jobs $m = 16n \ln n$. Then on average, each of the n processors handles $\mu = 16 \ln n$ jobs. With high probability every processor will have between half and twice the average load.
- Pf.
 - Let X_i, Y_{ij} be as before.
 - Applying Chernoff bounds with $\delta = 1/2$ yields $\Pr[X_i > 2\mu] < \left(\frac{e}{4}\right)^{16n \ln n}$ and $\Pr[X_i < \frac{1}{2}\mu] < e^{-\frac{1}{2}(\frac{1}{2})^2(16n \ln n)} = \frac{1}{n^2}$
- Union bound \Rightarrow every processor has load between half and twice the average with probability $\geq 1 - 2/n$. ■

Birthday Problem

- What is the smallest number of people you need in a group so that the probability of **2 or more** people having the same birthday is **greater than $1/2$** ?
- Answer: **23**

No. of people	23	30	40	60
Probability	.507	.706	.891	.994



Birthday Problem

- $A = \{\text{at least 2 people in the group have a common birthday}\}$
- $A' = \{\text{no one has common birthday}\}$

$$3 \text{ people} : P(A') = \frac{364}{365} \times \frac{363}{365}$$

23 people :

$$P(A') = \frac{364}{365} \times \frac{363}{365} \times \dots \times \frac{343}{365} = .498$$

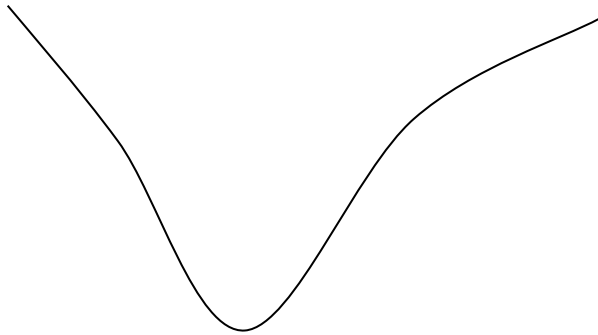
$$\text{so } P(A) = 1 - P(A') = 1 - .498 = .502$$

Gradient Descent

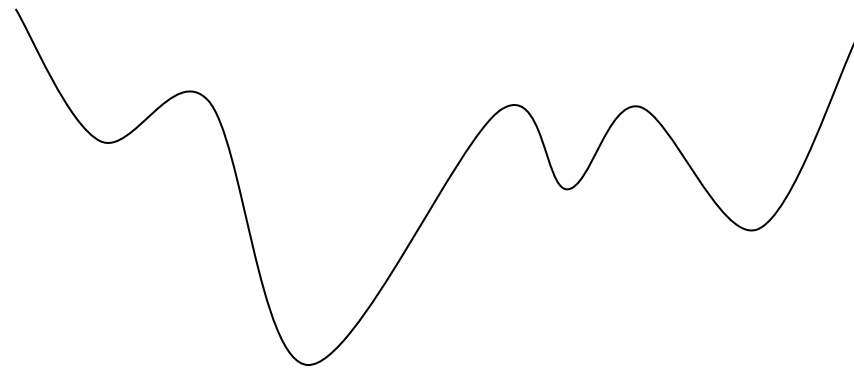
- Perceptron training rule may not converge if points are not linearly separable
- Gradient descent will try to fix this by changing the weights by the total error for all training points, rather than the individual
- If the data is not linearly separable, then it will converge to the best fit

Local Search

- Local search. Algorithm that explores the space of possible solutions in sequential fashion, moving from a current solution to a "nearby" one.
- Gradient descent. Let S denote current solution. If there is a neighbor S' of S with strictly lower cost, replace S with the neighbor whose cost is as small as possible. Otherwise, terminate the algorithm.



A funnel



A jagged funnel

Metropolis Algorithm

- Metropolis algorithm. [\[Metropolis, Rosenbluth, Rosenbluth, Teller, Teller 1953\]](#)
 - Simulate behavior of a physical system according to principles of statistical mechanics.
 - Globally biased toward "downhill" steps, but occasionally makes "uphill" steps to break out of local minima.
- Gibbs-Boltzmann function. The probability of finding a physical system in a state with energy E is proportional to $e^{-E/(kT)}$, where $T > 0$ is temperature and k is a constant.
 - For any temperature $T > 0$, function is monotone decreasing function of energy E .
 - System more likely to be in a lower energy state than higher one.
 - T large: high and low energy states have roughly same probability
 - T small: low energy states are much more probable

Metropolis Algorithm

- Metropolis algorithm.
 - Given a fixed temperature T , maintain current state S .
 - Randomly perturb current state S to new state $S' \in N(S)$.
 - If $E(S') \leq E(S)$, update current state to S'
Otherwise, update current state to S' with probability $e^{-\Delta E / (kT)}$, where $\Delta E = E(S') - E(S) > 0$.
- Theorem. Let $f_S(t)$ be fraction of first t steps in which simulation is in state S . Then, assuming some technical conditions, with probability 1:

$$\lim_{t \rightarrow \infty} f_S(t) = \frac{1}{Z} e^{-E(S) / (kT)},$$

$$\text{where } Z = \sum_{S \in N(S)} e^{-E(S) / (kT)}.$$

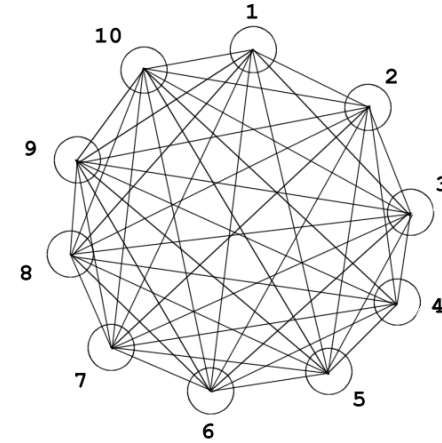
- Intuition. Simulation spends roughly the right amount of time in each state, according to Gibbs-Boltzmann equation.

Simulated Annealing

- Simulated annealing.
 - T large \Rightarrow probability of accepting an uphill move is large.
 - T small \Rightarrow uphill moves are almost never accepted.
 - Idea: turn knob to control T .
 - Cooling schedule: $T = T(i)$ at iteration i .
- Physical analog.
 - Take solid and raise it to high temperature, we do not expect it to maintain a nice crystal structure.
 - Take a molten solid and freeze it very abruptly, we do not expect to get a perfect crystal either.
 - Annealing: cool material gradually from high temperature, allowing it to reach equilibrium at succession of intermediate lower temperatures.

Hopfield Neural Networks

- Sub-type of recurrent neural nets
 - Fully recurrent
 - Weights are symmetric
 - Nodes can only be *on* or *off*
 - Random updating
- Learning: **Hebb rule** (cells that fire together wire together)
 - Biological equivalent to LTP and LTD
- auto-associative or
content-addressable memory

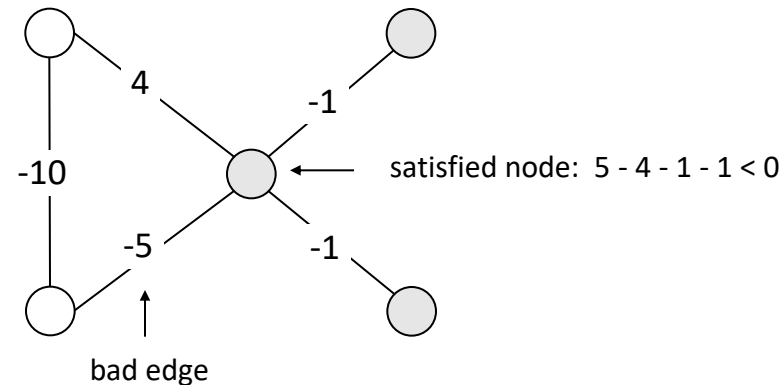


Hopfield Neural Networks

- Def. With respect to a configuration S , edge $e = (u, v)$ is **good** if $w_e s_u s_v < 0$. That is, if $w_e < 0$ then $s_u = s_v$; if $w_e > 0$, $s_u \neq s_v$.

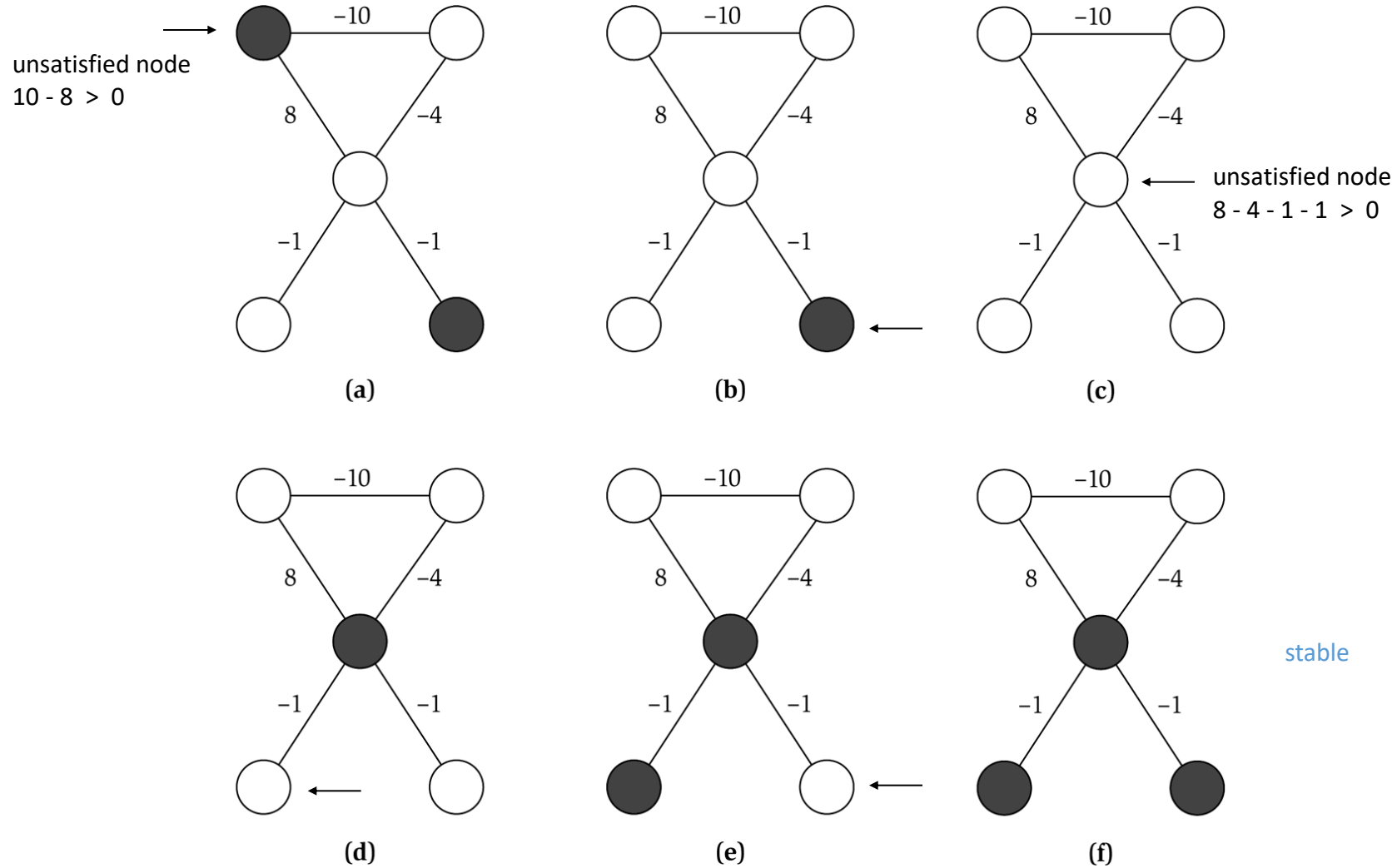
$$\sum_{v: e=(u,v) \in E} w_e s_u s_v \leq 0$$

- Def. With respect to a configuration S , a node u is **satisfied** if the weight of incident good edges \geq weight of incident bad edges.
- Def. A configuration is **stable** if all nodes are satisfied.



- Goal. Find a stable configuration, if such a configuration exists.

State Flipping Algorithm



Statistics, Machine Learning and Data Mining

- Statistics:
 - more theory-based
 - more focused on testing hypotheses
- Machine learning
 - more heuristic
 - focused on improving performance of a learning agent
 - also looks at real-time learning and robotics – areas not part of data mining

Basic Probability Formulas

- Product rule

$$P(A \wedge B) = P(A \mid B)P(B) = P(B \mid A)P(A)$$

- Sum rule

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- Bayes theorem

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Theorem of total probability, if event A_i is mutually exclusive and probability sum to 1

$$P(B) = \sum_{i=1}^n P(B \mid A_i)P(A_i)$$

Bayes Theorem

- Given a hypothesis h and data D which bears on the hypothesis:

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

- $P(h)$: independent probability of h : *prior probability*
- $P(D)$: independent probability of D
- $P(D/h)$: conditional probability of D given h : *likelihood*
- $P(h/D)$: conditional probability of h given D : *posterior probability*

Maximum A Posterior

- Based on Bayes Theorem, we can compute the *Maximum A Posterior* (MAP) hypothesis for the data
- We are interested in the best hypothesis for some space H given observed training data D .

$$\begin{aligned} h_{MAP} &\equiv \operatorname{argmax}_{h \in H} P(h \mid D) \\ &= \operatorname{argmax}_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} \\ &= \operatorname{argmax}_{h \in H} P(D \mid h)P(h) \end{aligned}$$

H : set of all hypothesis.

Note that we can drop $P(D)$ as the probability of the data is constant (and independent of the hypothesis).

Naïve Bayes Background

- There are two main methods for training a classifier:

a) Discriminative Classifiers

Examples: k-NN, decision trees, Neural Networks, SVM

b) Generative Classifiers

Example: Bayesian approaches (naive Bayes...)

Discriminative approach seems easier, as the task is easier; you don't need to model classes (observation distribution of features in those classes), just need to find where the query instance belongs to.

Naïve Bayes Classifier

- The *Naïve Bayes Assumption*: Assume that all features are independent **given the class label Y**
- Equationally speaking:

$$P(X_1, \dots, X_n | Y) = \prod_{i=1}^n P(X_i | Y)$$

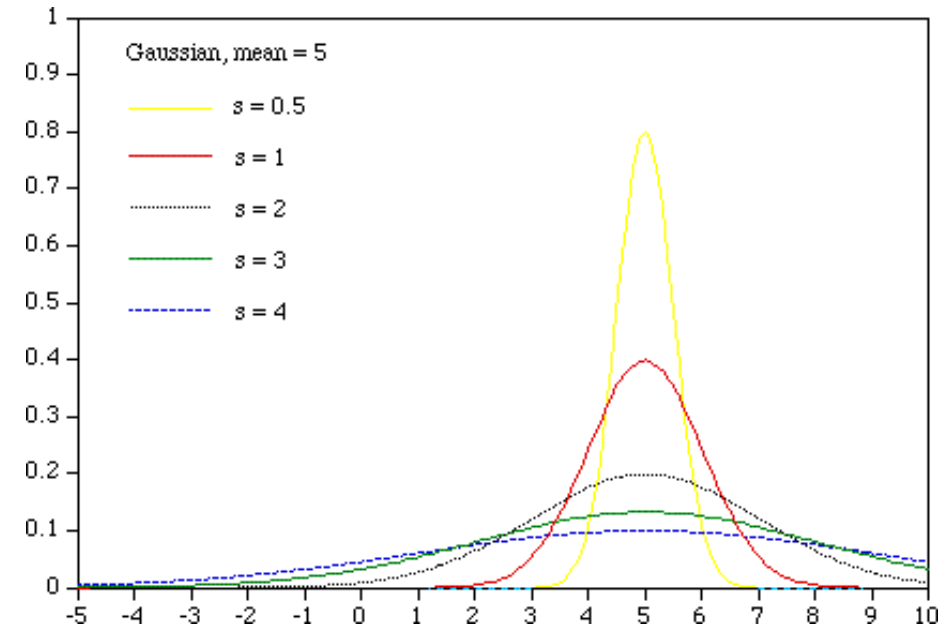
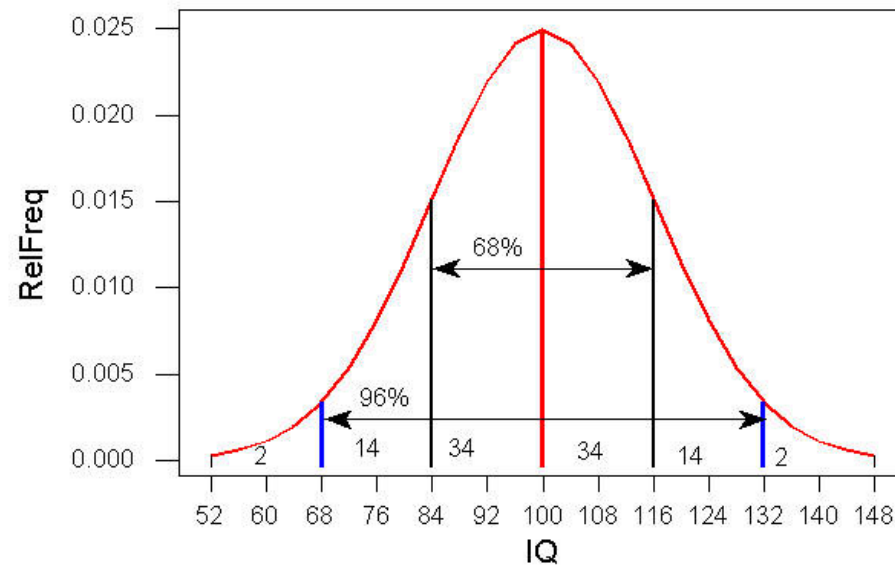
- (We will discuss the validity of this assumption later)

The Normal Distribution (Han & Kamber, 2006)

- The *normal* or *Gaussian* density:
 - applies to continuous, real-valued random variables
 - characterized by mean (average) m and standard deviation s
 - *density* at x is defined as
 - $(1/(s \sqrt{2\pi})) \exp(-(x-m)^2/2s^2)$
 - special case $m = 0, s = 1$: $a \exp(-x^2/b)$ for some constants $a, b > 0$
 - peaks at $x = m$, then dies off *exponentially* rapidly
 - the classic “bell-shaped curve”
 - exam scores, human body temperature,
 - remarks:
 - can control mean and standard deviation independently
 - can make as “broad” as we like, but always have *finite variance*

The Normal Distribution

(Han & Kamber, 2006)



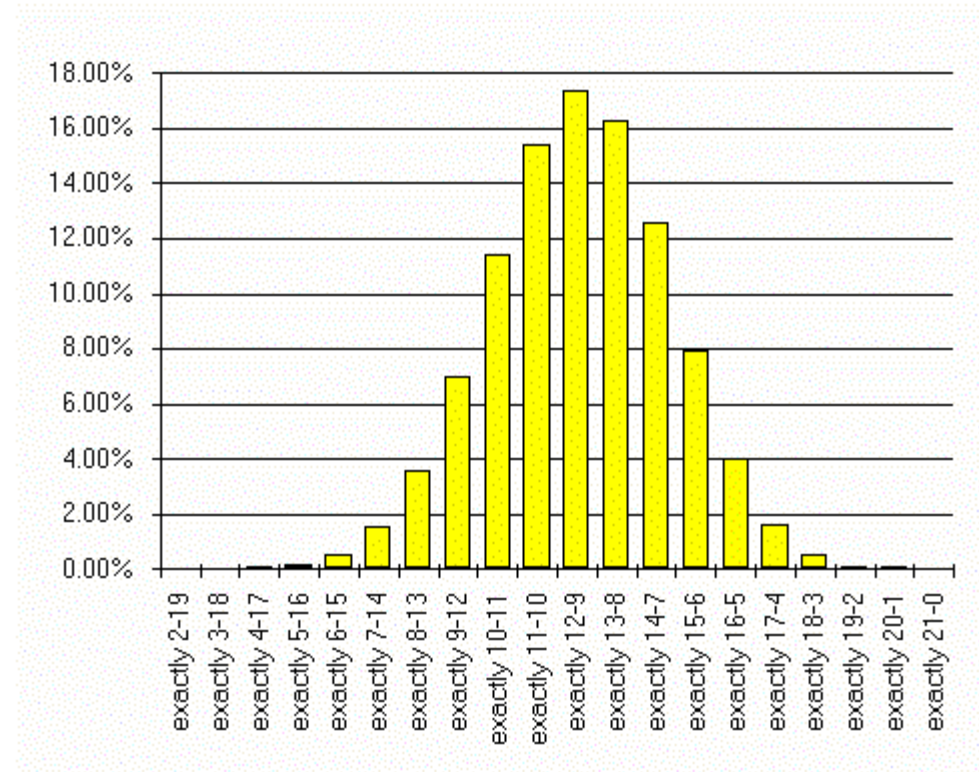
The Binomial Distribution

(Han & Kamber, 2006)

- coin with $\text{Pr}[\text{heads}] = p$, flip n times
- probability of getting exactly k heads:
 - $\text{choose}(n, k) p^k (1-p)^{n-k}$
- for large n and p *fixed*:
 - approximated well by a normal with
$$m = np, s = \sqrt{np(1-p)}$$
 - $s/m \rightarrow 0$ as n grows

The Binomial Distribution

(Han & Kamber, 2006)



[www.professionalgambler.com/ binomial.html](http://www.professionalgambler.com/binomial.html)

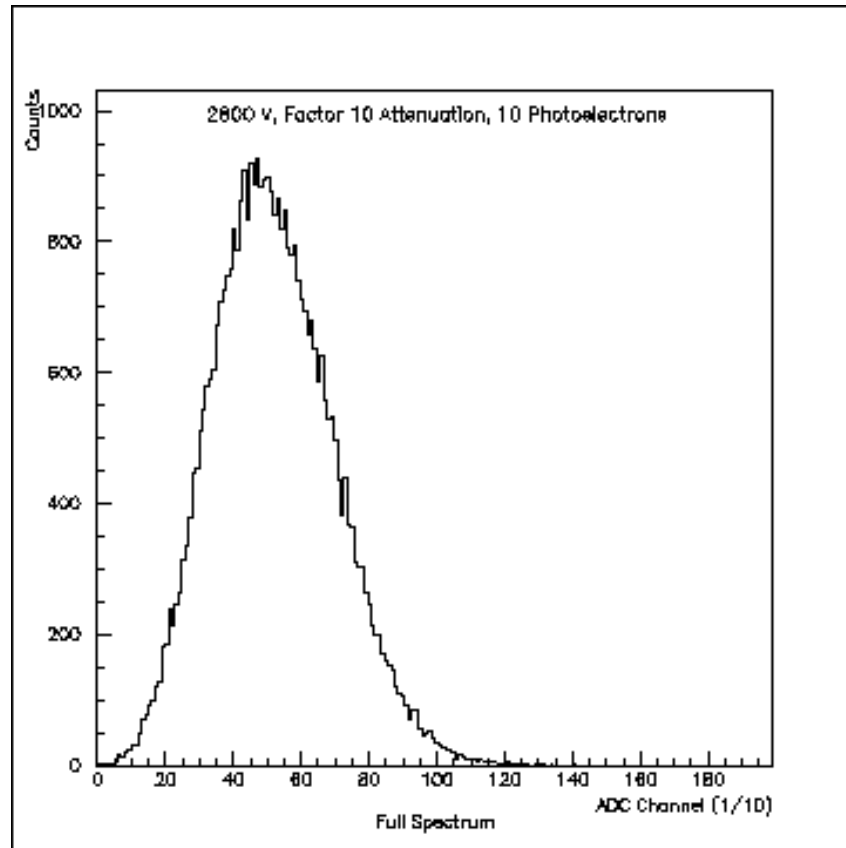
The Poisson Distribution

(Han & Kamber, 2006)

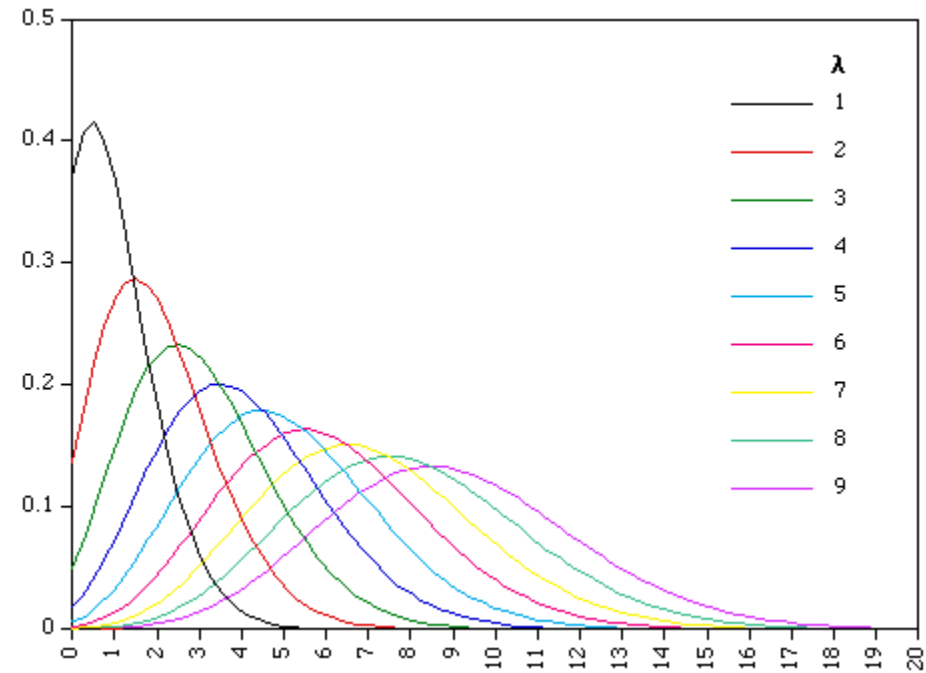
- like binomial, applies to variables taking on integer values > 0
- often used to model *counts* of events
 - number of phone calls placed in a given time period
 - number of times a neuron fires in a given time period
- single free parameter λ
- probability of exactly x events:
 - $\exp(-\lambda) \lambda^x / x!$
 - mean and variance are both λ
- binomial distribution with n large, $p = \lambda/n$ (λ fixed)
 - converges to Poisson with mean λ

The Poisson Distribution

(Han & Kamber, 2006)



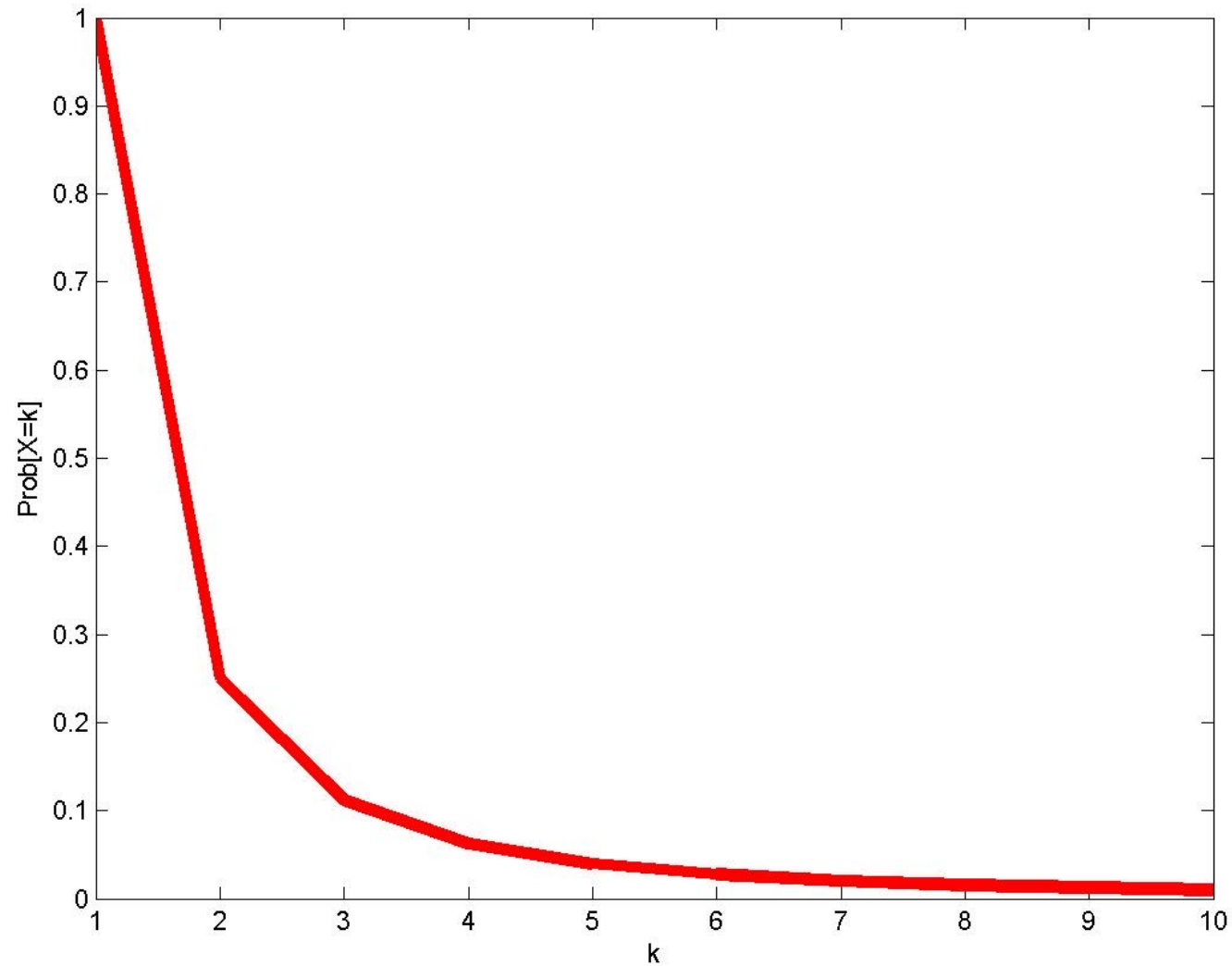
single photoelectron distribution



Fat-Tailed Distributions

- Power law distributions:
 - for variables assuming integer values > 0
 - $\text{Prob}[X=k] \sim Ck^{-\alpha}$
 - typically $0 < \alpha < 2$; smaller α gives heavier tail
- For binomial, normal, and Poisson distributions the tail probabilities approach 0 exponentially fast
- What kind of phenomena does this distribution model?
- What kind of process would generate it?

Power Law: Prob $[X=k] = k^{-2}$



Power Law

$$\text{Prob } [X=k] \sim Ck^{-\alpha}$$

$$\log \text{Prob } [X=k] \sim \log C - \alpha \log k$$

$$\text{Prob } [X=k] = k^{-2}$$

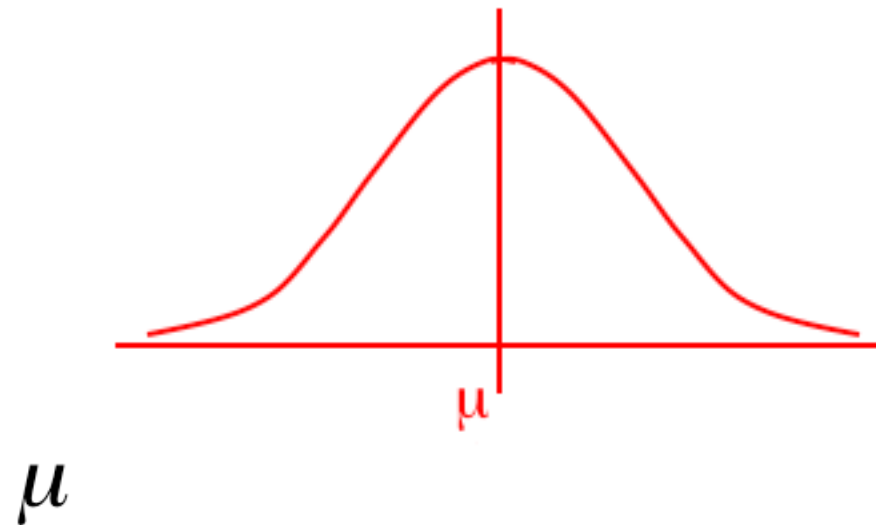
$$\log \text{Prob } [X=k] = -2 \log k$$

The Normal Distribution

- The **normal distribution** is the most important of all probability distributions. The probability density function of a **normal random variable** is given by:


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

- It looks like this:
- Bell shaped,
- Symmetrical around the mean ...



Exponential Distribution

- Another important continuous distribution is the ***exponential distribution*** which has this probability density function:

$$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$


- Note that $x \geq 0$. Time (for example) is a non-negative quantity; the exponential distribution is often used for time related phenomena such as the length of time between phone calls or between parts arriving at an assembly station. Note also that the mean and standard deviation are equal and to the inverse of the parameter of the distribution (lambda)

$$\mu = \sigma = \frac{1}{\lambda}$$

Poisson Distribution

- Distribution often used to model the number of incidences of some characteristic in time or space:
 - Arrivals of customers in a queue
 - Numbers of flaws in a roll of fabric
 - Number of typos per page of text.
- Distribution obtained as follows:
 - Break down the “area” into many small “pieces” (n pieces)
 - Each “piece” can have only 0 or 1 occurrences ($p=P(1)$)
 - Let $\lambda=np \equiv$ Average number of occurrences over “area”
 - $Y \equiv$ # occurrences in “area” is sum of 0^s & 1^s over “pieces”
 - $Y \sim \text{Bin}(n,p)$ with $p = \lambda/n$
 - Take limit of Binomial Distribution as $n \rightarrow \infty$ with $p = \lambda/n$

$$p(y) = P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!} \quad \lambda > 0, \quad y = 0, 1, 2, \dots$$

Negative Binomial Distribution

- Used to model the number of trials needed until the r^{th} Success (extension of Geometric distribution)
- Based on there being $r-1$ Successes in first $y-1$ trials, followed by a Success

$$p(y) = \binom{y-1}{r-1} p^r (1-p)^{y-r} = \frac{(y-1)!}{(r-1)!(y-r)!} p^r (1-p)^{y-r} =$$
$$= \frac{\Gamma(y)}{\Gamma(r)\Gamma(y-r+1)!} p^r (1-p)^{y-r} \quad y = r, r+1, \dots$$

$$E(Y) = \frac{r}{p} \quad V(Y) = \frac{r(1-p)}{p^2}$$

$$\Gamma(a) = \int_0^\infty z^{a-1} e^{-z/a} dz \quad \text{Note : } \Gamma(a) = (a-1)\Gamma(a)$$