Review of probability and statistics

STAT 37710 / CMSC 35300

Rebecca Willett and Yuxin Chen

Basic probability theory

Probability theory begins with three basic components. The set of all possible outcomes, denoted Ω. The collection of all sets of outcomes (events), denoted $\mathcal{A}$. And a probability measure $\mathbb{P}$. Specification of the triple $(\Omega, \mathcal{A}, \mathbb{P})$ defines the *probability space* which models a real-world measurement or experimental process.

Example:

$$
\begin{aligned}
\Omega &= \{\text{all outcomes of the roll of a die}\} \\
&= \{1, 2, 3, 4, 5, 6\} \\
\mathcal{A} &= \{\text{all possible sets of outcomes}\} \\
&= \{\{1\}, \ldots, \{6\}, \{1, 2\}, \ldots, \{5, 6\}, \ldots, \{1, 2, 3, 4, 5, 6\}\} \\
\mathbb{P} &= \text{probability of all sets/events}
\end{aligned}
$$

What is the probability of a given $\omega \in \Omega$, say $\omega = 3$?
What is the probability of the event $\omega \in \{1, 2, 3\}$?
The basic physical model here is that all six outcomes are equally probable.

Boolean algebra

$$
\begin{aligned}
A \cup B &= \{\omega : \omega \in A \text{ or } \omega \in B\} \text{ (union)} \\
A \cap B &= \{\omega : \omega \in A \text{ and } \omega \in B\} \text{ (intersection)} \\
\bar{A} &= \{\omega : \omega \notin A\} \text{ (complement)} \\
\overline{A \cup B} &= \{\bar{A} \cap \bar{B}\}
\end{aligned}
$$

Events are mutually exclusive if

$$A \cap B = \emptyset \text{ (empty set)}$$

Probability Measures

A probability measure is a positive measure $\mathbb{P}$ satisfying

$$\begin{aligned} \mathbb{P}(A) &\geq 0 \; \forall A \in \Omega \\ \mathbb{P}(\emptyset) &= 0 \; \forall A \in \Omega \\ \mathbb{P}(\Omega) &= 1 \\ \text{if } \mathbb{P}(A \cap B) &= \emptyset, \text{ then } \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) \end{aligned}$$

Show that the last condition also implies:

Union bound

In general

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$$

an inequality sometimes called the *union bound*.

Example: Dice

$$
\begin{aligned}
\Omega &= \{1,2,3,4,5,6\} \\
\mathbb{P}(\{\omega = i\}) &= 1/6 \text{ for } i = 1,2,3,4,5,6 \\
\mathbb{P}(\omega \in \{1,2\}) &= \mathbb{P}(\{\omega = 1\} \cup \{\omega = 2\}) \\
&= \mathbb{P}(\{\omega = 1\}) + \mathbb{P}(\{\omega = 2\}) = 1/3
\end{aligned}
$$

Conditional Probability

Given two events, A and B, what is the probability that A will occur given that B has occurred?

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)},\ \mathbb{P}(B) \neq 0$$

Example:

$$\begin{aligned} \Omega &= \{1,2,3,4,5,6\} \\ A &= \{1,2\} \\ B &= \{2,3\} \end{aligned}$$

The probability that A occurs, without knowledge of whether B has occurred, is $1/3$ (i.e. $\mathbb{P}(A) = 1/3$). But

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\{2\})}{\mathbb{P}(\{2,3\})} = \frac{1/6}{1/3} = 1/2$$

Independence

Two events are said to be independent if

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

Equivalently, A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Example: Dice

Suppose we have two dice. Then

$$\begin{aligned} \Omega &= \{\text{all pairs of outcomes of the roll two dice}\} \\ &= \{(1,1),(1,2),\ldots,(6,6)\} \end{aligned}$$

Let $A = \{\text{1st die is 1}\}$ and $B = \{\text{2nd die is 1}\}$. $\mathbb{P}(A) = 1/6$.

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(\{(1,1)\})}{\mathbb{P}(\{1\})} = \frac{1/36}{1/6} = 1/6$$

The value of one die does not influence the other.

Bayes' Rule

$$\begin{aligned} \mathbb{P}(A|B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \\ \mathbb{P}(B|A) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \\ \implies \mathbb{P}(B|A) &= \frac{\mathbb{P}(A|B) \cdot \mathbb{P}(B)}{\mathbb{P}(A)} \end{aligned}$$

Example: Genetic testing

Geneticists have determined that a particular genetic defect is related to a certain disease. Many people with the disease also have this defect, but there are disease-free people with the defect. The geneticists have found that 0.01% of the general population has the disease and that the defect is present in 50% of these cases. They also know that 0.1% of the population has the defect. What is the probability that a person with the defect has the disease?

Example: Genetic testing (cont.)

This is a simple application of Bayes' Rule. We are interested in two events: 'defect' and 'disease'. We know:

$$\begin{aligned}\mathbb{P}(\text{disease}) &= 0.0001\\ \mathbb{P}(\text{defect}|\text{disease}) &= 0.5\\ \mathbb{P}(\text{defect}) &= 0.001.\end{aligned}$$

We have everything we need to apply Bayes' Rule:

$$\begin{aligned}\mathbb{P}(\text{disease}|\text{defect}) &= \frac{\mathbb{P}(\text{defect}|\text{disease})\mathbb{P}(\text{disease})}{\mathbb{P}(\text{defect})}\\ &= \frac{0.5 \times 0.0001}{0.001} = 0.05\end{aligned}$$

In other words, if a person has the defect then the chance they have the disease is 5%. In the general population, on the other hand, the chance that a person has the disease is 0.01%. So this "genetic marker" is quite informative.

Random variables and probability distributions

A random variable X is a mapping of Ω to the real or complex numbers.

Example: Digital Thermometer

$$\begin{aligned} \Omega &= \{\text{weather patterns}\} \\ X &= \text{mercury level on a thermometer} \end{aligned}$$

For instance, real-valued random variable is a mapping $X : \Omega \to \mathbb{R}^n$, which means that for every $\omega \in \Omega$ there is a corresponding value $X(\omega) \in \mathbb{R}^n$. Since $\mathbb{P}$ specifies the probability of every subset of Ω, it also induces probabilities on events expressed in terms of X. For example, if X is a real-valued scalar (i.e. $n = 1$) random variable, then this is an event:

$$\{X \geq 0\} \equiv \{\omega : X(\omega) \geq 0\}$$

Therefore we can compute the probability of this event:

$$\mathbb{P}(X \geq 0) = \mathbb{P}(\{\omega : X(\omega) \geq 0\})$$

Probability Distribution Function and Cumulative Distribution Function

When X takes values in the real numbers, the probability distribution function is completely described by the cumulative distrubtion function:

$$P_X(x) := \mathbb{P}(X \leq x) \equiv \mathbb{P}(\omega : X(\omega) \leq x)$$

If $P_X(x)$ is differentiable, then the probability density function $p_X(x)$ is its derivative. Since P_X is a monotonic increasing function, $p_X(x) \geq 0$. By the Fundamental Theorem of Calculus we have

$$P_X(x) = \int_{-\infty}^{x} p_X(x)dx.$$

Note also that $\lim_{x\to\infty} P_X(x) = \int_{-\infty}^{\infty} p_X(x) = 1$, since *with probability* 1 the variable X takes on a finite value. Observe that $\mathbb{P}(x_1 < X \leq x_2) = \int_{x_1}^{x_2} p_X(x)dx$, which explains the term "density."

Example: Uniform on $[0, 1]$

$$P(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

$$p(x) = I_{[0,1]}(x) \equiv \begin{cases} 1, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

Standard Normal (Gaussian)

$$\text{mean } 0\text{, variance } 1\text{: } p(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \sim \mathcal{N}(0,1)$$

$$\text{mean } \mu\text{, variance } \sigma^2\text{: } p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \sim \mathcal{N}(\mu,\sigma^2)$$

CDF of Standard Normal

The cumulative distribution when $X \sim \mathcal{N}(\mu, \sigma^2)$ is

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt.$$

No closed-form expression! However, many people write the CDF in terms of the error function:

$$\operatorname{erf}(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-u^2} du$$

To see how the CDF can be expressed using the error function, set

$$\begin{aligned} s &\equiv \frac{t-\mu}{\sqrt{2\sigma^2}} \\ t &= \sqrt{2\sigma^2}s + \mu \\ dt &= \sqrt{2\sigma^2}ds \\ t = x &\implies s = \frac{x-\mu}{\sqrt{2\sigma^2}} \end{aligned}$$

We then have

$$
\begin{aligned}
P(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt \\
&= \frac{\sqrt{2\sigma^2}}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{(x-\mu)/\sqrt{2\sigma^2}} e^{-s^2} ds \\
&= 1/2 + \frac{1}{\sqrt{\pi}} \int_{0}^{(x-\mu)/\sqrt{2\sigma^2}} e^{-s^2} ds \\
&= 1/2 + \mathrm{erf}((x-\mu)/\sqrt{2\sigma^2})
\end{aligned}
$$

Probability Mass Function

Discrete random variables: mapping Ω to an enumerable set of values, such as $\{1, 2, \dots\}$. These are characterized by a *probability mass function* (pmf), which has the same interpretation as the density.

If X takes values in a set $\{x_1, x_2, \dots\}$ (which may be finite or infinite), then the pmf of X is given by

$$p_X(x) = \sum_i \mathbb{P}(X = x_i)\mathbf{1}_{x=x_i},$$

where $\mathbf{1}_{x=x_i}$ is the *indicator function* that takes a value 1 if $x = x_i$ and 0 otherwise. Note that $p_X(x) = \mathbb{P}(X = x)$.

Example: Binomial distribution

Suppose you toss a coin n times and count k heads. This number is a random variable X taking a value between 0 and n, and dependent on p, the probability of observing a head in a single toss. The binomial distribution gives the probability of observing k heads in the n trials, for $k \in \{0, \dots, n\}$, and has the following form:

$$p_X(x) = \sum_{k=0}^{n} \underbrace{\binom{n}{k} p^k (1-p)^{n-k}}_{\mathbb{P}(X=k)} \mathbf{1}_{x=k}$$

- $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ sequences of heads and tails that have exactly k heads,
- $p^k(1-p)^{n-k}$ is the probability of each such sequence

Shorthand: $X \sim \text{Bi}(n, p)$.

Example: Poisson random variable

Imagine standing by a street counting the number of cars go by in a 10-minute interval. The number of cars is a random variable X taking values between 0 and ∞ and dependent on λ, the average number of cars per 10-minute interval. The Poisson distribution gives the probability of observing k cars and has the following form:

$$\mathbb{P}(X = k) = p_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}, \quad \lambda > 0, k = 0, 1, \dots$$

Note that $p_X = \lim_{n\to\infty} \mathrm{Bi}(n, \lambda/n)$.

Expectation

The expected value of a random variable X is defined to be

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} xp(x)dx \quad \text{(continuous case)}$$

or

$$\mathbb{E}[X] = \sum_j x_j p(x_j) \quad \text{(discrete case)}.$$

More generally, if f is an arbitrary function, then

$$\mathbb{E}[f(X)] = \int_{-\infty}^{\infty} f(x)p(x)dx \quad \text{(continuous case)}$$

$$\mathbb{E}[X] = \sum_j f(x_j)p(x_j) \quad \text{(discrete case)}.$$

Example: Second moment

$$f(x) = x^2 \longrightarrow \mathbb{E}[X^2]$$

Example: Characteristic function (Fourier transform of density)

$$f(x) = e^{j\omega x} \longrightarrow \mathbb{E}[e^{j\omega X}]$$

Example: Indicator Function

$$f(x) = I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} \longrightarrow \mathbb{E}[I_A(X)] = \mathbb{P}(X \in A)$$

Cauchy-Schwarz Inequality:

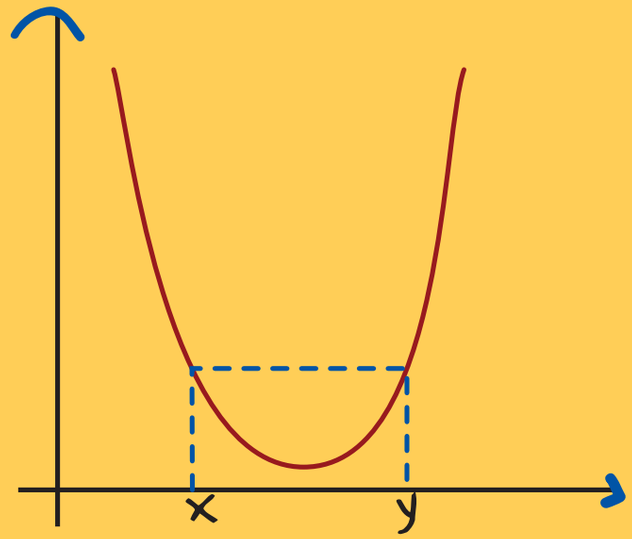
$$\mathbb{E}[|X \cdot Y|] \leq \left(\mathbb{E}[X^2] \cdot \mathbb{E}[Y^2]\right)^{1/2}$$

Convex function

A function ϕ is convex if

$$\lambda\phi(x) + (1-\lambda)\phi(y) \geq \phi(\lambda x + (1-\lambda)y)$$

for all $x, y \in \mathbb{R}$ and for all $\lambda \in [0, 1]$.



Jensen's Inequality

Suppose ϕ is a convex function. Then $\mathbb{E}[\phi(X)] \geq \phi(\mathbb{E}[X])$.

Example: Second moment

$\phi(x) = x^2 \Longrightarrow \mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$. (The second moment is always larger than the first moment squared.)

Jointly distributed random variables

Joint distribution:

$$P(x, y) = P(\{X \leq x\} \cap \{Y \leq y\})$$

Joint density:

$$P(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \underbrace{p(x, y)}_{\text{density}} dx dy$$

Statistically independent random variables:

$$P(x, y) = P(x) \cdot P(y) \text{ or } p(x, y) = p(x) \cdot p(y)$$

Uncorrelated random variables:

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

independent $\Rightarrow$ uncorrelated
uncorrelated $\not\Rightarrow$ independent

Conditional densities

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

if X and Y are independent, then

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

Conditional Expectation:

$$\mathbb{E}[X \,|\, Y = y] = \int_{-\infty}^{\infty} x p(x|y) dx$$

Note: $\mathbb{E}[X|Y]$ is a function of y, the value that the random variable Y takes.

Smoothing property of conditional expectation

So if Y is random, then so is $\mathbb{E}[X|Y]$, and it has the following expected value:

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[X|Y]] &= \int \left[\int x p(x|y) dx \right] p(y) dy \\
&= \int x \left[\int p(x|y) p(y) dy \right] dx \\
&= \int x \left[\int p(x,y) dy \right] dx \\
&= \int x p(x) dx \\
&= \mathbb{E}[X]
\end{aligned}$$

Example: Smoothing

Let L be a positive integer-valued random variable. Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed (iid) random variables each with mean m. Consider the partial sum:

$$S_L = \sum_{i=1}^{L} X_i$$

What is the expected value of S_L?

$$\begin{aligned}
\mathbb{E}[S_L|L] &= \mathbb{E}\left[\sum_{i=1}^{L} X_i|L\right] \\
&= \sum_{i=1}^{L} \mathbb{E}[X_i] = Lm \\
\mathbb{E}[S_L] &= \mathbb{E}[\mathbb{E}[S_L|L]] \\
&= \mathbb{E}[Lm] = m\mathbb{E}[L]
\end{aligned}$$

Random vectors

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Distribution:

$$P(x) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$$

Density:

$$P(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} \underbrace{p(\boldsymbol{t})}_{\text{density}} \, dt_1 \ldots dt_n$$

Expectation:

$$\begin{aligned} f &: \mathbb{R}^n \longrightarrow \mathbb{R}^m \\ \mathbb{E}[f(X)] &= \int_{\mathbb{R}^n} f(x)p(x)dx \end{aligned}$$

The Multivariate Normal Distribution

Notation: $X \sim \mathcal{N}(\mu, R)$.

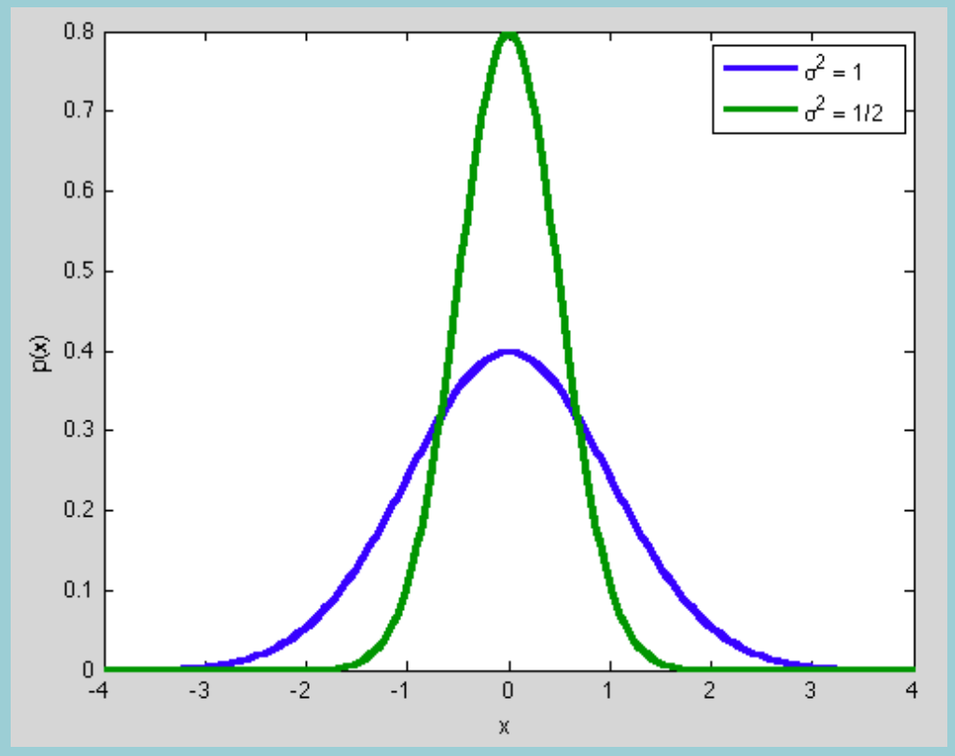
$$
\begin{aligned}
p(x) &= \frac{1}{(2\pi)^{n/2}|R|^{1/2}} e^{-\frac{1}{2}(x-\mu)^\top R^{-1}(x-\mu)} \sim \mathcal{N}(\mu, R) \\
\mu &= \mathbb{E}[X] \\
R &= \mathbb{E}[(x-\mu)(x-\mu)^\top] = \text{covariance matrix}
\end{aligned}
$$

Note: $R_{i,j} = \mathbb{E}[(x_i - \mu_i)(x_j - \mu_j)]$.

Example: 1d

$$R = \sigma^2$$

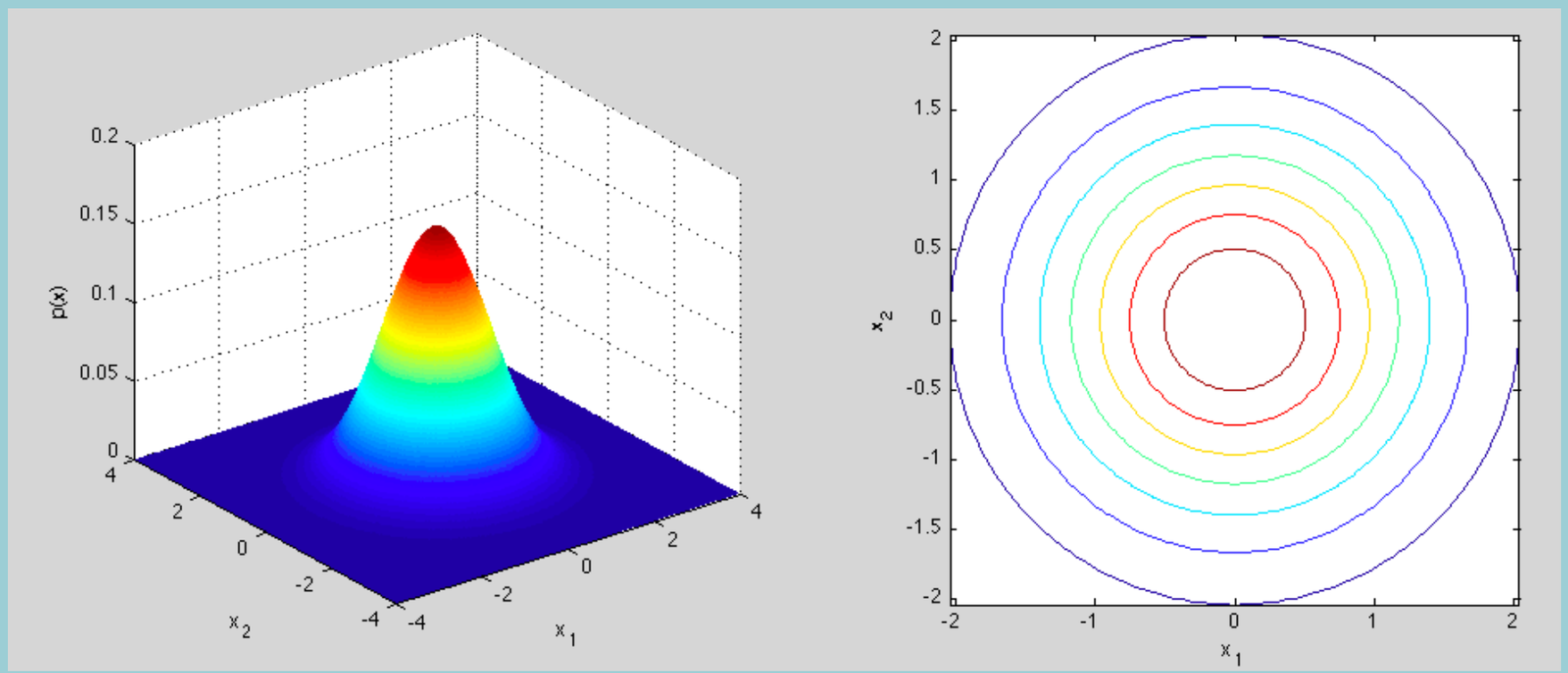
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Example: 2d symmetric

$$R = \sigma^2 \boldsymbol{I}_{2\times 2} = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

A contour of this density is a circle. That is, if we find all x such that $p(x) = \gamma$, then this is the set of all x such that $\|x - \mu\|^2 = \gamma\prime$.



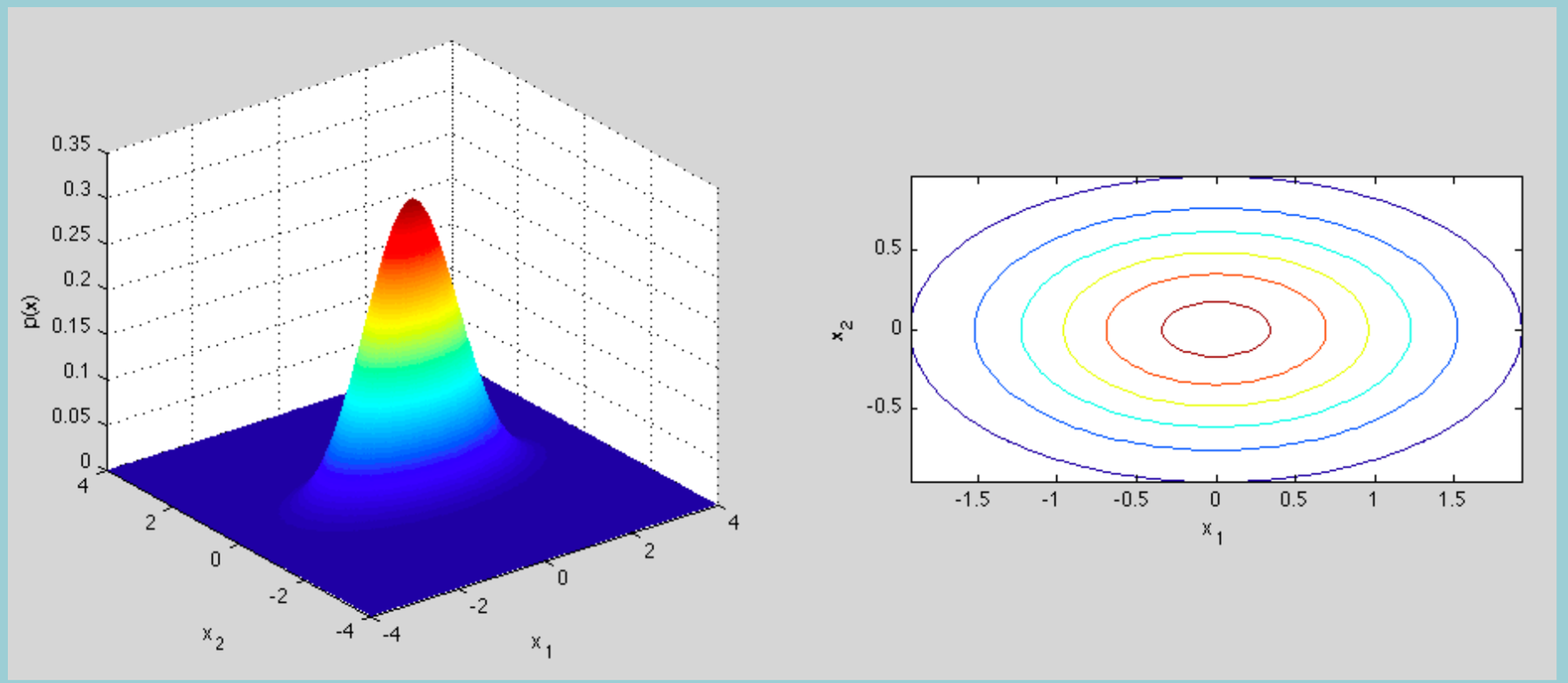
Example: 2d axis-aligned ellipse

$$R = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Here the density contours are ellipses whose axes align with the coordinate axes. To see this, note that

$$p(x) = \gamma \iff \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} = \gamma'$$

which is the equation for an ellipse.



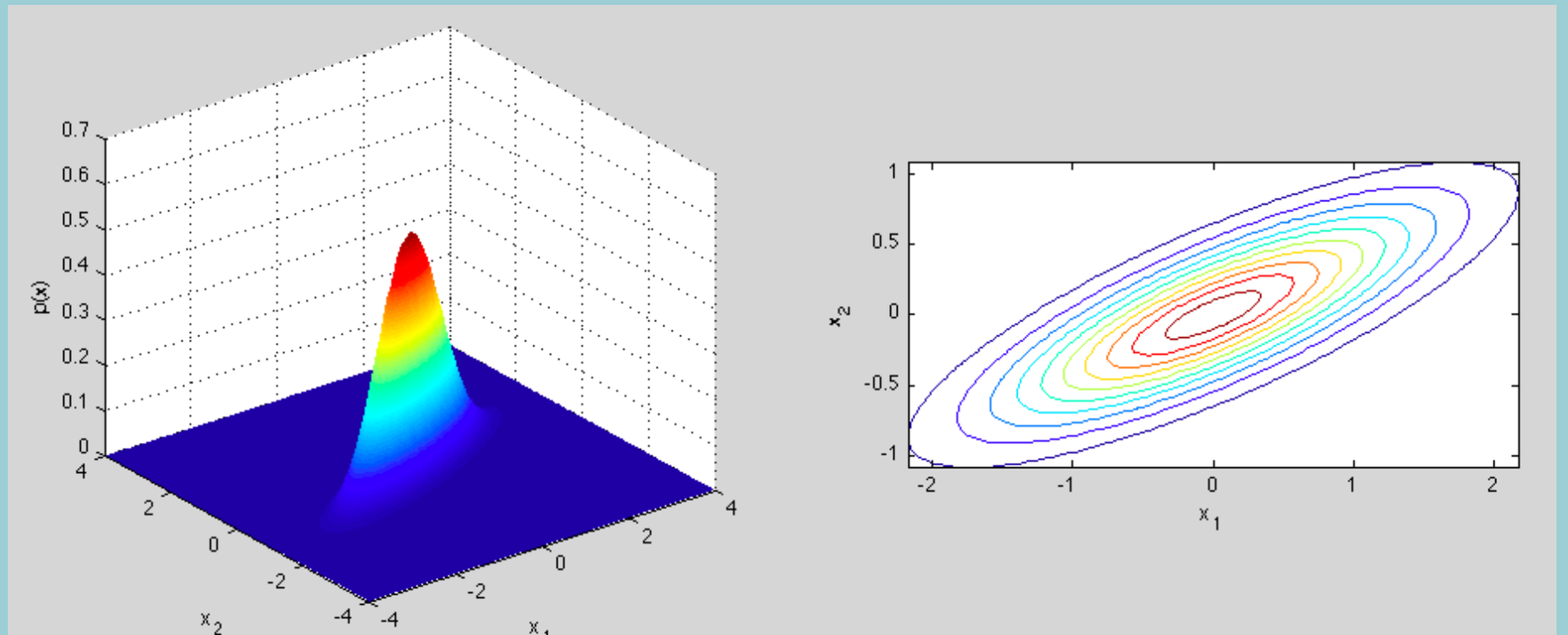
Example: 2d rotated ellipse

Here R is an arbitrary (positive definite) matrix. Here the density contours are ellipses with arbitrary orientation. To see this, first write $R = V\Lambda V^\top$; let $x\prime := V^\top x$ and $\mu\prime := V^\top \mu$. Then

$$(x-\mu)^\top R^{-1}(x-\mu) = (x-\mu)^\top V\Lambda^{-1}V^\top(x-\mu)$$

$$= (x\prime-\mu\prime)^\top\Lambda^{-1}(x\prime-\mu\prime) = \frac{(x_1\prime-\mu_1\prime)^2}{\lambda_1} + \frac{(x_2\prime-\mu_2\prime)^2}{\lambda_2}$$

which defines an ellipse in a rotated coordinate system, where V specifies the rotation.



Linear transformations of Gaussian Vectors

Suppose that we transform a multivariate normal vector X by applying a linear transformation (matrix) A:

$$Y = AX$$

(X, Y random, A deterministic). Now the characteristic function of Y is

$$\begin{aligned}
\phi(\omega) &= \mathbb{E}[e^{-j\omega^\top Y}] = \mathbb{E}[e^{-j\omega^\top AX}] \\
&= \underbrace{\exp\left\{ j\omega^\top A\mu - \frac{1}{2}\omega^\top ARA^\top \omega \right\}}
\end{aligned}$$

characteristic function of a $\mathcal{N}(A\mu, ARA^\top)$ random vector

$$\implies Y \sim \mathcal{N}(A\mu, ARA^\top)$$

Example: Weighted sum of Gaussian r.v.s

- X is vector of Gaussian r.v.s
- a is vector of weights
- $a^\top X$ is weighted sum

$$\begin{aligned} Y &= a^\top X \\ Y &\sim \mathcal{N}(a^\top \mu, a^\top R a) \end{aligned}$$

Covariance diagonalization

Let $X \sim \mathcal{N}(\mu, R)$ and let $v_1, \ldots, v_n$ be the eigenvectors of R
(Note: R is symmetric, positive semidefinite matrix)
Define $V = [v_1, \ldots, v_n]$.

$$V^\top X \sim \mathcal{N}(V^\top \mu, \underbrace{V^\top R V}_{\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)})$$

The transformation V diagonalizes the covariance matrix. Since the cross-correlations between elements of the transformed vector are identically zero and jointly Gaussian distributed, they are statistically independent.

Karhunen-Loeve transform

Thus, the transformation V decomposes the random vector X into a sum of independent components:

$$X = \sum_{i=1}^{n} (v_i^\top X) \cdot v_i$$

where the coefficients $v_i^\top x$ are independent Gaussian random variables

$$v_i^\top X \sim \mathcal{N}(v_i^\top \mu, \lambda_i)$$

This basic transformation is called the Karhunen-Loève transform, and it is a fundamental tool that is useful for decomposing random signals into key components.

Example:

We may interpret the component with the largest variance, $\max_i \lambda_i$, as the most important component in X.

Principal components analysis

Assume $\mu = 0$. (In practice, μ may not be zero, but we often estimate μ and then subtract this estimate from X before running PCA.)

Definition: Principal components analysis

The first r eigenvectors are called the first r principal components of X, and X's projection on these principal components is

$$X_r = \sum_{i=1}^{r} (v_i^\top X) v_i.$$

Note that this approximation involves only r scalar random variables $\{(v_i^\top X)\}_{i=1}^r$ rather than n. In fact, it is easy to show that among all r-term linear approximations of X in terms of r random variables, X_r has the smallest mean square error; that is if we let $\mathcal{S}_r$ denote all r-term linear approximations to X, then

$$\mathbb{E}[\|X - X_r\|^2] = \min_{Y_r \in \mathcal{S}_r} \mathbb{E}[\|X - Y_r\|^2]$$

Convergence of Sums of Independent Random Variables

Example: Synthia

A biologist is studying the new artificial lifeform called *synthia*. She is interested to see if the synthia cells can survive in cold conditions. To test synthia's hardiness, the biologist will conduct n independent experiments. She has grown n cell cultures under ideal conditions and then exposed each to cold conditions. The number of cells in each culture is measured before and after spending one day in cold conditions. The fraction of cells surviving the cold is recorded. Let $x_1, \ldots, x_n$ denote the recorded fractions. The average

$$\widehat{p} := \frac{1}{n} \sum_{i=1}^{n} x_i$$

is an estimator of the survival probability.

Understanding behavior of sums of independent random variables is extremely important. For instance, the biologist in the example above would like to know that the estimator is reasonably accurate. Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with variance $\sigma^2 < \infty$ and consider the average $\widehat{\mu} := \frac{1}{n} \sum_{i=1}^{n} X_i$.

- $\mathbb{E}[\widehat{\mu}] = \mathbb{E}[X]$ – same as RVs
- The variance of $\widehat{\mu}$ is σ^2/n – reduced by a factor of n

Lower variance means less uncertainty. So it is possible to reduce uncertainty by averaging. The more we average, the less the uncertainty (assuming, as we are, that the random variables are independent, which implies they are uncorrelated).

What about the distribution of the average $\widehat{\mu}$?

Central Limit Theorem

If X_i, $i = 1, \ldots, n$, are n independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathcal{N}(\mu, \sigma^2/n).$$

regardless of the form of the distribution of the variables.

Can we say something about the distribution even for finite values of n?

One approach is to calculate the distribution of the average explicitly. Recall that if the random variables have a density p_X, then the density of the sum $\sum_{i=1}^{n} X_i$ is the n-fold convolution of the density p_X with itself.

(Again this hinges on the assumption that the random variables are independent; it is easy to see by considering the characteristic function of the sum and recalling that multiplication of Fourier transforms is equivalent to convolution in the inverse domain).

However, this exact calculation can be sometimes difficult or impossible, if for instance we don't know the density p_X, and so sometimes probability bounds are more useful.

Let Z be a non-negative random variable and take $t > 0$.

Markov's Inequality

$$\begin{aligned} \mathbb{E}[Z] &\geq \mathbb{E}[Z\,\mathbf{1}_{Z\geq t}] \\ &\geq \mathbb{E}[t\,\mathbf{1}_{Z\geq t}] \;=\; t\,\mathbb{P}(Z \geq t) \end{aligned}$$

Now we can use this to get a bound on the probability 'tails' of Z.

Chebyshev's Inequality

$$\begin{aligned} \mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) &= P((Z - \mathbb{E}[Z])^2 \geq t^2) \\ &\leq \frac{\mathbb{E}[(Z - \mathbb{E}[Z])^2]}{t^2} \\ &= \frac{\mathsf{Var}(Z)}{t^2}\,, \end{aligned}$$

where $\mathsf{Var}(Z)$ denotes the variance of Z.

If we apply this to the average $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then we have

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \quad \leq \quad \frac{\sigma^2}{nt^2}$$

where μ and σ^2 are the mean and variance of the random variables $\{X_i\}$. This shows that not only is the variance reduced by averaging, but the tails of the distribution (probability of observing values a distance of more than t from the mean) are smaller.

Can we say more?

The tail bound given by Chebyshev's Inequality is loose, and much tighter bounds are possible under slightly stronger assumptions.

In particular, if the random variables $\{X_i\}$ are bounded or *sub-Gaussian* (meaning the tails of the probability distribution decay at least as fast as Gaussian tails), then the tails of the average converge exponentially fast in n. The simplest result of this form is for bounded random variables.

Hoeffding's Inequality

Let $X_1, X_2, ..., X_n$ be independent bounded random variables such that $X_i \in [a_i, b_i]$ with probability 1. Let $S_n = \sum_{i=1}^{n} X_i$. Then for any $t > 0$, we have

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geq t) \leq 2\, e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

If the random variables $\{X_i\}$ are binary-valued, then this result is usually referred to as the *Chernoff Bound*.

The proof of Hoeffding's Inequality, which relies on a clever generalization of Markov's inequality and some elementary concepts from convex analysis, is given in the last slides.

Now suppose that the random variables in the average $\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$ are bounded according to $a \leq X_i \leq b$. Let $c = (b-a)^2$. Then Hoeffding's Inequality implies

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \leq 2\, e^{-\frac{2nt^2}{c}}$$

In other words, the tails of the distribution of the average are tending to zero at an exponential rate in n, much faster than indicated by Chebeyshev's Inequality. Similar exponential tail bounds hold for averages of iid sub-Gaussian variables. Using tail bounds like these we can prove the so-called *laws of large numbers*.

Weak Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be iid random variables with $\mathbb{E}[|X_i|] < \infty$. Then $\widehat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to $\mu := \mathbb{E}[X_i]$; that is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}[|\widehat{\mu}_n - \mu| > \epsilon] = 0.$$

Strong Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be iid random variables with $\mathbb{E}[|X_i|] < \infty$. Then $\widehat{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$ converges almost surely to $\mu := \mathbb{E}[X_i]$; that is,

$$\mathbb{P}(\lim_{n\to\infty} \widehat{\mu}_n = \mu) = 1.$$

Example: Synthia revisited

The biologist has collected n observations, $x_1, \ldots, x_n$, each corresponding to the fraction of cells that survived in a given experiment. Her estimator of the survival rate is $\frac{1}{n}\sum_{i=1}^{n} x_i$. How confident can she be that this is an accurate estimator of the true survival rate? Let us model her observations as realizations of n iid random variables $X_1, \ldots, X_n$ with mean p and define $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We say that her estimator is probability approximately correct with non-negative parameters (ϵ, δ) if

$$\mathbb{P}(|\widehat{p} - p| > \epsilon) \ \leq \delta$$

The random variables are bounded between 0 and 1 and so the value of c in Hoeffding's inequality is equal to 1. For desired accuracy $\epsilon > 0$ and confidence $1 - \delta$, how many experiments will be sufficient?

Example: Synthia revisited (cont.)

From Hoeffding's inequality, we equate $\delta = 2\,\exp(-2n\epsilon^2)$ which yields $n \geq \frac{1}{2\epsilon^2}\log(2/\delta)$. Note that this requires no knowledge of the distribution of the $\{X_i\}$ apart from the fact that they are bounded. The result can be summarized as follows. If $n \geq \frac{1}{2\epsilon^2}\log(2/\delta)$, then the probability that her estimate is off the mark by more than ϵ is less than δ.

Proof of Hoeffding's inequality

Let X be any random variable and $s > 0$. Note that $\mathbb{P}(X \geq t) = \mathbb{P}(e^{sX} \geq e^{st}) \leq e^{-st}\mathbb{E}[e^{sX}]$, by using Markov's inequality, and note that e^{sx} is a non-negative monotone increasing function. For clever choices of s this can be quite a good bound. Let's look now at $\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$:

$$\begin{aligned}
\mathbb{P}(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t) &\leq e^{-st}\mathbb{E}\left[e^{s\left(\sum_{i=1}^{n}(X_i - E[X_i])\right)}\right] \\
&= e^{-st}\mathbb{E}\left[\prod_{i=1}^{n} e^{s(X_i - \mathbb{E}[X_i])}\right] \\
&= e^{-st}\prod_{i=1}^{n}\mathbb{E}\left[e^{s(X_i - \mathbb{E}[X_i])}\right] \ ,
\end{aligned}$$

where the last step follows from the independence of the X_i's. To complete the proof we need to find a good bound for $\mathbb{E}\left[e^{s(X_i - E[X_i])}\right]$.

Lemma

Let Z be a r.v. such that $\mathbb{E}[Z] = 0$ and $a \leq Z \leq b$ with probability one. Then

$$\mathbb{E}\left[e^{sZ}\right] \leq e^{\frac{s^2(b-a)^2}{8}} \ .$$

This upper bound is derived as follows. By the convexity of the exponential function,

$$e^{sz} \leq \frac{z-a}{b-a}e^{sb} + \frac{b-z}{b-a}e^{sa}, \text{ for } a \leq z \leq b \ .$$

Thus,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &\leq \mathbb{E}\left[\frac{Z-a}{b-a}\right]e^{sb} + \mathbb{E}\left[\frac{b-Z}{b-a}\right]e^{sa} \\
&= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \ , \text{ since } \mathbb{E}[Z] = 0 \\
&= (1-\lambda+\lambda e^{s(b-a)})e^{-\lambda s(b-a)} \ , \text{ where } \lambda = \frac{-a}{b-a}
\end{aligned}$$

Now let $u = s(b-a)$ and define

$$\phi(u) \equiv -\lambda u + \log(1 - \lambda + \lambda e^u) \ ,$$

so that

$$\mathbb{E}[e^{sZ}] \leq (1 - \lambda + \lambda e^{s(b-a)}) e^{-\lambda s(b-a)} = e^{\phi(u)} \ .$$

We want to find a good upper-bound on $e^{\phi(u)}$. Let's express $\phi(u)$ as its Taylor series with remainder:

$$\phi(u) = \phi(0) + u\phi\prime(0) + \frac{u^2}{2}\phi\prime\prime(v) \text{ for some } v \in [0, u] \ .$$

$$\begin{aligned}
\phi\prime(u) &= -\lambda + \frac{\lambda e^u}{1 - \lambda + \lambda e^u} \Rightarrow \phi\prime(0) = 0 \\
\phi\prime\prime(u) &= \frac{\lambda e^u}{1 - \lambda + \lambda e^u} - \frac{\lambda^2 e^{2u}}{(1 - \lambda + \lambda e^u)^2} \\
&= \frac{\lambda e^u}{1 - \lambda + \lambda e^u}(1 - \frac{\lambda e^u}{1 - \lambda + \lambda e^u}) \\
&= \rho(1 - \rho) \ ,
\end{aligned}$$

where $\rho = \frac{\lambda e^u}{1-\lambda+\lambda e^u}$.
Now note that $\rho(1 - \rho) \leq 1/4$, for any value of ρ (the maximum is attained when $\rho = 1/2$, therefore $\phi\prime\prime(u) \leq 1/4$. So finally we have $\phi(u) \leq \frac{u^2}{8} = \frac{s^2(b-a)^2}{8}$, and therefore

$$\mathbb{E}[e^{sZ}] \leq e^{\frac{s^2(b-a)^2}{8}} \ .$$

Now, we can apply this upper bound to derive Hoeffding's inequality.

$$\begin{aligned}
\mathbb{P}(S_n - \mathbb{E}[S_n] \geq t) &\leq e^{-st} \prod_{i=1}^{n} \mathbb{E}[e^{s(X_i - \mathbb{E}[X_i])}] \\
&\leq e^{-st} \prod_{i=1}^{n} e^{\frac{s^2(b_i-a_i)^2}{8}} \\
&= e^{-st} e^{s^2 \sum_{i=1}^{n} \frac{(b_i-a_i)^2}{8}} \\
&= e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}} \\
&\text{by choosing } \; s = \frac{4t}{\sum_{i=1}^{n}(b_i - a_i)^2}
\end{aligned}$$

The same result applies to the r.v.'s $-X_1, \ldots, -X_n$, and combining these two results yields the claim of the theorem.