ECE 830 Fall 2011 Statistical Signal Processing

instructor: R. Nowak

Lecture 2: Review of Probability Theory

Probabilistic models will be used throughout the course to represent noise, errors, and uncertainty in signal processing problems. This lecture reviews the basic notation, terminology and concepts of probability theory.

1 Basic Probability

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 1

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

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.

1.1 Probability Measures

Probability measures must satisfy the following properties:

1. $\mathbb{P}(A) \geq 0$, $\forall A \in \mathcal{A}$
2. $P(\Omega) = 1$
3. if $A \cap B = \emptyset$, then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$

Show that the last condition also implies that in general $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$, an inequality sometimes called the *union bound*.

1.2 Conditional Probability

Consider two events $A, B \in \mathcal{A}$. The (conditional) probability that A occurs given B occurs is

$$P(A|B) \;:=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Example 2

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

Recall that we say that A "occurs" if the outcome is either 1 *or* 2*. Now suppose you are told that B occurs. Then it seems more probable that A has also occurred.*

The probability that A occurs, without knowledge of whether B has occurred, is $1/3$*. That is,* $\mathbb{P}(A) = 1/3$*. From the formula above it is an easy calculation to see that*

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

So we see that knowledge of B provides important information about A. Does this make sense with your intuition?

1.3 Independence

Two events A and B are said to be independent if $P(A|B) = \mathbb{P}(A)$. In otherwords, B provides no information about whether A has occurred.

Example 3 *Supose we have two dice. Then*

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

Let $A = \{\textit{1st die is 1}\}$ *and* $B = \{\textit{2nd die is 1}\}$*.* $\mathbb{P}(A) = 1/6$*.* $P(A|B) = \frac{\mathbb{P}(\{(1,1)\})}{\mathbb{P}(\{1\})} = \frac{1/36}{1/6} = 1/6$*.*

As we know, the value of one die does not influence the other. The two outcomes are independent.

1.4 Bayes Rule

Again consider two events A and B. Recall the conditional probability $P(A|B)$. Bayes rule is a formula for the "inverse" conditional probability, $P(B|A)$. Bayes' Rule is

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

It is easy to verify by recalling the definition of conditional probability above. This inversion formula will play a central role in signal estimation problems later in the course.

Example 4 *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?*

This is a simple application of Bayes' Rule. We are interested in two events: 'defect' and 'disease'. Here is what the geneticists know: $\mathbb{P}(\textit{disease}) = 0.0001$*,* $P(\textit{defect}|\textit{disease}) = 0.5$*,* $\mathbb{P}(\textit{defect}) = 0.001$*. We have everything we need to apply Bayes' Rule:*

$$P(\textit{disease}|\textit{defect}) = \frac{P(\textit{defect}|\textit{disease})\,\mathbb{P}(\textit{disease})}{\mathbb{P}(\textit{defect})} = \frac{0.5 \times 0.0001}{0.001} = 0.05$$

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.

2 Random Variables

Often in applications the underlying probability space is not explicitly identified. Rather we work with random variables which are *mappings* from Ω to more concrete spaces such as $\mathbb{R}^n$.

Example 5 *A real-valued random variable is a mapping $X : \Omega \to \mathbb{R}$, which means that for every $\omega \in \Omega$ there is a corresponding value $X(\omega) \in \mathbb{R}$. Random variables can also be vectors, for example a mapping $X : \Omega \to \mathbb{R}^n$. 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 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:

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

More generally, for any set A we may consider the event $\{X \in A\}$ and its probability $\mathbb{P}(X \in A)$.

2.1 Cumulative Distributions and Densities

For real-valued scalar random variables it is common to consider the probabilities $\mathbb{P}(X \leq x)$ as a function of x. This function is called the *cumulative distribution function* of X, and it is often denoted by $F_X(x)$. The name can be understood by observing that the difference between $F_X(x_1)$ and $F_X(x_2)$ for any $x_2 > x_1$ is equal to the probability that X takes a value in the interval $(x_1, x_2]$; i.e.,

$$F_X(x_2) - F_X(x_1) = \mathbb{P}(X \leq x_2) - \mathbb{P}(X \leq x_1) = \mathbb{P}(x_1 < X \leq x_2)$$

The equality above follows by noting that $\{X \leq x_2\} = \{X \leq x_1\} \cup \{x_1 < X \leq x_2\}$ and these two sets are disjoint. Therefore by the basic property of probabilities we have that

$$\mathbb{P}(X \leq x_2) = \mathbb{P}(X \leq x_1) + \mathbb{P}(x_1 < X \leq x_2).$$

For continuous random variables we can consider the limit of the difference as $x_2 \to x_1$. Assume that the limit $\lim_{\Delta \to 0} \frac{F_X(x+\Delta) - F_X(x)}{\Delta} := p_X(x)$ exists, in other words F_X is differentiable at x. Since F_x is a monotonic increasing function, $p_X(x) \geq 0$. If F_x is differentiable everywhere, then by the Fundamental Theorem of Calculus we have

$$F_X(x) = \int_{-\infty}^{x} p_X(x)\, dx$$

Note also that $\lim_{x\to\infty} F_X(x) = \int_{-\infty}^{\infty} p_X(x) = 1$, since *with probability* 1 the variable X takes on a finite value. The function $p_X(x)$ is called the *probability density function* (pdf) of X. 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 6 *Recall the test statistic from the communication problem in Lecture 1*

$$t \;=\; \sum_{i=1}^{n} s_i x_i \;=\; \theta \|s\|_2^2 + \sum_{i=1}^{n} s_i \epsilon_i$$

If the errors are noise-like then we expect $z := \sum_{i=1}^{n} s_i \epsilon_i \approx 0$. Moreover as n increases it is reasonable to suppose that the approximation becomes more and more accurate since the individual errors may be randomly positive or negative and tend to cancel each other out. A reasonable probability density model for z is then $p(z) := \frac{1}{\sqrt{2\pi n}}\, e^{-\frac{1}{2}z^2/n}$, a Gaussian density with zero mean and standard deviation $1/\sqrt{n}$. As n gets larger the probability mass is more concentrated about 0. Since t is just a shifted version of z (shifted by the constant $\theta\|s\|_2^2$), the density for t is

$$p(t) \;=\; \frac{1}{\sqrt{2\pi n}}\, e^{-\frac{1}{2}(t - \theta\|s\|_2^2)^2/n}$$

This density is peaked about $\|s\|_2^2$ when $\theta = +1$ and about $-\|s\|_2^2$ when $\theta = -1$, so our test of whether t was larger or less than 0 is reasonable. Assuming this model we can easily calculate the probabilities of an incorrect decisions about which bit was sent; i.e., $\mathbb{P}(X < 0 \,|\, \theta = +1)$ and $\mathbb{P}(X > 0 \,|\, \theta = -1)$.

The general form of a Gaussian (or "Normal"*) density with mean μ and standard deviation σ is $\frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{1}{2}(z-\mu)^2/\sigma^2}$. Shorthand notation for this distribution is $\mathcal{N}(\mu, \sigma^2)$ and shorthand for indicating that a random variable X has this distribution is $X \sim \mathcal{N}(\mu, \sigma^2)$. So for our test statistic we could write $t \sim \mathcal{N}(\theta\|s\|_2^2, n^{-1})$. The "standard normal" distribution is $\mathcal{N}(0, 1)$.*

Scalar random variables that take values in a discrete set, such as $\{1, 2, \dots\}$, do not have a density function. Instead they have a *probability mass function* (pmf). A pmf 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 7 *One of the most common discrete distributions is the binomial distribution. Suppose you toss a coin n times and count the number of heads. This number is a random variable X taking a value between 0 and n, and of course it will depend 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 $k = 0, \dots, n$, and has the following form:*

$$p_X(x) = \sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k}\, \mathbf{1}_{x=k}$$

In other words, $\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is the binomial coefficient. It is easy to understand this formula. There are $\binom{n}{k}$ sequences of heads and tails that have exactly k heads, and the probability of each such sequence is $p^k(1-p)^{n-k}$. Our shorthand notation for this case is $X \sim Bi(n, p)$.

To simplify the notation, often we will denote probability functions by $F(x)$ or $p(x)$, dropping the subscript and using the choice of argument variable to indicate the underlying random variable.

3 Multivariate Distributions

In many signal processing problems we will need to consider vector-valued random variables. Such a random variable consists of many "variates" (the individual elements making up the vector which can be viewed as scalar random variables). Multivariate distributions describe the *joint* probability relationships among the variates. The most common multivariate distribution, and the one that will be used most often in the course, is the multivariate Gaussian (or Normal) distribution.

Suppose $X : \Omega \to \mathbb{R}^d$; that is X is a d-dimensional vector of scalar random variables. If the individual component random variables are independently distributed, then we have a trivial joint distribution which is just the product of the univariate distributions of each component. For example, if $X = [X_1\,,\, X_2]^T$ and X_1 and X_2 are independent and have densities, then the joint density of X is just $p_X(x) = p_{X_1}(x_1) p_{X_2}(x_2)$. Things are more interesting if the component random variables are not independent, in which case the multivariate distribution does not factor in this simple way. In fact, the joint distribution factorizes into univariate densities if and only if the component random variables are independent.

The multivariate Gaussian density is one model that can represent dependent random variables and it has the following form

$$p_X(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

where the argument $x \in \mathbb{R}^d$, $\mu \in \mathbb{R}^d$ is the *mean vector* of the density, and $\Sigma \in \mathbb{S}_+^d$ is the *covariance matrix*, where $\mathbb{S}_+^d$ denotes the cone of all positive semi-definite symmetric matrices (i.e., all symmetric matrices with non-negative eigenvalues). The term $|\Sigma|$ is the determinant of Σ. We will use $X \sim \mathcal{N}(\mu, \Sigma)$ as shorthand for "X is multivariate Gaussian distributed with mean μ and covariance Σ." It is easy to see that when $d = 1$ this reduces to the scalar Gaussian density function.

If Σ is a diagonal matrix (i.e., all off-diagonal entries are zero), then the component random variables are uncorrelated. Moreover, it is easy to verifiy that in that case the multivariate Gaussian density factorizes into univariate component densities, which means that uncorrelated Gaussian random variables are also independent. This is a special characteristic of the Gaussian density and in general being uncorrelated does not imply independence.

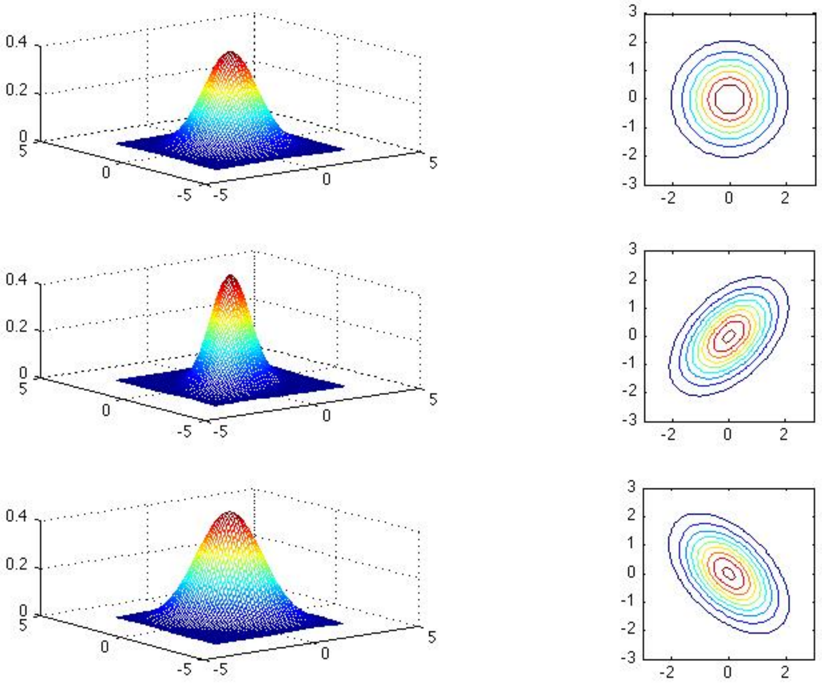

Figure 1: Multivariate Gaussian Distributions. The top row show the Gaussian density (left) and its contour plot (right) with mean $[0\ 0]^T$ and covariance $[1\ 0; 0\ 1]$. The second row is the same except that the covariance is $[1\ 0.5; 0.5\ 1]$; positive correlation. The third row is the same except that the covariance is $[1\ \ -0.5; -0.5\ 1]$; negative correlation.

Example 8 *Suppose that we have a linear system driven by a Gaussian white noise process. That is, a sequence of independent $N(0, 1)$ variables, denoted $X_1, \ldots, X_n$ is the input to the linear system. The output of the system can be expressed as $Y = \mathbf{H}\,X$, where $\mathbf{H}$ is a $n \times n$ matrix describing the action of the system on the input sequence $X = [X_1, \ldots, X_n]^T$. The joint distribution of X is $\mathcal{N}(0, I)$, where I denotes the $n \times n$ identity matrix. The joint distribution of the output Y is $\mathcal{N}(0, \mathbf{H}\mathbf{H}^T)$. The output is correlated due to the action of the system.*

Example 9 *Imagine we have a sensor network monitoring a manufacturing facility. The network consists of n nodes and its function is to monitor for failures in the system. Let $X = [X_1, \ldots, X_n]^T$ denote the set of* *scalar measurements produced by the network and suppose that it can modeled as a realization of a $\mathcal{N}(\mu, \Sigma)$ random vector for a particular choice of μ and Σ. Furthermore, assume that the set $B \subset \mathbb{R}^n$ denotes the set of all vectors indicative of failures. Then the probability of failure is given by $\mathbb{P}(X \in B)$, which can be calculated (numerically) by integrating the $\mathcal{N}(\mu, \Sigma)$ over the set B.*

It is sometimes necessary to consider *conditional* probabilities and expectations. Let X and Y be random variables. Suppose we observe $Y = y$. If X and Y are dependent, then knowledge that $Y = y$ tells us something about X. The conditional probability of X given $Y = y$ is defined according to the definition of conditional probability. If the random variables have a joint density or mass function, then the conditional density or mass function is defined as

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

where $p(y) = \int p(x,y)\,dx$ and x is the variable and y is fixed. Suppose we were interested in the probability that $X \in A$, for some set A, given $Y = y$. Then we have $P(X \in A \,|\, Y = y) = \int_A p(x|y)\,dx$.

4 Expectation

In addition to computing the probabilities that random variables fall into certain sets it is also useful to compute the average or expected value of random variables. If a random variable X has density $p_X(x)$, then the expectation of $f(X)$, where f is any function of X, is

$$\mathbb{E}[f(X)] \ = \ \int f(x)\, p_X(x)\, dx$$

If the random variable is discrete, then the expectation is

$$\mathbb{E}[f(X)] \ = \ \sum_i f(x_i)\, \mathbb{P}(X = x_i)$$

Here are some special cases.

mean: $\mu = \mathbb{E}[X]$

variance: $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2]$

probability: $\mathbb{P}(X \in A) = \mathbb{E}[\mathbf{1}_{\{X \in A\}}]$

characteristic function: $\phi(\omega) := \mathbb{E}[\exp(-i\omega X)]$

The characteristic function of X is the Fourier transform of its density.

If X is a d-dimensional random variable, $X = [X_1, \ldots, X_d]^T$, then the mean is

$$\mathbb{E}[X] \ = \ \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix}$$

and the *covariance matrix* is $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$, where $\Sigma_{i,j} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$, for $i, j = 1 \ldots, d$.

Note that if $Y = \mathbf{A}X$, where $\mathbf{A}$ is an $m \times n$ matrix, then the random vector Y has mean $\mathbb{E}[\mathbf{A}X] = \mathbf{A}\mu$ and covariance $\mathbb{E}[(\mathbf{A}X - \mathbf{A}\mu)(\mathbf{A}X - \mathbf{A}\mu)^T)] = \mathbf{A}\Sigma\mathbf{A}^T$. If X is multivariate Gaussian distributed, then so is Y (which can be shown using the characteristic function). In that case $Y \sim \mathcal{N}(\mathbf{A}\mu, \mathbf{A}\Sigma\mathbf{A}^T)$ (recall Example 8). This is a special property of the Gaussian distribution (generally linear transformations will alter the distributional characterstics).

We also sometimes need to use conditional expectations. For example, suppose X and Y are two dependent random variables. If we observe $Y = y$, then the expectation of $f(X)$ may differ from $\mathbb{E}[f(X)]$ (the unconditional expectation). The conditional expectation is denoted by $\mathbb{E}[f(X)\,|\,Y=y]$. If X and Y have a joint density, then the conditional expectation is computed using the conditional density as follows:

$$\mathbb{E}[f(X)\,|\,Y=y] \;=\; \int f(x)\,p(x|y)\,dx$$

If X and Y are independent random variables, then $\mathbb{E}[XY]=\mathbb{E}[X]\,\mathbb{E}[Y]$, a simple consequence of the fact that the joint density or mass function factorizes.

5 Convergence of Sums of Independent Random Variables

The most important form of statistic considered in this course is a sum of independent random variables.

Example 10 *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 $\widehat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$. First note that $\mathbb{E}[\widehat{\mu}]=\mathbb{E}[X]$. An easy calculation shows that the variance of $\widehat{\mu}$ is σ^2/n. So the average has the same mean value as the random variables and the variance is 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).

The argument above quantifies the effect of averaging on the variance, but often we would like to say more about the distribution of the average. The *Central Limit Theorem* is a classic result showing that the probability distribution of the average of n independent and identically distributed random variables with mean μ and variance $\sigma^2<\infty$ tends to a Gaussian distribution with mean μ and variance σ^2/n, regardless of the form of the distribution of the variables. By 'tends to' we mean in the limit as n tends to infinity.

In many applications we would like to say something more about the distributional characteristics 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 $t>0$. Then

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

The result $\mathbb{P}(Z\geq t)\leq \mathbb{E}[Z]/t$ is called *Markov's Inequality.* Now we can use this to get a bound on the probability 'tails' of Z. Let $t>0$

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

where Var(Z) denotes the variance of Z. This inequality is known as *Chebyshev's Inequality.* If we apply this to the average $\mu = \frac{1}{n}\sum_{i=1}^n X_i$, 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.

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 exponential fast in n. The simplest result of this form is for bounded random variables.

Theorem 1 *(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 next section.

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

$$\mathbb{P}(|\widehat{\mu} - \mu| \geq t) \quad \leq \quad 2\, e^{-\frac{2nt^2}{c}} \tag{1}$$

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.*

Theorem 2 *(Weak Law of Large Numbers). Let $X_1, X_2, ..., X_n$ be iid random variables with $\mathbb{E}[|X_i|] < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to $\mathbb{E}[X_i]$.*

Theorem 3 *(Strong Law of Large Numbers). Let $X_1, X_2, ..., X_n$ be iid random variables with $\mathbb{E}[|X_i|] < \infty$. Then $\frac{1}{n}\sum_{i=1}^n X_i$ converges almost surely $\mathbb{E}[X_i]$.*

Example 11 *Let us revisit the synthia experiments. 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 (1) above is equal to 1. For desired accuracy $\epsilon > 0$ and confidence $1 - \delta$, how many experiments will be sufficient? From (1) 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.

6 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 noting 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 $\sum_{i=1}^n X_i - \mathbb{E}[X_i]$. Then

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

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 $\mathbb{E}\left[e^{s(X_i - E[X_i])}\right]$.

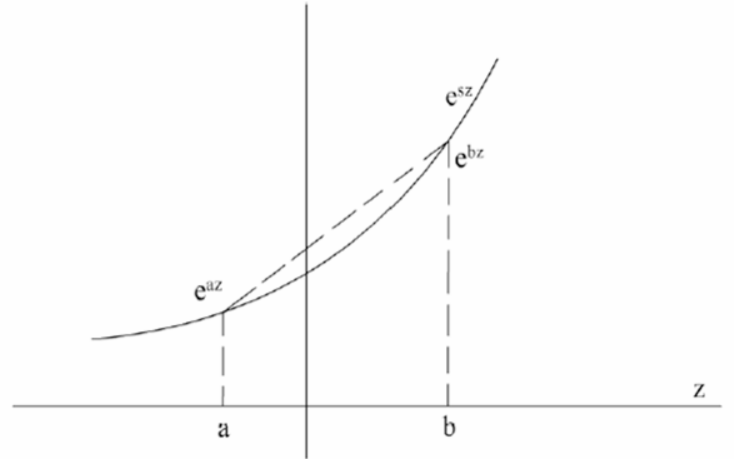

Figure 2: Convexity of exponential function.

Lemma 1 *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 (see Fig. 2),

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

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:

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

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

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 $\phi''(u) \leq 1/4$. 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.

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

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.