**EE-411: Fundamentals of inference and learning** **2022-2023**

# Lecture 1: All of probability

*Lecturer: Florent Krzakala*

In this lesson, we are learning as much probability theory that we can in a single lesson. Of course, the title "All of probablity" should be taken with a grain of salt.

## 1.1 Probability and probability distributions

Following the use in probability theory, we shall denote with capital letters, for instance $X$, a random variable, and with small letters, for instance, $x$, a particular realization of this random variable.

Let $X$ be a random variable. It can be a scalar (i.e. $X \in \mathbb{R}$), a complex variable (i.e. $X \in \mathbb{C}$) or even a real vector (i.e. $X \in \mathbb{R}^d$). What we mean, informally, by a random variable is that we have access to a system that give us particular values of $X$. For instance, a dice is a system that return fairly (hopefully) a number from 1 to 6 every time we use it.

Fundamentally, random variables have densities, which is the most important mathematical object associated to them:

**Definition 1.1 (Probability density functions)** *The probability density function (PDF), or density, is a non-negative function that describes the relative likelihood for this random variable to take on a given value such that*

$$\int p_X(x)dx = 1 \qquad \textit{Normalization}$$

$$p_X(x)dx = \mathbb{P}(X \in [x\,;x+dx])$$

$$p_X(x) \geq 0$$

*Often, it is written that $X \sim p_X(x)$ (which reads as $X$ is sampled from $p_X$).*

Careful: with continuous variable $p(x)dx$ is a probability, but NOT $p(x)$. In particular, $p(x)$ can be LARGER than 1.

**Definition 1.2 (Cumulative density functions)** *Let $X$ be a random variable. The cumulative density function is a real-valued function given by:*

$$F_X(x) = \int_{-\infty}^{x} p_X(u)\,\mathrm{d}u = \mathbb{P}(X \leq x)$$

**Definition 1.3 (Expectation or expected value)** *Let $X$ be a random variable of probability density function $p_X(x)$, then the expected value can be computed as:*

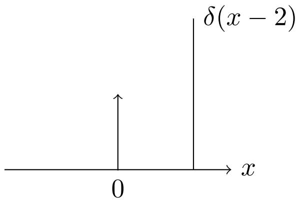
$$\mathbb{E}_X[g(X)] = \int_{-\infty}^{\infty} g(x) p_X(x) \, \mathrm{d}x$$

$\mathbb{E}_X(X)$, is named the "mean". $\mu_n = \mathbb{E}_X(X^n)$ is named the "n-th moment".

**Definition 1.4 (Variance and Standard deviation)** *The variance $\Delta$ and the standard deviation $\sigma$ of a random variable $X$ are defined as:*

$$\Delta = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \qquad \sigma = \sqrt{\Delta}$$

Let us describe some example of random variable and their density functions:



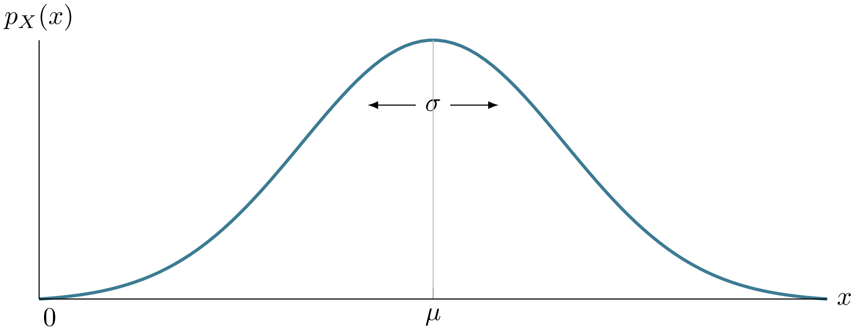
**Example 1 (Dirac distribution (or point-mass))** ■

In this example, the mean is 2 and the variance is 0. The pdf is given by

$$p_X(x) = \delta(x-2) \tag{1.1}$$

**Example 2 (Gaussian (or Normal) distribution)** *The gaussian distribution of mean $\mu$ and variance $\Delta = \sigma^2$ has the following form:*

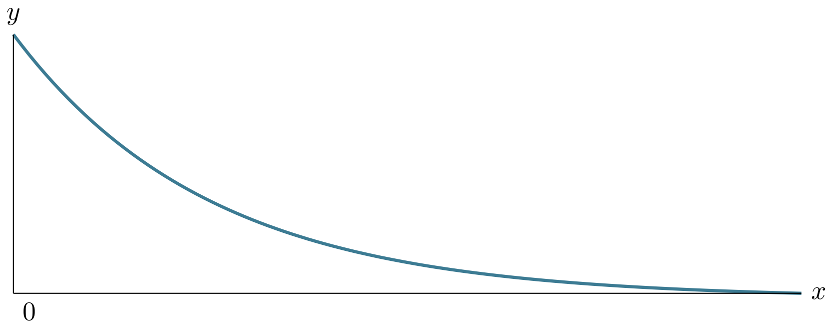
$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(x-\mu)^2}{2\sigma^2} \sim \mathcal{N}(\mu, \sigma^2)$$



■

**Example 3 (Exponential distribution)** *The exponential distribution of mean $\mu$ and a variance $\mu^2$ has the following form:*

$$p_X(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$



■

**R** We can add several PDF between them, but in this case we need to care about normalization. For example:

$$p_X(x) = \frac{1}{6}\sum_{i=1}^{6}\delta(x-i)$$

The exponential distribution is a probability distribution that describes the time between events in a Poisson process. For instance, the probability that a bulb breaks at any given time, say between $t$ and $t+\mathrm{d}t$ is $\lambda\,\mathrm{d}t$ with $\lambda = \frac{1}{\mu}$. Thus, we can obtain:

$$\mathbb{P}(\text{not broken between } t \text{ and } t+\mathrm{d}t) = 1-\lambda\,\mathrm{d}t$$

$$\begin{aligned}\mathbb{P}(\text{not broken between } 0 \text{ and } T\text{, breaking between } T \text{ and } T+dt) &\propto (1-\lambda\,\mathrm{d}t)^{\frac{T}{\mathrm{d}t}}\lambda\,\mathrm{d}t\\ &\simeq e^{-\lambda T}\lambda\,\mathrm{d}t\end{aligned}$$

Thus, we obtain the probability distribution with $\mu = \frac{1}{\lambda}$ when modeling such events.

**R** The mean and variance do not always exist. A power-law $x^{-k}$ has a well-defined mean over $x \in [1,\infty)$ only if $k > 2$ , and it has a finite variance only if $k > 3$.

## 1.2 Some basic properties

### 1.2.1 Densities of Transformations of Random Variables Using delta-function

It is very useful to use delta functions to transform and change variables. For instance, if $X$ is distributed as $p_X(x)$, and one wondering the distribution $p_Y(y)$ of the variable $Y = g(X)$, then one can write

$$p_Y(y) = \int dx p_X(x)\delta(y-g(x))$$

Combining this with the classic result for the delta function

$$\delta(f(x)) = \sum_i \frac{\delta(x-x_i)}{|f'(x_i)|} \quad \forall\, x_i \text{ such that } f(x_i) = 0$$

one finds that for all $x_i$ solutions of $g(x_i) = y$

$$p_Y(y) = \sum_i \int dx p_X(x) \frac{\delta(x - x_i)}{|g'(x_i)|} = \sum_i \frac{p_X(x_i)}{|g'(x_i)|}$$

**Proposition 1.5 (Change of variables)** *Let $g$ be a monotonous function, $p_X(x)$ a PDF and $y = f(x)$. Then the probability distribution of $y$ is expressed as:*

$$p_X(x) = p_Y(y) \left| \frac{\mathrm{d}y}{\mathrm{d}x} \right| \tag{1.2}$$

**Proof 1** *A simple "geometric" proof is based on the fact that the probability contained in a differential area must be invariant under change of variables. Or in other words:*

$$|p_Y(y)\,\mathrm{d}y| = |p_X(x)\,\mathrm{d}x|$$

*Further calculations lead to another formula:*

$$p_Y(y) = \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right| p_X(x) = \left|\frac{\mathrm{d}}{\mathrm{d}y}(x)\right| p_X(x) = \left|\frac{\mathrm{d}}{\mathrm{d}y}(g^{-1}(y))\right| p_X(g^{-1}(y)) = \frac{p_X(g^{-1}(y))}{|g'(g^{-1}(y))|}$$

A simple example is given by the following transformation: Say for instance that $X$ is distributed uniformly on $]0, 1]$, then the distribution of $Y = -\ln X$ is given by

$$p_Y(y) = |\frac{dx}{dy}| = e^{-y}$$

with $Y$ in $]0, \infty[$.

R If $f$ is not monotonic, we need to sum on all the $x$ variable.

$$\sum_{k=1}^{n(y)} \left| \frac{\mathrm{d}}{\mathrm{d}y} f_k^{-1}(y) \right| \dot{p}_X(f_k^{-1}(y))$$

with $n(y)$ the number of solutions of $f(x) = y$.

Another classic use of the delta function is to add variables. For instance if $Z = X + Y$ then

$$p_Z(z) = \int dxdy p_X(x) p_Y(y) \delta(z - (x + y))$$

## 1.3 Basic bounds

**Proposition 1.6 (Markov inequality)** *This inequality gives an upper bound for the probability that a non-negative function of a random variable is greater than or equal to some positive constant. If $X$ is a non-negative random variable and $a > 0$.*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{1.3}$$

**Proof 2** *The proof of this inequality is obtained directly with the definition of cumulative function:*

$$\begin{aligned}\mathbb{P}(X \geq a) &= \int_a^{+\infty} p_X(x)\,\mathrm{d}x \\ &\leq \int_a^{\infty} \frac{x}{a} p_X(x)\,\mathrm{d}x \quad \textit{because } x \geq a \\ &\leq \frac{\mathbb{E}[X]}{a} \quad \textit{because } x p_X(x) \geq 0\end{aligned}$$

**Proposition 1.7 (Chebyshev's inequality)** *This inequality uses the variance to bound the probability that a random variable, with mean and variance, deviates far from the mean:*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) \leq \frac{1}{k^2} \tag{1.4}$$

**Proof 3** *The Chebyshev's inequality follows from Markov's inequality by considering the random variable* $(X - \mathbb{E}[X])^2$ *and* $k^2\Delta$*:*

$$\begin{aligned}\mathbb{P}(|X - \mathbb{E}[X]| \geq k\sigma) &= \mathbb{P}((X - \mathbb{E}[X])^2 \geq k^2\Delta) \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{k^2\Delta} = \frac{\Delta}{k^2\Delta} \quad \textit{using Markov inequality} \\ &\leq \frac{1}{k^2}\end{aligned}$$

It is important to remember that these are bounds! Most of the time, the probability decay much faster than Chebyshev tells us, but Chebyshev tells us that it decays *at least* as $1/k^2$! This is what we call a worst case bound.

## 1.4 Cumulants, Connected moments, and moment generative functional

**Definition 1.8 (Moment generating function (MGF))** *Let* $X$ *be a random variable. The moment generating function is defined as:*

$$M_X(t) = \mathbb{E}[e^{tX}] = \underbrace{\int_{-\infty}^{\infty} \mathrm{d}x p_X(x) e^{tx}}_{\textit{Laplace transform}} \tag{1.5}$$

This function is called MGF because of the following property:

**Proposition 1.9**

$$\left.\frac{\partial^n M_X(t)}{\partial t^n}\right|_{t=0} = \int_{-\infty}^{\infty} \mathrm{d}x\; x^n p_x(x) = \mathbb{E}[X^n]$$

**Definition 1.10 (Characteristic function)** *The characteristic function of any real-valued random vari-*

*able completely defines its probability distribution. The function is defined as:*

$$\varphi(t) = \mathbb{E}[e^{itX}] = \underbrace{\int_{-\infty}^{\infty} \mathrm{d}x p_X(x) e^{itx}}_{\textit{Fourier transform}} \tag{1.6}$$

*Note that the characteristic function always exists* [1]. *Additionally, Levy continuity theorem ensures a sequence $X_n$ converges in distribution to $X$ if and only if the sequence of corresponding characteristic functions converges pointwise to the characteristic function of $X$.*

**Proposition 1.11** *As for the MGF, we have the following property:*

$$\left.\frac{\partial^n \varphi_X(t)}{\partial t^n}\right|_{t=0} = i^n \mathbb{E}[X^n]$$

**Definition 1.12 (Cumulant)** *The cumulant $K_n$ of a probability distribution is a set of quantities that provides alternatives to the moments of the distribution. They are defined via the cumulant generating function $K_X(t)$:*

$$K_X(t) = \ln \mathbb{E}[e^{tX}] = \ln M_X(t)$$

*Indeed, the cumulant is the power expansion of the cumulant generating function:*

$$K_X(t) = \sum_n \kappa_n \frac{t^n}{n!}$$

*Or in other words:*

$$\kappa_n = \frac{\partial^n K_X(t)}{\partial t^n}|_{t=0} = \frac{\partial^n \ln M_X(t)}{\partial t^n}|_{t=0} \tag{1.7}$$

With this definition we can compute the cumulants of any probability distribution. As examples, we will give the two first:

$$\kappa_1 = \frac{M_X'}{M_X}|_{t=0} = \mathbb{E}[X]$$

$$\kappa_2 = \frac{M_X''(0) M_X(0) - M_X'(0)^2}{M_X(0)^2} = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \Delta$$

However, then becomes more complicated after this. For instance $\kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2{\mu_1}^3$ and $\kappa_4 = \mu_4 - 4\mu_3\mu_1 - 3{\mu_2}^2 + 12\mu_2{\mu_1}^2 - 6{\mu_1}^4$ with $\mu_n = \mathbb{E}[X^n]$. Working over cumulants instead of moments has an advantage because of the following property:

**Proposition 1.13** *Let $X$ and $Y$ be independent random variables* [2]. *Then the cumulant of the sum is the sum of the corresponding cumulants, i.e.*

$$\kappa_n(X+Y) = \kappa_n(X) + \kappa_n(Y)$$

[1] This is not the case for the MGF that may not actually exist (this means that the integral diverges) if the integral is not well defined, which happens if the distribution does not decay fast enough "at infinity". The characteristic function, however, always exists: indeed we have $|p_X(x)e^{itx}| \leq p_X(x)$ so that $\int |p_X(x)e^{itx}| dx \leq \int p_X(x) dx = 1$. Since then $p_X(x)e^{itx}$ is absolutely integrable, it is also integrable.

[2] Two random variables $X$ and $Y$ are independent if $p_{X,Y}(x,y) = p_X(x)p_Y(y)$

**Proof 4** *Let $Z = X + Y$, with $X$ and $Y$ defined as before. Then, we have:*

$$p_Z(z) = \int \mathrm{d}x \int \mathrm{d}y p_X(x) p_Y(y) \delta(z - x - y)$$

*We now want to evaluate the MGF:*

$$\begin{aligned} M_Z(t) &= \int \mathrm{d}z p_Z(z) e^{tz} \\ &= \int \mathrm{d}z \int \mathrm{d}x \int \mathrm{d}y p_X(x) p_Y(y) e^{t(x+y)} \delta(z - x - y) \\ &= M_X(t) M_Y(t) \quad \textit{integrate over } z \end{aligned} \tag{1.8}$$

*If we take the logarithm of 1.8:*

$$\ln M_Z = \ln M_X + \ln M_Y \Leftrightarrow \kappa_n(Z) = \kappa_n(X) + \kappa_n(Y)$$

Careful: This is not true for moments.

$$\mu_n(X + Y) \neq \mu_n(X) + \mu_n(Y)$$

## 1.5 Convergence of a sequence of random variable

### 1.5.1 Convergence in law

Roughly speaking, a sequence of random variables converges in distribution to a random variable X if

$$\lim_{n\to\infty} \Pr(X_n \in A) = \Pr(X \in A) \text{for every} A \tag{1.9}$$

Convergence in law (or weak convergence, or convergence in distribution) distribution may be denoted in a different way as the author of a book sees fit:

$$X_n \xrightarrow{d} X, \quad X_n \xrightarrow{\mathcal{D}} X, \quad X_n \xrightarrow{\mathcal{L}} X, \quad X_n \xrightarrow{d} \mathcal{L}_X, X_n \rightsquigarrow X, \quad X_n \Rightarrow X, \quad \mathcal{L}(X_n) \to \mathcal{L}(X) \tag{1.10}$$

A fundamental theorem (Levy continuity theorem) tells us that a sequence $X_n$ converges in distribution to $X$ if and only if the sequence of corresponding characteristic functions converges pointwise to the characteristic function of $X$.

### 1.5.2 Convergence in probability

Convergence in probablity happens when a sequence of random variable converge to a well defined value. A sequence $X_n$ of random variables converges in probability towards the random variable X if for all $\epsilon > 0$

$$\lim_{n\to\infty} \Pr\left(|X_n - X| > \varepsilon\right) = 0. \tag{1.11}$$

It is denoted as

$$X_n \xrightarrow{p} X, \quad X_n \xrightarrow{P} X, \quad \underset{n\to\infty}{plim}\, X_n = X. \tag{1.12}$$

Let us now see the two most important examples of these two types of convergence:

### 1.5.3 Law of large numbers

Let $(X_n)$ be a sequence of i.i.d. random variables with mean $\mu$ (that is assumed to exist). Then the empirical mean $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$ converge in probability to $\mu$: $X_n \xrightarrow{p} X$.

This means that for any $\varepsilon > 0$, we have $\lim\limits_{n\to+\infty} \mathbf{P}\left(|Y_n - \mu| > \varepsilon\right) = 0$.

**Proof 5** *We shall assume, for simplicity, that the variance $\Delta$ of the distribution exists. In this case, the proof is a simple consequence of the additivity of cumulants and Chebyshev inequality. Indeed the random variable $Y_n = \frac{X_1+X_2+\cdots+X_n}{n}$ has an expected mean $\mu$ and variance $n\Delta_X/n^2 = \Delta_X/n$ . Thus for any $n$, we have*

$$\mathbb{P}\left(\left|\frac{X_1+X_2+\cdots+X_n}{n} - E(X)\right| \geqslant \varepsilon\right) \leqslant \frac{\Delta_X}{n\,\varepsilon^2}$$

.

*Note that the Law of large number is also valid when the variance does not exists. In this case, the proof is more complicated, and works by using the characteristic function and Levy continuity theorem.*

Informally, this means that if the variance exists then:

$$\frac{1}{n}\sum_{k=1}^{n} X_k \approx \mu \pm \sqrt{\frac{\Delta}{n}}$$

### 1.5.4 Central limit theorem

The central limit theorem establishes the convergence in law of the sum of a series of random variables towards the normal, or Gaussian, law. Intuitively, this result asserts that any sum of independent random variables tends in some cases towards a Gaussian random variable:

Let $(X_n)$ be a sequence of i.i.d. random variables with mean $\mu$ and variance $\Delta$. Consider the empirical mean $Y_n = \frac{1}{n}\sum_{k=1}^{n} X_k$. Then the random variable

$$S_n = \frac{\sqrt{n}(Y_n - \mu)}{\sqrt{\Delta}}$$

converges in law to a Gaussian Normal random variable $S_n \rightsquigarrow X\mathcal{N}(0,1)$.

**Proof 6** *As with the previous theorem, we can simplify the proof by assuming that all cumulants of $X$ exists. In this case, all the $S_n$ tends to zero with $n$ except for the second one $\kappa_2^{S_n} \to 1$. Therefore the characteristic function is given by $e^{-t^2/2}$. This is the Fourier transform of a Gaussian random variable, and Levy continuity theorem thus proves that $S_n = \frac{\sqrt{n}(Y_n-\mu)}{\sqrt{\Delta}}$ in law.*

*Again, the theorem does not acutally requires all the cumulants to exists, but only the mean and variance. In this case the proof is slightly (but not quite) more complicated.*

## 1.6 Chernoff bounds

### 1.6.1 Generic bound and Cramér-theorem

There is a very important theorem, due to Chernoff and Cramér, that allows us to improve A LOT over the simple Chebyshev bound for averages over many variables:

**Theorem 1.14 (Chernoff-type Bound and Cramér-Chernoff theorem)** *Let $X_1, X_2, \dots, X_N$ be independent, identically distributed random variables, and let $K_X(t)$ be their cumulant generating function $K_X(t) = \ln(\mathbb{E}[e^{tX}])$. Then many standard concentration inequalities can be derived from the following single basic result: For any $\lambda > 0$:*

$$\mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq a\right) \leq e^{-N(\lambda a - K_X(\lambda))} \tag{1.13}$$

*Moreover, the Cramér-Chernoff theorem states that this bounds asymptotically achieves the best possible exponent if we optimize our choice of $\lambda$:*

$$\lim_{N\to\infty} -\frac{\ln \mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq a\right)}{N} = \sup_{\lambda>0}(\lambda a - K_X(\lambda)) \tag{1.14}$$

**Proof 7** *The proof the Chernoff bound is a simple application of Markov's inequality: first we write (for positive $\lambda$)*

$$\mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq a\right) = \mathbb{P}\left(\sum_i X_i \geq Na\right) = \mathbb{P}\left(\lambda \sum_i X_i \geq \lambda N a\right) = \mathbb{P}\left(e^{\lambda \sum_i X_i} \geq e^{\lambda N a}\right)$$

*Then, we apply Markov and get*

$$\mathbb{P}\left(e^{\lambda \sum_i X_i} \geq e^{\lambda N a}\right) \leq \frac{\mathbb{E}\left[e^{\lambda \sum_i X_i}\right]}{e^{\lambda N a}} = \left(\mathbb{E}\left[e^{\lambda X}\right]\right)^N e^{-\lambda N a} = e^{-N(\lambda a - K_X(\lambda))}$$

### 1.6.2 Sub-Gaussian variables and the Hoeffding bound

Let us consider the simplest case where all $X_i$ are Gaussians with mean $m$ and variance $\Delta$. In this case, one can compute the cumulant generating function that reads $K_X(t) = mt + \frac{\Delta}{2}t^2$. We can use the bound to estimate

$$\mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq m + \epsilon\right) \leq e^{-N(\lambda(m+\epsilon) - K_X(\lambda))}$$

The best $\lambda$ is given by $\lambda^{\text{best}} = \epsilon/\Delta$ so that

$$\mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq m + \epsilon\right) \leq e^{-N\frac{\epsilon^2}{2\Delta}}$$

This is much better than what we get from Chebyshev! Let us ask for instance the probability that we could reach a result that off by an amount$\epsilon$:

$$\mathbb{P}(\hat{m} \geq m + \epsilon) \leq e^{-n\frac{\epsilon^2}{2\Delta}} \tag{1.15}$$

$$\mathbb{P}(\hat{m} \leq m - \epsilon) \leq e^{-n\frac{\epsilon^2}{2\Delta}} \tag{1.16}$$

and so

$$\mathbb{P}(|\hat{m}-m|> \epsilon) \leq 2e^{-n\frac{\epsilon^2}{2\Delta}} \tag{1.17}$$

Using $\delta = 2e^{-n\frac{\epsilon^2}{2\Delta}}$, or equivalently

$$n = \frac{2\Delta}{\epsilon^2}\log\frac{2}{\delta}$$

and solving for $\epsilon$ leads to the following statement:

**Theorem 1.15 (PAC learning for Gaussian's mean)** *With probability at least $1-\delta$, we have*

$$|\hat{m}-m|\leq \sqrt{\frac{2\Delta\log\frac{2}{\delta}}{N}} \tag{1.18}$$

This is an example of PAC learning = "probably approximately correct learning". This is a fundamental concept that will come back many time during the lecture.

Note that **confidence is cheap** (delta decay exponentially with $N$) while **accuracy is expensive** ($\epsilon$ decay as the inverse of the square root of $n$).

Motivated by the structure of this example, we are led to introduce the following definition:

**Definition 1.16 (Sub-Gaussians variables)** *A random variable $X$ with mean $m = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number $\sigma$ such that*

$$\mathbb{E}[e^{\lambda(X-m)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}$$

*for all $\lambda \in \mathbb{R}$*

The constant $\sigma$ is referred to as the sub-Gaussian parameter ; for instance, we say that $X$ is sub-Gaussian with parameter $\sigma$ when the condition holds. Naturally, any Gaussian variable with variance $\sigma^2$ is sub-Gaussian with parameter $\sigma$. In addition, a large number of non-Gaussian random variables also satisfy the condition.

For instance, one can show that any bounded random variable between $[a;b]$ is sub-Gaussian with parameter $\sigma = (b-a)/2$ (this is left as exercise). In fact, when a distribution has tails that decay at least like a Gaussian, then the variable is sub-Gaussian. More formally: if there is a constant $c \geq 1$ and a Gaussian random variable $Z \sim \mathcal{N}(0,\sigma^2)$ such that $\mathbb{P}[|X|\geq s] \leq c\mathbb{P}[|Z|\geq s]$ for all $s \geq 0$, then the variable is Sub-Gaussian with parameter $\sigma$.

In this case, the application of the Chernoff bounds leads to the so-called Chernoff-Hoeffding bound:

**Theorem 1.17 (Hoeffding bound)** *Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed sub-Gaussian random variables of parameter $\sigma$. Then*

$$\mathbb{P}\left(\frac{1}{N}\sum_i X_i \geq m+\epsilon\right) \leq e^{-N\frac{\epsilon^2}{2\sigma^2}} \tag{1.19}$$

If we apply the Hoeffding bound to the Bernoulli random variables of the previous example, we get the following results:

$$\mathbb{P}\left(\hat{p} \geq p+\epsilon\right) \leq e^{-2N\epsilon^2}$$

This is much better than what we get from Chebyshev! Let us ask for instance the probability that we could reach a result that larger than the true $p$ by an amount that is $k/\sqrt{N}$ then

$$\mathbb{P}\left(\hat{p} \geq p + \frac{k}{\sqrt{N}}\right) \leq e^{-2k^2}$$

This is much stronger than the previous result! We see that the probability to make a large mistake decay exponentially, so indeed, we know that the error, with high probability, should not be larger than $k$ times the variance of the estimator. This is why we do trust Monte Carlo simulation! so that

**Theorem 1.18 (PAC learning for Bernoulli variables)** *With probability at least $1-\delta$, we have*

$$|\hat{p} - p| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2N}} \tag{1.20}$$

### 1.6.3 Chernoff bound for Bernoulli variables

We could get another very precise result by directly applying the Chernoff bound to a Bernoulli variable:

**Theorem 1.19 (Chernoff bound for Bernoulli variables)** *Let consider a random variable $\hat{p} = \frac{1}{N}\sum_{i=1}^{N} X_i$, with $X_i$ independent Bernoulli random variables, then:*

$$P\left(\hat{p} \geq p + \varepsilon\right) \leq e^{-ND_{KL}(p+\varepsilon||p)} \tag{1.21}$$

This is an example of the Sanov theorem (which is somehow equivalent to the Chernoff one) that illustrates the importance of the Kullback-Leibler divergence $D_{KL}(p+\varepsilon||p)$.

**Proof 8** *Take the exponential and raise to power $N$:*

$$P\left(\hat{p} \geq p + \varepsilon\right) = P\left(e^{t\hat{p}} \geq e^{t(p+\varepsilon)}\right) = P\left(e^{t\sum_i X_i} \geq e^{Nt(p+\varepsilon)}\right)$$

*Then, we apply Markov's inequality:*

$$P\left(\hat{p} \geq p + \varepsilon\right) \leq \frac{\mathbb{E}\left[e^{t\sum_i X_i}\right]}{e^{N(p+\varepsilon)t}} = \frac{\mathbb{E}\left[e^{tX}\right]^N}{e^{N(p+\varepsilon)t}}$$

*where we have used independence of $X_i$ and the numerator is the MGF. For a Bernoulli variable $\mathbb{E}[e^{tX}] = pe^t + (1-p)$, therefore:*

$$P\left(\hat{p} \geq p + \varepsilon\right) \leq \frac{\left(pe^t + (1-p)\right)^N}{\left(e^{(p+\varepsilon)t}\right)^N}$$

*Minimizing the right hand side with respect to $t$, one gets $t^* : e^{t^*} = \frac{(1-p)(p+\varepsilon)}{p(1-p-\varepsilon)}$. Inserting this in the inequality:*

$$P\left(\hat{p} \geq p+\varepsilon\right) \leq \left[\left(\frac{p}{p+\varepsilon}\right)^{p+\epsilon}\left(\frac{1-p}{1-p-\varepsilon}\right)^{1-p-\varepsilon}\right]^{N} = e^{-ND_{KL}(p+\varepsilon||p)}$$

*which is the Chernov bound, where the Kullback-Leibler divergence (or relative entropy) has been defined: D˙KL*

$$D_{KL}(p||q) = \int p(x)\log\left(\frac{p(x)}{q(x)}\right)dx$$

*The probability of making a mistake decays exponentially, so Markov and Chebyshev are very poor bounds for this distribution.*

**Proposition 1.20 (Gibbs' inequality)** $D_{KL}(p||q) \geq 0$.

**Proof 9** *Start from*

$$x-1 \geq \log(x) \quad \forall\ x>0$$

*then*

$$-D_{KL}(p||q) = \int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx \leq \int dx p(x)\left[\frac{q(x)}{p(x)}-1\right] = \int dx q(x) - \int dx p(x) = 0$$

*since $q(x)$ and $p(x)$ are normalized probability distributions. Therefore:*

$$D_{KL}(p||q) \geq 0$$

R Gibbs' inequality says that the Kullback-Leiber divergence is always positive. In fact, the proof says more: it says that $D_{KL}(p||q) = 0$ if and only if $p = q$. This is because the inequality we used ($x-1 \geq \log(x)$) is true if and only if $x = 1$, or in other terms $q(x) = p(x)$ for all $x$.

This fact says that the Kullback-Leiber divergence measures how different are two probability distributions $p$ and $q$, because it is equal to zero only when they are the same. However, it is not a distance in the mathematical sense, because it is not symmetric: $D_{KL}(p||q) \neq D_{KL}(q||p)$ in general. Nevertheless, it is very useful in comparing probability distributions.