Lecture Notes

Probability and Beyond

January 24, 2017

ARSHAK MINASYAN

1 Basic Probability

This lecture is an overview of basic probability concepts, notations that we will be using along the course.

1.1 Probability Spaces

A triple $(\Omega, \mathcal{F}, \mathbb{P})$ is called a probability space, where Ω is the set of outcomes, $\mathcal{F}$ is a set of all possible events and $\mathbb{P}$ is function from $\mathcal{F}$ to $[0,1]$ that assigns probabilities to events, precisely $\mathbb{P} : \mathcal{F} \to [0,1]$. We restrict ourselves with some assumptions on $\Omega, \mathcal{F}$ and $\mathbb{P}$: $\mathcal{F}$ is a σ-algebra of set Ω and $\mathbb{P}$ is a *probability measure.*

We define a measure μ as follows, it is a function $\mu : \mathcal{F} \to \mathbb{R}$ with the following properties:

- $\mu(A) \geq \mu(\varnothing)$ for all $A \in \mathcal{F}$
- if $A_i \in \mathcal{F}$ is a countable (finite or countably infinite) sequence of disjoint sets, then

$$\mu(\cup A_i) = \sum_i \mu(A_i) \tag{1}$$

A measure becomes *probability measure* if $\mu(\Omega) = 1$. The following theorem is stated without proof and contains the basic properties of measure μ defined on $(\Omega, \mathcal{F})$.

Theorem 1.1. *Let μ be a measure on $(\Omega, \mathcal{F})$ then*

(a) ***monotonicity.*** *If $A \subset B$ then $\mu(A) \leq \mu(B)$.*

(b) ***sub-additivity.*** *If $A \subset \cup_{m=1}^{\infty} A_m$ then $\mu(A) \leq \sum_{m=1}^{\infty} \mu(A_m)$.*

(c) ***continuity from below.*** *If $A_i \uparrow A$, i.e. $A_1 \subset A_2 \subset \ldots, \cup A_i = A$ then $\mu(A_i) \uparrow \mu(A)$.*

(d) ***continuity from above.*** *If $A_i \downarrow A$, i.e. $A_1 \supset A_2 \supset \ldots, \cap A_i = A$ and $\mu(A_1) < \infty$ then $\mu(A_i) \downarrow \mu(A)$.*

The probability spaces can be either discrete or continuous. If the set of outcomes Ω is countable, i.e. either finite or countably infinite, then the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is discrete, otherwise it is continuous.

The natural example of discrete probability space is $\Omega = \{0, 1\}$. More unintuitive one could be $\Omega = \mathbb{N}$ (set of natural numbers). The set of natural numbers is indeed infinite, but it is countable, hence it is still discrete.

The probability measure on $\Omega = \{0, 1\}$ is, in general, $\mu(\{1\}) = p$ and $\mu(\{0\}) = q = 1 - p$, where $p \in [0, 1]$. It is straightforward to check that the given measure is indeed a probability measure.

Remark 1.1. *It is very natural to think of the mixture of above described two spaces, i.e. the space is somewhere continuous and somewhere discrete.*

1.2 Random Variables

Probability spaces become more interesting when we define random variables on them. A random variable X on Ω is a measurable function $X : \Omega \to T$, where T is some set of values that random variable can possibly take. $\mathcal{F}$-measurable or just measurable means that for every Borel set $B \subset T$ it holds $X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{F}$. Another simple, but useful, example random variable is *indicator function* of a set $A \in \mathcal{F}$:

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases} \tag{2}$$

Any random variable X induces a probability measure on $\mathbb{R}$ known as distribution of random variable X, i.e. $\mu(A) = \mathbb{P}(X \in A)$ for Borel sets A. The distribution of random variable X is usually described by giving its *distribution function*, $F(x) = \mathbb{P}(X \leq x)$.

Theorem 1.2. *Any distribution function $F(x)$ has the following properties:*

(a) F is non-decreasing

(b) $\lim_{x\to+\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

(c) F is right continuous, we denote it as $\lim_{y\downarrow x} F(y) = F(x)$.

(d) Define $F(x-) = \lim_{y\uparrow x} F(y)$, then $F(x-) = \mathbb{P}(X < x)$.

(e) $\mathbb{P}(X = x) = \mathbb{P}(X \leq x) - \mathbb{P}(X < x) = F(x) - F(x-)$.

Proof.

(a)

■

The following exercise is known as St Petersburg paradox.

Exercise 1.1. *A fair coin is tossed repeatedly. Let T be the number of tosses until the first head. You are offered the following prospect, which you may accept on payment of a fee. If $T = k$, say, then you will receive 2^k roubles. What would be a 'fair' fee to ask of you?*

1.3 Independence of random variables

1.4 Convergence of random variables

There are mainly 4 types of convergence: in probability, almost surely, weak convergence (in distribution) and in mean.

1.5 Law of Large Numbers

1.6 Central Limit Theorem

1.7 Distributions

2 Probability Inequalities

This section contains a number of well known and useful probability inequalities.

Theorem 2.1. ***Markov's inequality***
Let X be a non-negative random variable and suppose that $\mathbb{E}X$ exists. For any $t > 0$,

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t} \tag{3}$$

Proof. Since $X > 0$ a.s.

$$\mathbb{E}X = \int_0^{+\infty} xp(x)\,dx = \int_0^t xp(x)\,dx + \int_t^{+\infty} xp(x)\,dx \geq t\int_t^{+\infty} p(x)\,dx = t\mathbb{P}(X > t)$$

■

Chebyshev's inequality is follows straight from Markov's inequality by simple plugging-in $|X - \mathbb{E}X|$ instead of X. More precisely it looks as follows

Theorem 2.2. ***Chebyshev's inequality***
Let $\mu = \mathbb{E}X$ and $\mathbb{V}ar(X) = \sigma^2 < +\infty$, then for any $t > 0$

$$\mathbb{P}(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2} \tag{4}$$

Proof. Once again, plug in $|X - \mathbb{E}X|$ instead of X in (3). ■

As we have introduced the law of large numbers, one can easily prove the weak law by applying Chebyshev's inequality to it. (do it as an exercise)

Another modification of Markov's inequality is known as Chernoff's inequality or bound.

Theorem 2.3. ***Chernoff's bound***
Let X is a random variable, then

$$\mathbb{P}(X \geq \epsilon) \leq \inf_{t \geq 0} e^{-t\epsilon}\mathbb{E}e^{tX}. \tag{5}$$

Proof. For any $t > 0$,

$$\mathbb{P}(X \geq \epsilon) = \mathbb{P}(e^{tX} \geq e^{t\epsilon}) \leq e^{-t\epsilon}\mathbb{E}e^{tX}. \tag{6}$$

And since, it is true for any $t \geq 0$, then it should be true for the infimum, which completes the proof. ■

A more sharper inequality of this kind is Hoeffding's inequality, but before presenting Hoeffding's inequality we introduce one auxiliary lemma. This lemma plays crucial role in the proof of Hoeffding's inequality and is very useful in many other cases.

Lemma. *Suppose $a \leq X \leq b$ a.s. and $\mathbb{E}X = \mu$, then*

$$\mathbb{E}e^{tX} \leq e^{t\mu} \cdot e^{\frac{t^2(b-a)^2}{8}} \tag{7}$$

Proof. Without loss of generality (WLOG) we can assume that $\mu = 0$, in other words, from proof for $\mu = 0$ it is straightforward to prove for random variable X non-zero mean. Then, from $a \leq X \leq b$ a.s. one can write

$$X = \alpha b + (1 - \alpha)a,$$

where $\alpha = \frac{X-a}{b-a}$. Since, the function e^{tx} is convex in x, then

$$e^{tX} \leq \alpha e^{tb} + (1 - \alpha)e^{ta} = \frac{X - a}{b - a}e^{tb} + \frac{b - X}{b - a}e^{ta} \tag{8}$$

Taking the expected value of both sides and using $\mathbb{E}X = 0$ yields

$$\mathbb{E}e^{tX} \leq -\frac{a}{b - a}e^{tb} + \frac{b}{b - a}e^{ta} \stackrel{\text{def}}{=} e^{g(u)}, \tag{9}$$

where $u = t(b - a)$ and $g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u)$ with $\gamma = -a/(b - a)$. Note that $g(0) = g'(0) = 0$ and $g''(u) \leq 1/4$ for all $u > 0$. To see the later, we explicitly compute the second derivative

$$g''(u) = \frac{\gamma(1 - \gamma)e^u}{(1 - \gamma + \gamma e^u)}$$

This function obtains it's maximum when it's derivative is zero (due to convexity), which implies that

$$e^u = \frac{1 - \gamma}{\gamma} \implies g''(u) \leq \frac{(1 - \gamma)^2}{(2 - 2\gamma)^2} = \frac{1}{4}.$$

Then, using Taylor series and Taylor's theorem we obtain that there exists a point $x_\circ \in [0, u]$ such that

$$g(u) = g(0) + g'(0) \cdot u + g''(x_\circ)\frac{u^2}{2} \leq \frac{1}{4} \cdot \frac{t^2(b - a)^2}{2} = \frac{t^2(b - a)^2}{8}. \tag{10}$$

Hence,

$$\mathbb{E}e^{tX} \leq e^{g(u)} \leq e^{\frac{t^2(b-a)^2}{8}}, \tag{11}$$

which leads to the end of the proof. ■

Theorem 2.4. ***Hoeffding's inequality***
Let $Y_1, \ldots, Y_n$ be iid observations with $\mathbb{E}Y_i = \mu$ and $a \leq Y_i \leq b$ a.s. Then, for any $\epsilon > 0$

$$\mathbb{P}(|\overline{Y}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{12}$$

Proof. Note that

$$\mathbb{P}(|\overline{Y}_n - \mu| \geq \epsilon) = \mathbb{P}(\overline{Y}_n - \mu \geq \epsilon) + \mathbb{P}(-\overline{Y}_n + \mu \geq \epsilon)$$

Examine $\mathbb{P}(\overline{Y}_n - \mu \geq \epsilon)$

$$\mathbb{P}(\overline{Y}_n - \mu \geq \epsilon) = \mathbb{P}(Y_1 + \dots Y_n \geq \epsilon n + \mu n) = \mathbb{P}\left(e^t \sum_{i=1}^{n} Y_i \geq e^{tn(\epsilon+\mu)}\right) \leq \tag{13}$$

$$e^{-tn(\epsilon+\mu)} \cdot \mathbb{E}e^{t\sum_{i=1}^{n} Y_i} = e^{-tn(\epsilon+\mu)} \cdot \left(\mathbb{E}e^{tY_i}\right)^n \leq e^{-tn(\epsilon+\mu)} \cdot e^{tn\mu} \cdot e^{\frac{t^2 n(b-a)^2}{8}}. \tag{14}$$

And since the inequality

$$\mathbb{P}(\overline{Y}_n - \mu \geq \epsilon) \leq e^{-tn\epsilon} \cdot e^{\frac{t^2 n(b-a)^2}{8}} \tag{15}$$

is true for any $t \geq 0$, hence we are allowed to take the infimum with respect to t, i.e.

$$\mathbb{P}(\overline{Y}_n - \mu \geq \epsilon) \leq \inf_{t \geq 0} e^{-tn\epsilon} \cdot e^{\frac{t^2 n(b-a)^2}{8}} \tag{16}$$

We see a parabola with positive coefficient of t^2, hence the minimum is obtained at $t^* = \frac{4\epsilon}{(b-a)^2}$ which implies

$$\mathbb{P}(\overline{Y}_n - \mu \geq \epsilon) \leq \inf_{t \geq 0} e^{-tn\epsilon} \cdot e^{\frac{t^2 n(b-a)^2}{8}} = e^{-\frac{2n\epsilon^2}{(b-a)^2}}. \tag{17}$$

Applying the same argument to $\mathbb{P}(-\overline{Y}_n + \mu \geq \epsilon)$ yields the result. ■

Remark 2.1. *First of all, note that here we have two "degrees of freedom": ϵ and n, which means that by varying the tolerance level and sample size one can obtain different results. Suppose we fix ϵ then one can take n that goes to infinity which will lead the rhs vanishes to zero. In case when $n\epsilon^2 \asymp C \implies \epsilon \asymp \frac{1}{\sqrt{n}}$ we get a constant rhs, more precisely, a number proportional to $e^{-2/(b-a)^2}$, which is very typical in probability theory.*

Exercise 2.1. *In the proof of Hoeffding's inequality we have used the fact that Y_is are independent, what if we get rid of this condition? Repeat the proof and comment on your findings.*

A generalization of Hoeffding's inequality is known as McDiarmid inequality.

Theorem 2.5. ***McDiarmid's inequality***
Let $X_1, \dots, X_n$ be independent random variables. Further, let f be a function of $X_1, \dots X_n$ that satisfies $\forall i$,

$$\sup_{x_1,\dots,x_n,x_i'} |f(x_1, \dots x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, \dots x_{i-1}, x_i', x_{i+1}, \dots, x_n)| \leq c_i \tag{18}$$

Then

$$\mathbb{P}\left(f(X_1, \dots, X_n) - \mathbb{E}f(X_1 \dots X_n) \geq \epsilon\right) \leq \exp\left\{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right\} \tag{19}$$

Proof.

■

Another interesting and useful inequality in probability theory is Bernstein inequality. We will describe it in the simplest framework.

Theorem 2.6. ***Bernstein's inequality***
Let $\xi_1, \ldots, \xi_n$ be independent and bounded random variables with zero mean and $|\xi_i| \leq M$ a.s., for all $i \in \{1, \ldots, n\}$ and some constant M. Let $\sigma^2 = \mathbb{V}ar(\overline{\xi_n})$, where $\overline{\xi_n}$ is the sample mean of random variables $\xi_1, \ldots, \xi_n$. Then,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i \geq \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2\sigma^2 + 2/3M\epsilon}} \tag{20}$$

Proof.

■

2.1 Application to High-Dimensional Geometry

2.1.1 Initial observations

We start with a very simple but not intuitive (at least, for people who I asked) fact.

- Consider the map $x \to \lambda x$ and a convex figure D. Applying this map to every point of the figure one can easily conclude that the volume of D changes as follows $|D| \to \lambda^n |D|$, because $x \in \mathbb{R}^n$. Let $\lambda = 1 + \varepsilon, \varepsilon > 0 \implies (1+\varepsilon)^n$ which can be enormously large in case of $n \gg 1$.

  The great example of this is a watermelon in $\mathbb{R}^{1000}$. Assume that we live in 1000-dimensional world and we bought a watermelon with a radius of one meter. As all normal people (remark: they live in $\mathbb{R}^3$) we get out the watermelon rind a thickness of 1 cm. Let's see what we have left

  $$\lambda = \frac{1}{1-10^{-2}} \implies \frac{V}{V_-} = \left[1 + \frac{10^{-2}}{1-10^{-2}}\right]^{1000} \approx 2.3 \times 10^4, \tag{21}$$

  where V is the volume of initial watermelon and V_- is the volume of the watermelon without rind. Thus,

  $$\frac{V - V_-}{V_-} \approx 2.3 \times 10^4, \tag{22}$$

  which mean that the volume of rind is 20000 times as much as the volume of what we have left. In this way, we have obtained the concentration property of n-dimensional ball (defined in the Euclidean space with ℓ_2 norm), i.e. almost everything is near the border.

- Let $F(x) \leq 1$ for $|x| \leq 1$, where $x \in \mathbb{R}^n$, $n \gg 1$, then if we measure the the function value at a random point x such that $|x| \leq 1$, then $|x|(|\cdot|$ can be any norm, if not specified) is approximately be equal to 1, which, in turn, means that the value of $F(x)$ is also close to 1.

- In general, we have:

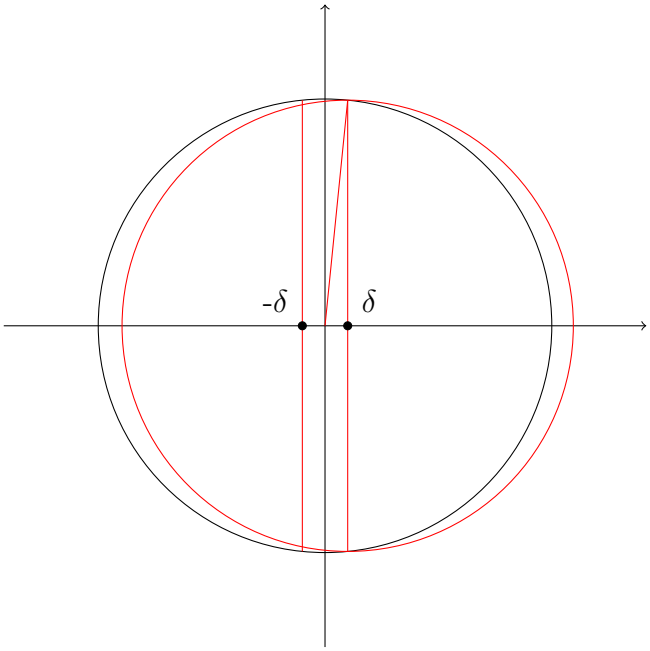
  $$f \in C(\mathbb{B}^n, \mathbb{R}) \text{ и } f(\partial\mathbb{B}^n) = \text{const}, \text{ otherwise} — \text{arbitrary function}, \tag{23}$$

  then from the observer's point of view the function is a constant with great probability, while it may be very picky and far from constant.

2.1.2 Detailed Estimates

- **Volume of cuts**

  Let $\delta \in [0,1]$



Assume we cut a tiny strip near the center of a ball in $\mathbb{R}^n$. The radius of the red circle (see above picture) equals $r = (1-\delta^2)^{\frac{1}{2}}$ and since $1-\delta \leq (1-\delta^2)^{\frac{1}{2}} \implies 2\delta(1-\delta) \leq 0$, which is only true for $\delta \in [0,1]$. So, we construct the cut showed in the picture and calculate the ratio of this tiny strip through δ to the initial ball's volume.

$$R = \frac{\frac{1}{2}(1-\delta^2)^{\frac{n}{2}}}{1^n} = \frac{1}{2}e^{\frac{n}{2}\log(1-\delta^2)} < \frac{1}{2}e^{-\frac{1}{2}\delta^2 n}. \tag{24}$$

For $n \gg 1$ we see that $R \to 0$, which means that for large enough n the cut contains almost no volume, hence we have concentration of measure in a tiny strip around center. Analogically for the left part, i.e. $-\delta$.

Exercise 2.2. *Repeat the same steps (as for volume) for area.*
Hint: Use that the area of hypersphere with raduis R in $\mathbb{R}^n$ equals to nC_nR^{n-1}, where C_n is some constant.

Hence, using the above exercise we get

$$\mathbb{P}_n(x \in \Omega_\delta) > 1 - 2 \cdot e^{-\frac{1}{2}\delta^2 n}, \tag{25}$$

where Ω_δ is the figure (split) around center.

- **Orthogonality of random vectors on $\mathbb{S}^n$**

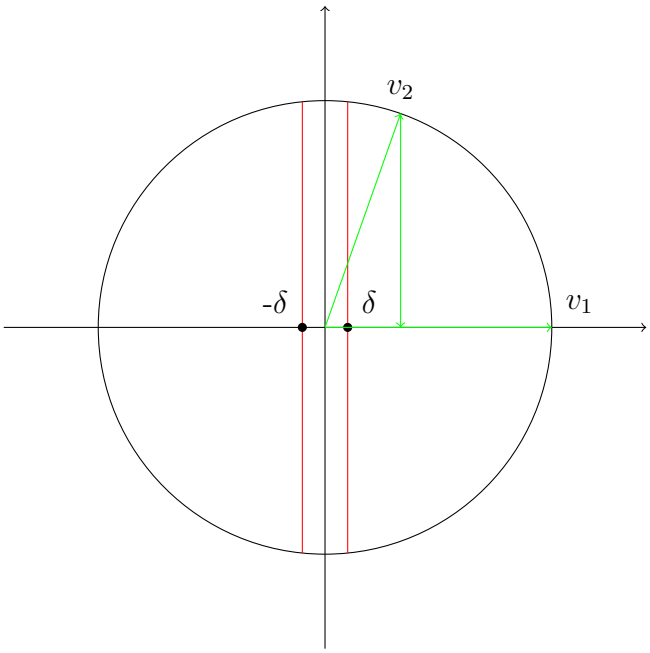
  The claim is the following: random vectors uniformly distributed on a n-dimensional unit ball are almost always orthogonal. We threat "almost always orthogonal" as follows

$$\mathbb{P}_n(|\langle v_1, v_2\rangle| > \delta) \to 0, \tag{26}$$

Alternatively, we write it as $\langle v_1, v_2 \rangle \approx 0$.

Fixing one vector, say v_1 as it was shown in the picture and randomly choose the second vector v_2 on the unit sphere. Let's estimate the probability that the scalar product is more that a given constant δ.

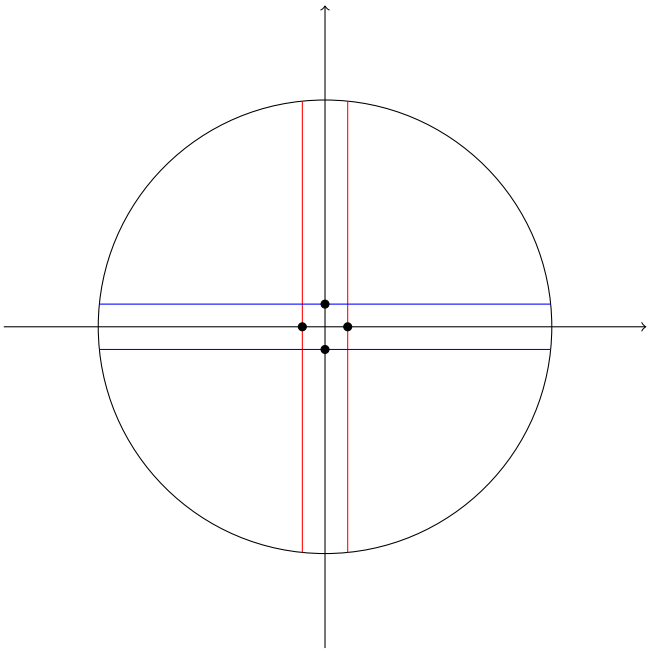
$$\mathbb{P}_n(|\langle v_1, v_2 \rangle| > \delta) < 2 \cdot e^{-\frac{1}{2}\delta^2 n}. \tag{27}$$



Hence, the scalar product of two uniformly distributed random vectors is greater than arbitrary small constant δ with very low probability $\asymp Ce^{-\delta^2 n}$.

Remark 2.2. *The intuition of this phenomenon is the following: the orthogonal space of first vector v_1 has dimension $n-1$, and if a random vector belongs to this space then the scalar product is zero. The rest*

Exercise 2.3. *What if we cut the way it is shown below? Does that mean that almost-all-volume is concentrated in a small square around the sphere center?*



2.1.3 Non-linear Law of Large Numbers

Let $f : S^n \to \mathbb{R}$, then for function f we define the median function M_f as follows

$$|\{x \in S^n \,|\, f(x) \geq M_f\}| \geq \frac{1}{2} \tag{28}$$

$$|\{x \in S^n \,|\, f(x) \leq M_f\}| \geq \frac{1}{2} \tag{29}$$

Assume that $f \in \mathrm{Lip}_1(S^n, \mathbb{R}), n \gg 1$. Note the Lipschitz condition is just an example of regularity condition, there are many other conditions with weaker restrictions on function class. What is important that the function does not have sharp changes, like Dirichlet function.

We claim that for $x_1, x_2 \in S^n$ it holds $f(x_1) \approx f(x_2) \approx M_f$. Then,

$$\Pr\{|f(x) - M_f| > \delta\} < 2 \cdot e^{-\frac{1}{2}\delta^2 n} \tag{30}$$

The above result can be easily extended for a sphere with radius r and $f \in \mathrm{Lip}_L(S^n, \mathbb{R})$

$$\Pr\{|f(x) - M_f| > \delta\} < 2 \cdot e^{-\frac{1}{2}\left(\frac{\delta}{rL}\right)^2 n}. \tag{31}$$

The last inequality is known as non-linear law of large numbers for Lipschitz function f. Recall the standard law of large numbers: let $\xi_1, \xi_2, \ldots \xi_n$ is a sequence of i.i.d. random variables and$\mathbb{E}\xi_i = \mu$, тогда

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \xrightarrow{\mathbb{P}} \mu, \tag{32}$$

or

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu\right| < \varepsilon\right) = 1 \tag{33}$$

Example 2.1. *The temperature in auditorium, from the observer's point of view is constant, while it varies from point to point.*

In probability theory the random variable $S_n = \frac{1}{n}\left(\sum_{i=1}^n x_1 + \dots x_n\right)$ plays crucial role. Consider a sphere $x_1^2 + \dots x_n^2 = \sigma^2 n$. Obviously S_n is Lipschitz function, because $|\nabla S_n| = \frac{1}{\sqrt{n}} = L$ (Lipschitz constant). The radius of sphere $r_n = \sigma\sqrt{n}$, then

$$\Pr\left\{|S_n - 0| > \delta\right\} < 2 \cdot \exp\left(-\frac{1}{2}\left(\frac{\delta}{\sigma}\right)^2 n\right). \tag{34}$$

The intuition behind the last inequality is the following: typical deviations of S_n from 0 observed when $\delta \asymp \frac{1}{\sqrt{n}}$. In probability theory the expression $\frac{1}{\sqrt{n}}$ appears very often, in particular, it is a typical rate of convergence. If δ vanishes with *higher* speed, say, $\frac{1}{n}$ or $\frac{1}{n^2}$, then the right hand side vanishes and we don't observe errors of such size.

2.1.4 Ball in $\mathbb{R}^n$. Central Limit Theorem (CLT)

Recall the Central Limit Theorem (CLT) from section 1. Let $\xi_1, \dots, \xi_n$ i.i.d. with $\mathbb{E}\xi_i = \mu, \mathbb{E}\xi_i^2 < \infty$ and $\mathbb{V}ar\xi = \sigma^2$, then

$$\frac{\xi_1 + \dots + \xi_n - n\mu}{\sigma \cdot \sqrt{n}} \xrightarrow{w} \mathcal{N}(0,1) \tag{35}$$

Consider a sphere with volume of 1, i.e. we want it to a probability measure, then $r_n \asymp \sqrt{n}$. Since,

$$|B^n(r_n)| = \frac{\pi^{n/2}}{\Gamma(n/2+1)} r_n^n = 1, \tag{36}$$

then using the asymptotic expression for Gamma function and Stirling formula one can obtain

$$r_n = \frac{\sqrt[n]{\Gamma\left(\frac{n}{2}+1\right)}}{\sqrt{\pi}} \sim \frac{1}{\sqrt{\pi}}\sqrt{\frac{n}{2e}} \cdot \sqrt[n]{\pi n} \to \sqrt{\frac{n}{2\pi e}}, \text{ as } n \to \infty. \tag{37}$$

Loosely speaking, the sphere with unit volume is huge, in terms of linear sizes, i.e. $r_n \asymp \sqrt{n}$, and volume is 1. Analogically it holds for a unit cube as well, i.e. the volume of cube is 1, but the main diagonal has length of $\sqrt{n}$.

The physical explanation of this fact is as follows: suppose we have n molecules in the auditorium and they are interacting somehow. The kinetic energy then is

$$\frac{1}{2}mv_1^2 + \dots \frac{1}{2}mv_n^2 \asymp \sigma^2 n, \tag{38}$$

where, the last approximation is motivated by the fact kinetic energy is proportional to the number of molecules. Moreover, (38) is a $3n$-dimensional sphere, the radius of which is of order $r \asymp \sigma\sqrt{n}$.

Exercise 2.4. *Project the n-dimensional sphere on a line? Is it something familiar to you?*

2.2 sub-Gaussian class

An important special case of random variables is so called sub-Gaussian random variables. In this section we will present basic properties of sub-Gaussian random variables as well as the structure of the space of sub-Gaussian random variables. They play a crucial role in a

Definition. We call a random variable X with $\mathbb{E}X = \mu$ to be sub-Gaussian if there exists such $\sigma > 0$ that

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\sigma^2}{2}} \tag{39}$$

holds for all $\lambda \in \mathbb{R}$.

Thus, the condition for X to be sub-Gaussian says that there exists $\sigma > 0$ such that the Laplace transform of X is dominated by the Laplace transform of a Gaussian random variable with mean μ and variance σ^2. When $\mu = 0$ the random variable X is usually called σ-sub-Gaussian or sub-Gaussian with parameter σ.

2.3 sub-Exponential class

- definition
- some properties
- the following problem: Prove that

$$\mathbb{P}(|X-\mu| > t) \leq \begin{cases} 2e^{-\frac{t^2}{2\sigma^2}}, t \in [0, \frac{\sigma^2}{b}) \\ 2e^{-\frac{t}{2b}}, t \geq \frac{\sigma^2}{b} \end{cases} \tag{40}$$

3 Bayesian Statistics

We start this section with an interesting and simple exercise.

Example 3.1. *Consider two doctors doing two type of surgeries, and the first doctor has better performance on both surgeries, separately. Is it possible that the second doctor has better performance overall? If yes, bring an example, otherwise prove that it is impossible.*

3.1 Maximum Likelihood

Consider the maximum likelihood function: given an independent and identically distributed (i.i.d.) set of data from a density function f_θ with an unknown parameter $\theta \in \Theta$, the associated likelihood function is

$$\mathcal{L}(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i) \tag{41}$$

The standard parameter estimation technique is to maximize the (log) likelihood with respect to parameter $\theta \in \Theta$. Usually $\Theta = \mathbb{R}$ which leads to , for more complicated cases one can use Lagrange multipliers along with Karush-Kuhn-Tucker conditions to obtain

$$\tilde{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta \,|\, x_1, \ldots, x_n). \tag{42}$$

Note that in some cases it might be convenient to maximize the logarithm of the likelihood function. The estimator obtained this way is known as maximum likelihood estimator (MLE).

Exercise 3.1. *Estimate the probability of getting a tail p of an unfair coin given n experiments. What do you think is bad in this estimate?*

3.2 Bayesian approach

The major input of Bayesian approach, compared to the standard likelihood approach, is that it modifies the likelihood into a *posterior* distribution, taking into account what we believe about the parameter θ *a priori*.

Applying Bayes rule and letting $\mathbf{X} = \{x_1, \ldots, x_n\}$ we can obtain

$$\pi(\theta \,|\, \mathbf{X}) = \frac{p(\mathbf{X} \,|\, \theta)\pi_0(\theta)}{\int p(\mathbf{X} \,|\, \theta)\pi_0(\theta)\, d\theta}, \tag{43}$$

where $\pi_0(\theta)$ is prior distribution, $p(\mathbf{X}, \theta)$ is the likelihood of $\mathbf{X}$ given for θ. $p(\mathbf{X} \,|\, \theta)$ is a joint conditional probability of $\mathbf{X}$ given θ, and it can be decomposed as the product of one-dimensional densities (assuming iid x_i's), i.e.

$$p(\mathbf{X} \,|\, \theta) = \prod_{i=1}^{n} p_\theta(x_i), \tag{44}$$

which is exactly what we have in (41).

Note that the expression (43) results in a distribution function of θ known as posterior distribution of θ given data $\mathbf{X}$.

In order to obtain a point estimate from the distribution function one can maximize the probability density function (pdf) of posterior distribution

$$\hat{\theta} = \arg\max_{\theta} \pi(\theta \mid \mathbf{X}) = \arg\max_{\theta} \left[\overbrace{p(\mathbf{X} \mid \theta)}^{\text{likelihood}} \underbrace{\pi_0(\theta)}_{\text{prior}} \right] \stackrel{\text{def}}{=} \hat{\theta}_{\text{MAP}}, \tag{45}$$

which is called maximum a posteriori (MAP) estimator; or take the expected value of it

$$\hat{\theta} = \mathbb{E}\pi(\theta \mid \mathbf{X}) = \frac{\int \theta p(\mathbf{X} \mid \theta) \pi_0(\theta)\, d\theta}{\int p(\mathbf{X} \mid \theta) \pi_0(\theta)\, d\theta} \tag{46}$$

Remark 3.1. *$\hat{\theta}_{MAP}$ converges to $\hat{\theta}_{MLE}$ when $n \to \infty$.*

Exercise 3.2. *Estimate the probability of getting a tail p of an unfair coin (maximizing the posterior probability) given n experiments using a prior for p, $p \sim \mathcal{B}eta(\alpha, \beta)$ with density*

$$p_{\alpha,\beta}(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \quad B(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}, \tag{47}$$

which implies $p_{\alpha,\beta}(x) \propto x^{\alpha-1}(1-x)^{\beta-1}$. Comment on what you obtain.

3.3 Prior distributions

In this section we introduce an important class of distributions, which is commonly used in Bayesian inference.