Chapter 7. Basic Probability Theory

I-Liang Chern

October 20, 2016

What's kind of matrices satisfying RIP

- Random matrices with
  - iid Gaussian entries
  - iid Bernoulli entries $(+/-1)$
  - iid subgaussian entries
  - random Fourier ensemble
  - random ensemble in bounded orthogonal systems
- In each case, $m = O(s \ln N)$, they satisfy RIP with very high probability $(1 - e^{-Cm})$..

This is a note from S. Foucart and H. Rauhut, A Mathematical Introduction to Compressive Sensing, Springer 2013.

Outline of this Chapter: basic probability

- Basic probability theory
- Moments and tail
- Concentration inequalities

Basic notion of probability theory

- Probability space
- Random variables
- Expectation and variance
- Sum of independent random variables

Probability space

- A probability space is a triple $(\Omega, \mathcal{F}, P)$, Ω: sample space, $\mathcal{F}$: the set of events, P: the probability measure.
- The sample space is the set of all possible outcomes.
- The collection of events $\mathcal{F}$ should be a σ-algebra:
  1. $\emptyset \in \mathcal{F}$;
  2. If $E \in \mathcal{F}$, so is $E^c \in \mathcal{F}$;
  3. $\mathcal{F}$ is closed under countable union, i.e. if $E_i \in \mathcal{F}$, $i = 1, 2, \cdots$, then $\cup_{i=1}^{\infty} E_i \in \mathcal{F}$.
- The probability measure $P : \mathcal{F} \to [0, 1]$ satisfies
  1. $P(\Omega) = 1$;
  2. If E_i are mutually exclusive (i.e. $E_i \cap E_j = \phi$), then $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

Examples

1. Bernoulli trial: a trial which has only two outcomes: success or fail. We represent it as $\Omega = \{1, 0\}$. The collect of events $\mathcal{F} = 2^\Omega = \{\phi, \{0\}, \{1\}, \{0, 1\}\}$. The probability

$$P(\{1\}) = p, \quad P(\{0\}) = 1 - p, \quad 0 \leq p \leq 1.$$

   If we denote the outcome of a Bernoulli trial by x, i.e. $x = 1$ or 0, then $P(x) = p^x(1-p)^{1-x}$.

2. Binomial trials: Let us perform Bernoulli trials n times *independently*. An outcome has the form $(x_1, x_2, ..., x_n)$, where $x_i = 0$ or 1 is the outcome of the ith trial. There are 2^n outcomes. The sample space $\Omega = \{(x_1, ..., x_n) | x_i = 0 \text{ or } 1\}$. The collection of events $\mathcal{F} = 2^\Omega$ is indeed the collection of all subsets of Ω. The probability

$$P(\{(x_1, ..., x_n)\}) := p^{\sum x_i}(1-p)^{n - \sum x_i}.$$

Property of probability measure

1. $P(E^c) = 1 - P(E)$;
2. If $E \subset F$, then $P(E) \leq P(F)$;
3. If $\{E_n, n \geq 1\}$ is either increasing (i.e. $E_n \subset E_{n+1}$) or decreasing to E, then

$$\lim_{n\to\infty} P(E_n) = P(E).$$

Independence and conditional probability

- Let $A, B \in \mathcal{F}$ and $P(B) \neq 0$. The conditional probability $P(A|B)$ (the probability of A given B) is defined to be

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

- Two events A and B are called *independent* if $P(A \cap B) = P(A)P(B)$. In this case $P(A|B) = P(A)$ and $P(B|A) = P(B)$.

Random variables

Outline:

- Discrete random variables
- Continuous random variables
- Expectation and variances

Discrete random variables

- A discrete random variable is a mapping $\mathbf{x} : \Omega \to \{a_1, a_2, \cdots\}$, denoted by $\Omega_{\mathbf{x}}$.
- The random variable $\mathbf{x}$ induces a probability on the discrete set $\Omega_{\mathbf{x}} := \{a_1, a_2, \cdots\}$ with probability $P(\{a_k\}) := P_{\mathbf{x}}(\{\mathbf{x} = a_k\})$ and with σ-algebra $\mathcal{F}_{\mathbf{x}}$ which is $2^{\Omega_{\mathbf{x}}}$, the collection of all subsets of $\Omega_{\mathbf{x}}$. .
- We call the function $a_k \mapsto P_{\mathbf{x}}(a_k)$ the probability mass function of $\mathbf{x}$.
- Once we have $(\Omega_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}}, P_{\mathbf{x}})$, we can just deal with this probability space if we only concern with $\mathbf{x}$, and forget the original probability space $(\Omega, \mathcal{F}, P)$.

Binomial random variable

- Let $\mathbf{x}_k$ be the kth $(k = 1, ..., n)$ outcome of the n independent Bernoulli trials. Clearly, $\mathbf{x}_k$ is a random variable.
- Let $S_n = \sum_{i=1}^{n} \mathbf{x}_i$ be the number of successes in n Bernoulli trials. We see that S_n is also a random variable.
- The sample space that S_n induces is $\Omega_{S_n} = \{0, 1, ..., n\}$.

$$P_{S_n}(S_n = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k (1-p)^{n-k}.$$

- The binomial distribution models the following uncertainty:
  - the number of successes in n independent Bernoulli trials;
  - the relative increase or decrease of a stock in a day;

Poisson random variable

- The Poisson random variable $\mathbf{x}$ takes values $0,\ 1,\ 2, \ldots$ with probability

$$P(\mathbf{x} = k) = e^{-\lambda} \frac{\lambda^k}{k!},$$

  where $\lambda > 0$ is a parameter.
- The sample space is $\Omega = \{0, 1, 2, \ldots\}$. The probability satisfies

$$P(\Omega) = \sum_{k=0}^{\infty} P(\mathbf{x} = k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = 1.$$

- The Poisson random process can be used to model the following uncertainties:
  - the number of customers visiting a specific counter in a day;
  - the number of particles hitting a specific radiation detector in certain period of time;
  - the number of phone calls of a specific phone in a week.
- The parameter λ is different for different cases. It can be estimated by experiments.

Continuous Random Variables

- A continuous random variable is a (Borel) measurable mapping from $\Omega \to \mathbb{R}$. This means that $\mathbf{x}^{-1}([a, b)) \in \mathcal{F}$ for any $[a, b)$.
- It induces a probability space $(\Omega_{\mathbf{x}}, \mathcal{F}_{\mathbf{x}}, P_{\mathbf{x}})$ by
  - $\Omega_{\mathbf{x}} = \mathbf{x}(\Omega)$;
  - $\mathcal{F}_{\mathbf{x}} = \{A \subset \mathbb{R} \,|\, \mathbf{x}^{-1}(A) \in \mathcal{F}\}$
  - $P_{\mathbf{x}}(A) := P(\mathbf{x}^{-1}(A))$.
- In particular, define

$$F_{\mathbf{x}}(x) := P_{\mathbf{x}}((-\infty, x)) := P(\{\mathbf{x} < x\}).$$

called the (cumulative) distribution function. Its derivative $p_{\mathbf{x}}(x)$ w.r.t. dx is called the probability density function:

$$p_{\mathbf{x}}(x) = \frac{dF_{\mathbf{x}}(x)}{dx}.$$

In other word,

$$P(\{a \leq \mathbf{x} < b\}) = F_{\mathbf{x}}(b) - F_{\mathbf{x}}(a) = \int_a^b p_{\mathbf{x}}(x)\, dx.$$

- Thus, $\mathbf{x}$ can be completely characterized by the density function $p_{\mathbf{x}}$ on $\mathbb{R}$.

Gaussian distribution

- The density function of Gaussian distribution is

$$p(x) := \frac{1}{\sqrt{2\pi}\sigma} e^{-|x-\mu|^2/2\sigma^2}, \; -\infty < x < \infty.$$

- A random variable $\mathbf{x}$ with the above probability density function is called a Gaussian random variable and is denoted by

$$\mathbf{x} \sim N(\mu, \sigma).$$

- The Gaussian distribution is used to model:
  - the motion of a big particle (called Brownian particle) in water;
  - the limit of binomial distribution with infinite many trials.

- Exponential distribution. The density is

$$p(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

  The exponential distribution is used to model
  - the length of a telephone call;
  - the length to the next earthquake.
- Laplace distribution The density function is given by

$$p(x) = \frac{\lambda}{2} e^{-\lambda |x|}.$$

  - It is used to model some noise in images.
  - It is used as a prior in Baysian regression (LASSO).

Remarks

- The probability mass function which takes discrete values on $\mathbb{R}$ can be viewed as a special case of probability density function by introducing the notion of delta function.
- The definition of the delta function is

$$\int_{-\infty}^{\infty} \delta(x-a) f(x)\, dx = f(a).$$

- Thus, a discrete random random variable $\mathbf{x}$ with value a_i with probability p_i has the probability density function

$$p(x) = \sum_i p_i \delta(x - a_i).$$

$$\int_a^b p(x)\, dx = \sum_{a<a_i<b} p_i.$$

Expectation

- Given a random variable $\mathbf{x}$ with pdf $p(x)$, we define the expectation of $\mathbf{x}$ by

$$E(\mathbf{x}) = \int xp(x)\,dx.$$

- If f is a continuous function on $\mathbb{R}$, then $f(\mathbf{x})$ is again a random variable. Its expectation is denoted by $E(f(\mathbf{x}))$. We have

$$E(f(\mathbf{x})) = \int f(x)p(x)\,dx.$$

- The kth moment of $\mathbf{x}$ is defined to be:

$$m_k := E(|\mathbf{x}|^k).$$

- In particular, the first and second moments have special names:
    - mean: $\mu := E(\mathbf{x})$
    - variance: $\mathsf{Var}(\mathbf{x}) := E((\mathbf{x} - E(\mathbf{x}))^2)$.

  The variance measures the spread out of values of a random variable.

Examples

1. Bernoulli distribution: mean $\mu = p$, variance $\sigma^2 = p(1-p)$.
2. Binomial distribution S_n: the mean $\mu = np$, variance: $\sigma^2 = np(1-p)$.
3. Poisson distribution: mean $\mu = \lambda$, variance $\sigma^2 = \lambda$.
4. Normal distribution $N(\mu, \sigma)$: mean μ, variance σ^2.
5. Uniform distribution: mean $\mu = (a+b)/2$, variance $\sigma^2 = (b-a)^2/12$.

Joint Probability

- Let $\mathbf{x}$ and $\mathbf{y}$ be two random variables on $(\Omega, \mathcal{F}, P)$. The joint probability distribution of $(\mathbf{x}, \mathbf{y})$ is the measure on $\mathbb{R}^2$ defined by

$$\mu(A) := P((\mathbf{x}, \mathbf{y}) \in A) \text{ for any Borel set } A.$$

The derivative of μ w.r.t. the Lebesgue measure $dx\,dy$ is called the joint probability density:

$$P((\mathbf{x}, \mathbf{y}) \in A) = \int_A p_{(\mathbf{x},\mathbf{y})}(x, y)\,dx\,dy.$$

Independent random variables

- Two random variables $\mathbf{x}$ and $\mathbf{y}$ are called independent if the events $(a < \mathbf{x} < b)$ and $(c < \mathbf{y} < d)$ are independent for any $a < b$ and $c < d$. If $\mathbf{x}$ and $\mathbf{y}$ are independent, then by taking $A = (a,b) \times (c,d)$, we can show that the joint probability

$$\int_{(a,b)\times(c,d)} p_{(\mathbf{x},\mathbf{y})}(x,y)\,dx\,dy = P(a < \mathbf{x} < b \text{ and } c < \mathbf{y} < d)$$
$$= P(a < \mathbf{x} < b)P(c < \mathbf{y} < d) = \left(\int_a^b p_{\mathbf{x}}(x)\,dx\right)\left(\int_c^d p_{\mathbf{y}}(y)\,dy\right)$$

  This yields that

$$p_{(\mathbf{x},\mathbf{y})}(x,y) = p_{\mathbf{x}}(x)\,p_{\mathbf{y}}(y).$$

- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then $E[\mathbf{x}\mathbf{y}] = E[\mathbf{x}]E[\mathbf{y}]$.
- The covariance of $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\text{cov}[\mathbf{x},\mathbf{y}] := E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{y} - E[\mathbf{y}])].$$

  Two random variables are called uncorrelated if their covariance is 0.

Sum of independent random variables

- If $\mathbf{x}$ and $\mathbf{y}$ are independent, then
$$F_{\mathbf{x}+\mathbf{y}}(z) = P(\mathbf{x}+\mathbf{y} < z) = \int_{-\infty}^{\infty} dy \int_{-\infty}^{z-y} dx\, p_{\mathbf{x}}(x) p_{\mathbf{y}}(y)$$
We differentiate this in z and get
$$p_{\mathbf{x}+\mathbf{y}}(z) = \int_{-\infty}^{\infty} dy\, p_{\mathbf{x}}(z-y) p_{\mathbf{y}}(y) := p_{\mathbf{x}} * p_{\mathbf{y}}(z).$$
- Sum of n independent identical distributed (iid) random variables $\{\mathbf{x}_i\}_{i=1}^n$ with mean μ and variance σ^2. Their average is
$$\bar{\mathbf{x}}_n := \frac{1}{n}(\mathbf{x}_1 + \cdots + \mathbf{x}_n)$$
which has mean μ and variance σ^2/n:
$$E[(\bar{\mathbf{x}}_n - \mu)^2] = E\left[\left(\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i - \mu\right)^2\right] = E\left[\left(\frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \mu)\right)^2\right]$$
$$= E\left[\frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n (\mathbf{x}_i - \mu)(\mathbf{x}_j - \mu)\right] = E\left[\frac{1}{n^2}\sum_{i=1}^n (\mathbf{x}_i - \mu)^2\right] = \frac{\sigma^2}{n}$$

Limit of sum of ranviables

- Moments and Tails
- Gaussian, Subgaussian, Subexponential distributions
- Law of large numbers, central limit theorem
- Concentration inequalities

Tails and Moments: Markov inequality

Theorem (Markov's inequality)

Let $\mathbf{x}$ *be a random variable. Then*

$$P(|\mathbf{x}| \geq t) \leq \frac{E[|\mathbf{x}|]}{t} \quad \textit{for all } t > 0.$$

Proof. Note that $P(|\mathbf{x}| \geq t) = E[I_{|\mathbf{x}| \geq t}]$, where

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

is called an indicator function supported on A, which satisfies

$$I_{\{|\mathbf{x}| \geq t\}} \leq \frac{|\mathbf{x}|}{t}.$$

Thus,

$$P(|\mathbf{x}| \geq t) = E[I_{\{|\mathbf{x}| \geq t\}}] \leq \frac{E[|x|]}{t}.$$

Remarks.

- For $p > 0$,

$$P(|\mathbf{x}| \geq t) = P(|\mathbf{x}|^p \geq t^p) \leq \frac{E[|\mathbf{x}|^p]}{t^p}.$$

- For $p = 2$, apply Markov's inequality to $\mathbf{x} - \mu$, we obtain Chebyshev inequality:

$$P(|\mathbf{x} - \mu|^2 \geq t^2) \leq \frac{\sigma^2}{t^2}$$

- For $\theta > 0$,

$$P(\mathbf{x} \geq t) = P(\exp(\theta \mathbf{x}) \geq \exp(\theta t)) \leq \exp(-\theta t) E[\exp(\theta \mathbf{x})].$$

Law of large numbers

Theorem (Law of large numbers)

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ *be i.i.d. with mean* μ *and variance* σ^2*. Then the sample average* $\bar{\mathbf{x}}_n := \frac{1}{n}(\mathbf{x}_1 + \cdots + \mathbf{x}_n)$ *converges in probability to its expected value:*

$$\bar{\mathbf{x}}_n - \mu \xrightarrow{P} 0 \quad \text{as} \quad n \to \infty.$$

That is, for any $\epsilon > 0$*,*

$$\lim_{n \to \infty} P(|\bar{\mathbf{x}}_n - \mu| \geq \epsilon) = 0.$$

Proof.

1. Using iid of $\mathbf{x}_j$, the mean and variance of $\bar{\mathbf{x}}_n$ are: $E[\bar{\mathbf{x}}_n] = \mu$, while

$$\sigma^2(\bar{\mathbf{x}}_n) = E[(\bar{\mathbf{x}}_n - \mu)^2] = E\left[\frac{1}{n^2}\left(\sum_{j=1}^{n}(\mathbf{x}_j - \mu)\right)^2\right]$$

$$= \frac{1}{n^2}\sum_{j,k=1}^{n} E[(\mathbf{x}_j - \mu)(\mathbf{x}_k - \mu)] = \frac{1}{n^2}\sum_{j=1}^{n} E[(\mathbf{x}_j - \mu)^2] = \frac{1}{n}\sigma^2$$

2. We apply the Chebyshev's inequality to $\bar{\mathbf{x}}_n$:

$$P(|\bar{\mathbf{x}}_n - \mu| \geq \epsilon) = P(|\bar{\mathbf{x}}_n - \mu|^2 \geq \epsilon^2) \leq \frac{\sigma(\bar{\mathbf{x}}_n)^2}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0 \text{ as } n \to \infty.$$

Remarks.

- The condition on variance can be removed. But we use Chebyshev inequality to prove this theorem, which uses the assumption of finite variance.
- No convergent rate here. The concentration inequality provides rate estimate, which needs tail control.

Tail probability can be controlled by moments

Lemma (Markov)

For $p > 0$,

$$P(|\mathbf{x}| \geq t) = P(|\mathbf{x}|^p \geq t^p) \leq \frac{E[|\mathbf{x}|^p]}{t^p}.$$

Proposition

If $\mathbf{x}$ *is a random variable satisfying*

$$E[|\mathbf{x}|^p] \leq \alpha^p \beta p^{p/2} \quad \text{for all } p \geq 2,$$

then

$$P(|\mathbf{x}| \geq e^{1/2}\alpha u) \leq \beta e^{-u^2/2} \quad \text{for all } u \geq \sqrt{2}.$$

1. Use Markov inequality
$$P(|\mathbf{x}| \geq \sqrt{e}\alpha u) \leq \frac{E[|\mathbf{x}|^p]}{(\sqrt{e}\alpha u)^p} \leq \beta \left(\frac{\alpha\sqrt{p}}{\sqrt{e}\alpha u}\right)^p.$$
2. Choosing $p = u^2$, we get the estimate.

Moments can be controlled by tail probability

Proposition

The moments of a random variable $\mathbf{x}$ *can be expressed as*

$$E[|\mathbf{x}|^p] = p\int_0^\infty P(|\mathbf{x}| \geq t)t^{p-1}\,dt, \quad p > 0.$$

Proof.

1. Use Fubini theorem:

$$\begin{aligned} E[|\mathbf{x}|^p] &= \int_\Omega |\mathbf{x}|^p \, dP = \int_\Omega \int_0^{|\mathbf{x}|^p} 1 \, dx \, dP = \int_\Omega \int_0^\infty I_{\{|\mathbf{x}|^p \geq x\}} \, dx \, dP \\ &= \int_0^\infty \int_\Omega I_{\{|\mathbf{x}|^p \geq x\}} \, dP \, dx = \int_0^\infty P(|\mathbf{x}|^p \geq x) \, dx \\ &= p\int_0^\infty P(|\mathbf{x}|^p \geq t^p)t^{p-1} \, dt = p\int_0^\infty P(|\mathbf{x}| \geq t)t^{p-1} \, dt \end{aligned}$$

2. Here $I_{\{|\mathbf{x}|^p \geq x\}}$ is a random variable which is 1 as $|\mathbf{x}|^p \geq x$ and 0 otherwise.

Moments can be controlled by tail probability

Proposition

Suppose $\mathbf{x}$ *is a random variable satisfying*

$$P(|\mathbf{x}| \geq e^{1/2}\alpha u) \leq \beta e^{-u^2/2} \text{ for all } u > 0,$$

then for all $p > 0$,

$$E[|\mathbf{x}|^p] \leq \beta\alpha^p(2e)^{p/2}\Gamma\left(\frac{p}{2}+1\right).$$

Proposition

If $P(|\mathbf{x}| \geq t) \leq \beta e^{-\kappa t^2}$ *for all* $t > 0$, *then*

$$E[|\mathbf{x}|^n] \leq \frac{n\beta}{2}\kappa^{-n/2}\Gamma\left(\frac{n}{2}\right).$$

Subgaussian and subexponential distributions

Definition

A random variable $\mathbf{x}$ is called subgaussian if there exist constants $\beta, \kappa > 0$ such that

$$P(|\mathbf{x}| \geq t) \leq \beta e^{-\kappa t^2} \quad \text{for all } t > 0;$$

It is called subexponential if there exist constants $\beta, \kappa > 0$ such that

$$P(|\mathbf{x}| \geq t) \leq \beta e^{-\kappa t} \quad \text{for all } t > 0.$$

Notice that $\mathbf{x}$ is subgaussian if and only if $\mathbf{x}^2$ is subexponential.

Subgaussian

Proposition

A random variable is subgaussian if and only if $\exists\, c, C > 0$ *such that*

$$E[\exp(c\mathbf{x}^2)] \le C.$$

Proof. ($\Rightarrow$)

1. Estimating moments from tail, we get $E[\mathbf{x}^{2n}] \le \beta\kappa^{-n} n!$.
2. Expand exponential function

$$E[\exp(c\mathbf{x}^2)] = 1 + \sum_{n=1}^{\infty} \frac{c^n E[\mathbf{x}^{2n}]}{n!} \le 1 + \beta \sum_{n=1}^{\infty} \frac{c^n \kappa^{-n} n!}{n!} \le C.$$

($\Leftarrow$) From Markov inequality

$$P(|\mathbf{x}| \ge t) = P(\exp(c\mathbf{x}^2) \ge e^{ct^2}) \le E[\exp(c\mathbf{x}^2)]e^{-ct^2} \le Ce^{-ct^2}.$$

Subgaussian with mean 0

Proposition

A random variable $\mathbf{x}$ *is subgaussian with* $E\mathbf{x} = 0$ *if and only if* $\exists\ c > 0$ *such that* $E[\exp(\theta \mathbf{x})] \leq \exp(c\theta^2)$ *for all* $\theta \in \mathbb{R}$.

$(\Leftarrow)$

1. Apply Markov inequality

$$P(\mathbf{x} \geq t) = P(\exp(\theta \mathbf{x}) \geq \exp(\theta t)) \leq E[\exp(\theta \mathbf{x})] e^{-\theta t} \leq e^{c\theta^2 - \theta t}.$$

   Optimal θ yields $P(\mathbf{x} \geq t) \leq e^{-t^2/(4c)}$.

2. Repeating this for $-\mathbf{x}$, we also get $P(-\mathbf{x} \geq t) \leq e^{-t^2/(4c)}$. Thus,

$$P(|\mathbf{x}| \geq t) = P(\mathbf{x} \geq t) + P(-\mathbf{x} \geq t) \leq 2e^{-t^2/(4c)}.$$

3. To show $E[\mathbf{x}] = 0$, we use

$$1 + \theta E[\mathbf{x}] \leq E[\exp(\theta \mathbf{x})] \leq e^{c\theta^2}$$

   Take $\theta \to 0$, we obtain $E[\mathbf{x}] = 0$.

$(\Rightarrow)$

1. It is enough to prove the statement for $\theta \geq 0$. For $\theta < 0$, we replace $\mathbf{x}$ by $-\mathbf{x}$.
2. For $\theta < \theta_0$ small, expand $\exp$, use $E[\mathbf{x}] = 0$, moment estimate via tail and Stirling formula:

$$E[\exp(\theta \mathbf{x})] = 1 + \sum_{n=2}^{\infty} \frac{\theta^n E[\mathbf{x}^n]}{n!} \leq 1 + \beta \sum_{n=2}^{\infty} \frac{\theta^n C^n \kappa^{-n/2} n^{n/2}}{n!}$$

$$\leq 1 + \theta^2 \frac{\beta (Ce)^2}{\sqrt{2\pi}\kappa} \sum_{n=0}^{\infty} \left(\frac{Ce\theta_0}{\sqrt{\kappa}} \right)^n \leq 1 + \theta^2 \frac{\beta (Ce)^2}{\sqrt{2\pi}\kappa} \frac{1}{1 - \frac{Ce\theta_0}{\sqrt{\kappa}}} = 1 + c_1 \theta^2 \leq \exp(c_1 \theta^2).$$

   Here, $\theta_0 = \sqrt{\kappa}/(2Ce)$ and satisfies $Ce\theta_0 \kappa^{-1/2} < 1$.
3. For $\theta > \theta_0$, we aim at proving $E[\exp(\theta \mathbf{x} - c_2 \theta^2)] \leq 1$. Here, $c_2 = 1/(4c)$.

$$E[\exp(\theta \mathbf{x} - c_2 \theta^2)] = E[\exp(-q^2 + \frac{\mathbf{x}^2}{4c_2})] \leq E[\exp(\frac{\mathbf{x}^2}{4c_2})] \leq C.$$

4. Define $\rho = \ln(C)\theta_0^{-2}$ yields

$$E[\exp(\theta \mathbf{x})] \leq C e^{c_2 \theta^2} = C e^{(-\rho + (\rho + c_2))\theta^2} \leq C e^{-\rho \theta_0^2} e^{(\rho + c_2)\theta^2} \leq e^{(\rho + c_2)\theta^2}$$

   Setting $c_3 = \max(c_1, c_2 + \rho)$. This completes the proof.

Bounded random variable

Corollary

If a random variable $\mathbf{x}$ *has mean* 0 *and* $|\mathbf{x}| \leq B$ *almost surely, then*

$$E[\exp(\theta \mathbf{x})] \leq \exp(B^2 \theta^2 / 2)$$

Proof.

1. We write $\mathbf{x} = (-B)t + (1-t)B$, where $t = (B - \mathbf{x})/2B$ is a random variable, $0 \leq t \leq 1$ and $E[t] = 1/2$..
2. By Jensen inequality: $e^{\theta \mathbf{x}} \leq t e^{-B\theta} + (1-t) e^{B\theta}$, taking expectation,

$$E[\exp(\theta \mathbf{x})] \leq \frac{1}{2} e^{-B\theta} + \frac{1}{2} e^{B\theta} = \sum_{k=0}^{\infty} \frac{(\theta B)^2 n}{(2n)!} \leq \sum_{k=0}^{\infty} \frac{(\theta B)^2 n}{(2^n n!} = \exp(B^2 \theta^2 / 2).$$

Exponential decay tails, cumulant function $\ln E[\exp(\theta \mathbf{x})]$

In the case when the tail decays fast, the corresponding moment information can be grouped into $\exp[\theta \mathbf{x}]$. The function $C_{\mathbf{x}}(\theta) := \ln E[\exp(\theta \mathbf{x})]$ is called the cumulant function of $\mathbf{x}$.

Example of $C_{\mathbf{x}}$

Let $g \sim N(0,1)$. Then

$$E[\exp(ag^2 + \theta g)] = \frac{1}{\sqrt{1-2a}} \exp\left(\frac{\theta^2}{2(1-2a)}\right), \text{ for } a < 1/2, \theta \in \mathbb{R}.$$

This is from

$$E[\exp(ag^2 + \theta g)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(ax^2 + \theta x) \exp(-x^2/2)\, dx$$

In particular,

$$E[\exp(\theta g)] = \exp\left(\frac{\theta^2}{2}\right).$$

On the other hand,

$$E[\exp(\theta g)] = \sum_{j=0}^{\infty} \frac{\theta^j E[g^j]}{j!} = \sum_{n=0}^{\infty} \frac{\theta^{2n} E[g^{2n}]}{(2n)!}$$

By comparing the two expansions for $E[\exp(\theta \mathbf{x})]$, we obtain

$$E[g^{2n+1}] = 0, \quad E[g^{2n}] = \frac{(2n)!}{2^n n!}, \quad n = 0, 1, \ldots$$

$$C_g(\theta) := \ln E[\exp(\theta g)] = \frac{\theta^2}{2}.$$

Examples of $C_{\mathbf{x}}$

Let the random variable $\mathbf{x}$ have the pdf $\chi_{[-B,B]}/(2B)$. Then

$$E[\exp(\theta\mathbf{x})] = \frac{1}{2B}\int_{-B}^{B} \exp(\theta x)\,dx = \frac{e^{B\theta} - e^{-B\theta}}{2B\theta} = \sum_{n=0}^{\infty} \frac{(B\theta)^{2n}}{(2n+1)!}.$$

On the other hand, $E[\exp(\theta\mathbf{x})] = \sum_{k=0}^{\infty} \frac{\theta^k E[\mathbf{x}^k]}{k!}$. By comparing these two expansions, we obtain

$$E[\mathbf{x}^{2n+1}] = 0, \quad E[\mathbf{x}^{2n}] = \frac{B^{2n}}{2n+1}, \quad n = 0, 1, \ldots$$

$$C_{\mathbf{x}}(\theta) = \ln\left(e^{B\theta} - e^{-B\theta}\right) - \ln\theta + C.$$

Examples of C_{x}

The pdf of Rademacher distribution is $p_\epsilon(x) = (\delta(x+1) + \delta(x-1))/2$.

$$E[\exp(\theta\epsilon)] = \frac{1}{2}\int e^{\theta x}(\delta(x+1) + \delta(x-1))\,dx = \frac{e^\theta + e^{-\theta}}{2} = \sum_{n=0}^{\infty} \frac{\theta^{2n}}{(2n)!}.$$

Thus, we get

$$E[\epsilon^{2n+1}] = 0, \quad E[\epsilon^{2n}] = 1, \quad n = 1, 2, \ldots$$

$$C_\epsilon(\theta) = \ln(e^\theta + e^{-\theta}) + C.$$

Concentration inequalities

Motivations: In the law of large numbers, if the distribution function satisfies certain growth condition, e.g. decay exponentially fast, or even finite support, then we have sharp estimate how fast $\bar{\mathbf{x}}_n$ tends to μ. This is the large deviation theory below. The rate is controlled by the cumulant function $C_{\mathbf{x}}(\theta) := \ln E[\exp(\theta \mathbf{x})]$.

Theorem (Cremér's theorem)

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ be independent random variables with cumulant-generating function $C_{\mathbf{x}_\ell}$. Then for $t > 0$,

$$P\left(\frac{1}{n}\sum_{\ell=1}^{n} \mathbf{x}_\ell \geq x\right) \leq \exp(-nI(x)),$$

$$I(x) := \sup_{\theta > 0}\left[\theta x - \frac{1}{n}\sum_{\ell=1}^{n} C_{\mathbf{x}_\ell}(\theta)\right].$$

Proof.

1. By Markov's inequality and independence of $\mathbf{x}_\ell$,

$$
\begin{aligned}
P(\bar{\mathbf{x}}_n \geq x) &= P(\exp(\theta \bar{\mathbf{x}}_n) \geq \exp(\theta x)) \leq e^{-\theta x} E[\exp(\theta \bar{\mathbf{x}}_n)] = e^{-\theta x} E\left[\exp\left(\sum_{\ell=1}^n \frac{\theta \mathbf{x}_\ell}{n}\right)\right] \\
&= e^{-\theta x} E\left[\prod_{\ell=1}^n \exp\left(\frac{\theta \mathbf{x}_\ell}{n}\right)\right] = e^{-\theta x} \prod_{\ell=1}^n E\left[\exp\left(\frac{\theta \mathbf{x}_\ell}{n}\right)\right] \\
&= e^{-\theta x} \exp\left(\sum_{\ell=1}^n \ln E\left[\exp\left(\frac{\theta \mathbf{x}_\ell}{n}\right)\right]\right) = \exp\left(-\theta x + \sum_{\ell=1}^n C_{\mathbf{x}_\ell}\left(\frac{\theta}{n}\right)\right) \\
&= \exp\left(-n\left(\theta' x - \frac{1}{n}\sum_{\ell=1}^n C_{\mathbf{x}_\ell}(\theta')\right)\right) \leq \exp\left(-nI(x)\right)
\end{aligned}
$$

Here,

$$
I(x) = \sup_{\theta > 0} \left[\theta x - \frac{1}{n} \sum_{\ell=1}^n C_{\mathbf{x}_\ell}(\theta)\right].
$$

Remark. If each $C_{\mathbf{x}_\ell}$ is subgaussian with 0 mean, then $C_{\mathbf{x}_\ell}(\theta) \leq c\theta^2$. This leads to $I(x) \geq x^2/4c$

Hoeffding inequality

Theorem (Hoeffding inequality)

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ *be independent random variables with mean* 0 *and* $\mathbf{x}_\ell \in [-B, B]$ *almost surely for* $\ell = 1, \ldots, n$. *Then*

$$P(\bar{\mathbf{x}}_n > x) \le \exp(-nI(x)).$$

Here,

$$I(x) := \frac{x^2}{\frac{1}{2n}\sum_{i=1}^{n}(2B)^2} = \frac{x^2}{2B^2}$$

Proof.

1. Let $S_n = \sum_{\ell=1}^{n} \mathbf{x}_\ell$. Use Markov's inequality

$$
\begin{aligned}
P(S_n > t) \le e^{-t\theta} E[\exp(\theta S_n)] = e^{-t\theta} E[\prod_{\ell=1}^{n} \exp(\theta \mathbf{x}_\ell)] = e^{-t\theta} \prod_{\ell=1}^{n} E[\exp(\theta \mathbf{x}_\ell)] \\
\le e^{-t\theta} \prod_{\ell=1}^{n} \exp\left(\frac{B^2}{2}\theta^2\right) = \exp\left(-t\theta + \sum_{\ell=1}^{n}\left(\frac{B^2}{2}\theta^2\right)\right).
\end{aligned}
$$

2. Let write $t = nx$, taking convex conjugate:

$$
I(x) := \sup_{\theta}\left(x\theta - \frac{1}{2}B^2\theta^2\right) = \frac{x^2}{2B^2}.
$$

we get

$$
P(S_n > nx) \le \exp(-n(x\theta - \frac{B^2}{2}\theta^2)) \le \exp(-nI(x)).
$$

Bernstein's inequality

Theorem (Bernstein's inequality)

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ *be independent random variables with mean* 0 *and variance* σ_ℓ^2*, and* $\mathbf{x}_\ell \in [-B, B]$ *almost surely for* $\ell = 1, ..., n$*. Then*

$$P(\sum_{\ell=1}^{n} \mathbf{x}_\ell > t) \leq \exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right),$$

$$P(|\sum_{\ell=1}^{n} \mathbf{x}_\ell| > t) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Bt/3}\right),$$

where $\sigma^2 = \sum_{\ell=1}^{n} \sigma_\ell^2$*.*

Remark.

In Hoeffding's inequality, we do not use the variance information. The concentration estimation is

$$P(\bar{\mathbf{x}}_n > \epsilon) \leq \exp\left(-n\frac{\epsilon^2}{2B^2}\right).$$

In Bernstein inequality, we use the variance information. Let $\bar{\sigma}^2 := \frac{1}{n}\sum_{\ell=1}^{n}\sigma_\ell^2$. The Bernstein inequality reads

$$P(\bar{\mathbf{x}}_n > \epsilon) \leq \exp\left(-n\frac{\epsilon^2}{\bar{\sigma}^2 + \frac{B\epsilon}{3}}\right)$$

Comparing the denominators, Bernstein's inequality is sharper, provided $\bar{\sigma} < B$.

Proof.

1. For a random variable $\mathbf{x}_\ell$ which has mean 0, variance σ_ℓ^2 and $|\mathbf{x}| \leq B$ almost surely, its moment generating function $E[\exp(\theta \mathbf{x}_\ell)]$ satisfies

$$E[\exp(\theta \mathbf{x}_\ell)] = E\left[\sum_{k=0}^{\infty} \frac{\theta^k \mathbf{x}_\ell^k}{k!}\right] \leq E\left[1 + \sum_{k=2}^{\infty} \frac{\theta^k |\mathbf{x}_\ell|^2 B^{k-2}}{k!}\right]$$
$$\leq 1 + \frac{\theta^2 \sigma_\ell^2}{2} \sum_{k=2}^{\infty} \frac{2(\theta B)^{k-2}}{k!} = 1 + \frac{\theta^2 \sigma_\ell^2}{2} F_\ell(\theta) \leq \exp(\theta^2 \sigma_\ell^2 F_\ell(\theta)/2).$$

   Here,

$$F_\ell(\theta) = \sum_{k=2}^{\infty} \frac{2(\theta B)^{k-2}}{k!} \leq \sum_{k=2}^{\infty} \frac{(\theta B)^{k-2}}{3^{k-2}} = \frac{1}{1 - B\theta/3} := \frac{1}{1 - R\theta}$$

   where $R := B/3$. We require $0 \leq \theta < 1/R$.

2. Let $S_n = \sum_{\ell=1}^n \mathbf{x}_\ell$. Using Cramer theorem,

$$P(S_n > t) \leq e^{-t\theta} \prod_{\ell=1}^{n} E[\exp(\theta \mathbf{x}_\ell)] \leq e^{-\theta t} \exp\left(\sum_{\ell=1}^{n} \theta^2 \sigma_\ell^2 F_\ell(\theta)/2\right)$$
$$\leq \exp\left(-\theta t + \frac{\sigma^2}{2} \frac{\theta^2}{1 - R\theta}\right)$$

   Here, $\sigma^2 = \sum_{\ell=1}^n \sigma_\ell$.

Proof (Cont.)

3. Choose $\theta = t/(\sigma^2 + Rt)$, which satisfies $\theta < 1/R$. We then get

$$P(S_n > t) = \exp\left(\frac{t^2\sigma^2}{2(\sigma^2+Rt)^2}\frac{1}{1-\frac{Rt}{\sigma^2+Rt}} - \frac{t^2}{\sigma^2+Rt}\right) \leq \exp\left(-\frac{t^2}{2(\sigma^2+Rt)}\right). \tag{0.1}$$

4. Replacing $\mathbf{x}_\ell$ by $-\mathbf{x}_\ell$ yields the same estimate. We then get the estimate for $P(|S_n| > t)$.

Bernstein inequality for subexponential r.v.

We can extend Bernstein's inequality to random variables without bound, but decay exponentially fast. That is, those subexponential random variables.

Corollary

Let $\mathbf{x}_1, ..., \mathbf{x}_n$ *be independent mean* 0 *subexponential random variables, i.e.* $P(|\mathbf{x}_\ell| \geq t) \leq \beta e^{-\kappa t}$ *for some constant* $\beta, \kappa > 0$ *for all* $t > 0$*. Then*

$$P(|\sum_{\ell=1}^{n} \mathbf{x}_\ell| \geq t) \leq 2 \exp\left(-\frac{(\kappa t)^2/2}{2\beta n + \kappa t}\right).$$

Proof.

1. For subexponential random variable $\mathbf{x}$, for $k \geq 2$,

$$E[|\mathbf{x}|^k] = k\int_0^\infty P(|\mathbf{x}| \geq t)t^{k-1}\,dt \leq \beta k\int_0^\infty e^{-\kappa t}t^{k-1}\,dt$$
$$= \beta k\kappa^{-k}\int_0^\infty e^{-u}u^{k-1}\,du = \beta\kappa^{-k}k!.$$

2. Using this estimate and $E[\mathbf{x}_\ell] = 0$, we get

$$E[\exp(\theta\mathbf{x}_\ell)] = E\left[\sum_{k=0}^\infty \frac{\theta^k\mathbf{x}_\ell^k}{k!}\right] \leq 1 + \sum_{k=2}^\infty \frac{\theta^k\beta\kappa^{-k}k!}{k!}$$
$$= 1 + \beta\frac{\theta^2\kappa^{-2}}{1-\theta\kappa^{-1}} \leq \exp\left(\beta\frac{\theta^2\kappa^{-2}}{1-\theta\kappa^{-1}}\right).$$

3. Using Cramer's inequality, we have

$$P(S_n \geq t) \leq \exp\left(-\theta t + \frac{n\beta}{\kappa^2}\frac{\theta^2}{1-\kappa^{-1}\theta}\right)$$

Comparing this formula and (0.1) with $R = 1/\kappa$, $\sigma^2 = 2n\beta\kappa^{-2}$, we get

$$P(S_n \geq t) \leq \exp\left(-\frac{t^2}{2(\sigma^2+Rt)}\right) = \exp\left(-\frac{t^2}{2(2n\beta\kappa^{-2}+\kappa^{-1}t)}\right)$$
$$= \exp\left(-\frac{(\kappa t)^2/2}{2n\beta+\kappa t}\right).$$