Advanced mathematical statistics August 30, 2025

Part I. Review of Basic Probabilities

Instructor: Lin Liu

1 Plan of this semester

Our plan is to (hopefully) cover the following topics in theoretical statistics:

- Classical parametric statistics [**? ?**]: including sufficiency/ancillary statistics, exponential families, moment estimators, maximum likelihood estimation (MLE), M/Z-estimation, Le Cam's theory, Bernstein-von Mises theory for Bayesian inference, etc. Here the parameter space Θ is of finite dimensional compared to the sample size n.

- Nonparametric statistics [**? ?**]: minimax lower and upper bound for nonparametric statistical problems, including both function and functional estimation. By nonparametric statistics, we mean that the parameter space Θ is infinite dimensional. e.g. $\Theta = L^2$. Nonparametric statistics are motivated to avoid model misspecification bias in classical parametric statistics. Here we will also cover some high-dimensional statistics with sparsity constraint, as sparsity constraint can be viewed as restricting the L_0 norms of the Fourier coefficients of the Fourier expansion of a function in L^2.

- Semiparametric statistics [**? ?**]: semiparametric efficiency and influence functions. By semiparametric statistics, we mean that the parameter θ has two components (μ, ν), where μ belongs to a finite-dimensional space and ν, the nuisance parameter, belongs to an infinite-dimensional space. The parameter of interest is μ. We will use a lot of functional analysis in this chapter.

- Some other topics: Bayesian nonparametrics [**?**], robust statistics, statistical-computational gap (if time permitted). For statistical-computational gap, we mean that the known statistically optimal procedure might belong to computational complexity class like Exp or NP-complete etc., but the known polynomial time procedure cannot achieve the statistical optimality after years of effort. Such gap might be intrinsic and strong evidence can be shown by the so-called sum-of-square (SoS) or Lasserre relaxation hierarchy [**? ?**], a fascinated topic in the intersection between mathematical statistics, theoretical computer science and real algebraic geometry.

The above plan is for last year. I might change the above plan as we go along.

2 Review of some basic probabilistic facts

Note 1. I choose not to develop expectations in a more measure-theoretically rigorous way in this class due to time limit. Several very good references (at least to me) in this regard are David Pollard's "A user's guide to measure theoretical probability" (as a statistician), Patrick Billingsley's "Probability and Measure" (as a mathematician) or Amir Dembo's lecture notes (as an applied probabilist). I highly recommend Chapter 1 and Chapter 4 of Dembo's notes for self-studying because he develops everything in a fast pace.

2.1 Expectation, variance, covariance, and moment

X is a random variable and $X \sim P_X$, where P_X denotes the probability distribution function/probability measure. In undergraduate statistics courses, we usually define expectations/moments of X through probability density/mass function (provided that it exists) by

$$\mathbb{E}(X) := \int x f_X(x) dx.$$

If $f_X(x)$ is a p.d.f., this integral is taken with respect to (w.r.t.) the Lebesgue measure, whereas if $f_X(x)$ is a p.m.f., this integral is taken with respect to the counting measure of a discrete probability distribution.

But a more general way of defining expectation is the following

$$\mathbb{E}(X) := \int x dF_X(x)$$

where F_X is the c.d.f. because the c.d.f. always exists, unlike p.d.f. or p.m.f. Or if you really want to be measure-theoretically rigorous, expectation can be defined as

$$\mathbb{E}(X) := \int_\Omega X(\omega) P_X(d\omega)$$

where the integral is taken w.r.t. the probability measure P_X over the sample space Ω. However, to really understand the meaning of the above definition, a measure-theory based probability course is needed.

Note 2. Another common identity is the following (in case you haven't seen it): denote the c.d.f. of X as $F_X(x) \equiv P(X \leq x)$. We can define c.d.f. using expectation by introducing the indicator function

$$\mathbb{1}\{X \in A\} := \begin{cases} 1 & X \in A \\ 0 & \text{Otherwise.} \end{cases}$$

Then $F_X(x) \equiv \mathbb{E}\left[\mathbb{1}\{X \leq x\}\right]$.

In mathematical analysis, $\mathbb{1}\{X \in A\}$ is called characteristic function [**?**]. But because characteristic function has another meaning in probability, we rename it as the indicator function.

For instance, in measure-theoretic probability, we often start with a probability space $(\Omega, \mathcal{F}, P)$ with Ω the sample space, $\mathcal{F}$ a σ-algebra on Ω, and P a probability measure. The reason that we need a σ-algebra rather than an algebra is we need to operate on union/intersection of countably many rather than finitely many sets in probability theory. We can subsequently define a random variable $X \equiv X(\omega)$ as a measurable mapping $X : \Omega \to \mathbb{X}$ from the sample space Ω to the random variable space $\mathbb{X}$. Again, the random variable space should also be a measurable space, equipped with a σ-algebra $\mathcal{X}$ and written as a tuple $(\mathbb{X}, \mathcal{X})$. The most common random variable space will be $(\mathbb{R}, \mathcal{B})$, where $\mathbb{R}$ denotes the field of real numbers and $\mathcal{B}$ the Borel σ-algebra, i.e. σ-algebras generated from open sets in $\mathbb{X}$, which are open intervals when $\mathbb{X} = \mathbb{R}$.

Variance, covariance and moment can be similarly defined based on expectation. We have the following:

Definition 3. An r-th moment of a r.v. $X \sim P_X$ is $\mathbb{E}\left(X^r\right)$, if it exists. Variance is the central second moment: $\mathsf{var}\left(X\right) := \mathbb{E}\{(X - \mathbb{E}(X))^2\}$. Covariance between two r.v.s X_1, X_2 is defined as $\mathsf{cov}\left(X_1, X_2\right) := \mathbb{E}\{(X_1 - \mathbb{E}(X_1))\,(X_2 - \mathbb{E}(X_2))\}$. When $X_1 = X_2 = X$ almost surely (with probability 1 [w.p.1]), $\mathsf{cov}(X_1, X_2) \equiv \mathsf{var}\left(X\right)$.

For variance and covariance, we also have the following useful identities:

Fact 1.

$$\mathsf{var}\left(X\right) \equiv \mathbb{E}\left(X^2\right) - \{\mathbb{E}\left(X\right)\}^2,$$
$$\mathsf{cov}\left(X_1, X_2\right) \equiv \mathbb{E}\left(X_1 X_2\right) - \mathbb{E}\left(X_1\right)\mathbb{E}\left(X_2\right).$$

2.1.1 Expectation is a linear operator/functional

Expectation is a linear operator/functional: for two random variables $X_1 \sim P_{X_1}, X_2 \sim P_{X_2}$, $\mathbb{E}(aX_1 + bX_2) = a\mathbb{E}(X_1) + b\mathbb{E}(X_2)$ for $a, b \in \mathbb{R}$ (here the left hand side [LHS] expectation is to be understood as taken w.r.t. the joint distribution of X_1, X_2). In this class, a "functional" simply means a measurable map $\mathcal{L} : \mathscr{F} \to \mathbb{R}^m$, where $\mathscr{F}$ is some normed function space, e.g. a Hilbert space like L^2. A linear functional further requires $\mathcal{L}(f_1 + f_2) = \mathcal{L}(f_1) + \mathcal{L}(f_2)$ and $\mathcal{L}(cf) = c\mathcal{L}(f)$, where $c \in \mathbb{R}$, $f, f_1, f_2 \in \mathscr{F}$.

Note 4. One question you can think about now is whether

$$\mathbb{E}\left(\sum_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} \mathbb{E}\left(X_i\right)?$$

We will come back to this later in this course. Unlike the measure theoretical aspects that I try to down-tone, whether integral and limit can be exchanged will be done more carefully.

However, variance is not "linear". Variance can be viewed as a "quadratic functional", which we might discuss in later lectures.

$$\mathsf{var}\left(X_1 \pm X_2\right) = \mathsf{var}\left(X_1\right) + \mathsf{var}\left(X_2\right) \pm 2\mathsf{cov}\left(X_1, X_2\right).$$

If $\mathsf{cov}\left(X_1, X_2\right) = 0$ (e.g. implied by $X_1 \perp\!\!\!\perp X_2$ [$\perp\!\!\!\perp$ denotes independence]), then

$$\mathsf{var}\left(X_1 \pm X_2\right) = \mathsf{var}\left(X_1\right) + \mathsf{var}\left(X_2\right).$$

In general, we can still control (i.e. upper bound) variance without computing covariance. The following inequality turns out to be quite useful in research, because we sometimes only need a bound on the variance rather than its exact value.

$$\mathsf{var}\left(X_1 \pm X_2\right) \leq 2\mathsf{var}\left(X_1\right) + 2\mathsf{var}\left(X_2\right)$$

the proof of which follows from the trivial inequality $(a \pm b)^2 \leq 2a^2 + 2b^2$.

2.1.2 Geometric interpretation of expectation and variance

We have the following variational characterization of expectation and variance:

Proposition 1. *If* $\mathbb{E}(X^2) < \infty$,

$$\mathsf{var}\,(X) \equiv \min_{c \in all\ constant\ functions} \mathbb{E}\,(X-c)^2 \ \ and\ \mathbb{E}\,(X) \equiv \mathop{\mathsf{arg\ min}}_{c \in all\ constant\ functions} \mathbb{E}\,(X-c)^2$$

Proof. Take derivative w.r.t. c and solve c such that the derivative is zero. You will find $c = \mathbb{E}\,(X)$. Then evaluate the second derivative at $c = \mathbb{E}\,(X)$, and it should be nonnegative. □

Remark 5. By Proposition 1, we have the following geometric interpretation of $\mathbb{E}\,(X)$: $\mathbb{E}\,(X)$ is the $L_2(P_X)$-projection of the r.v. X onto the space of all constant functions $\{c\}$, i.e. the closest constant c to X in $L_2(P_X)$ distance between X and c, i.e. $\{\mathbb{E}\,(X-c)^2\}^{1/2}$.

3 Conditioning

Conditioning is the heart and soul of statistics – almost every statistical analysis is conditioning on something, e.g. assumptions, and sometimes data. Conditional probability distribution functions or conditional probability measures, however, are more trickier to define than their marginal counterpart. In fact, people have argued that this is the reason why measure theory should be used in probability.

As we have seen in the previous section, probability distribution/measure can be defined by taking expectation over an indicator function, we only consider conditional expectation in this section.

Unfortunately, because we do not have enough time to cover all the measure theoretical issues, we content ourselves with the following definition which does not rely on the existence of p.d.f. or p.m.f. and requires minimal knowledge on measure theory (Lebesgue decomposition theorem, Radon-Nikodym [i.e. change of measure]). For rigorous development, please read Chapter 4 of Amir Dembo's lecture notes (if you haven't seen measure theory yet, you also need to read Chapter 1). First, consider two random variables X and Y with joint distributed with distribution function $P_{X,Y}$.

Definition 6. The conditional expectation $\mathbb{E}[Y|X]$ is the almost surely (w.p.1.) unique function $g(X)$ that uncorrelates the residual $Y - g(X)$ from all measurable functions $h(X)$ of X, i.e.

$$\mathbb{E}\left\{(Y - g(X))\, h(X)\right\} = 0.$$

Note however that this definition is not constructive. But it has all the properties that a conditional expectation needs to satisfy. For the sake of understanding, we can check if the above definition is true if the conditional p.d.f.'s of $Y|X=x$ exist.

$$\begin{aligned}\mathbb{E}\left[(Y-\mathbb{E}(Y|X))h(X)\right] &= \int\int yh(x)f_{X,Y}(x,y)dxdy - \int \mathbb{E}(Y|X=x)h(x)f_X(x)dx \\ &= \int\int yh(x)f_{X,Y}(x,y)dxdy - \int\left\{\int yf_{Y|X}(y|X=x)dy\right\}h(x)f_X(x)dx \\ &= \int\int yh(x)f_{X,Y}(x,y)dxdy - \int\int yh(x)f_{Y|X}(y|X=x)f_X(x)dydx \\ &= \int\int yh(x)f_{X,Y}(x,y)dxdy - \int\int yh(x)f_{X,Y}(x,y)dydx \\ &= 0.\end{aligned}$$

Note 7. $\mathbb{E}[Y|X]$ is a random variable, but there is no randomness in $\mathbb{E}[Y|X=x]$.

Based on Definition 6, we can prove the following properties of conditional expectation:

Proposition 2.

1. $\mathbb{E}[Y|X]$ *is unique almost surely.*
2. $\mathbb{E}\left(k(X)Y|X\right)=k(X)\mathbb{E}\left(Y|X\right)$
3. *Tower law* $\mathbb{E}\left(Y\right)=\mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)$
4. $\mathsf{var}\left(Y\right)=\mathbb{E}\left(\mathsf{var}\left(Y|X\right)\right)+\mathsf{var}\left(\mathbb{E}\left(Y|X\right)\right)$
5. $\mathsf{cov}\left(Y_1,Y_2\right)=\mathbb{E}\left(\mathsf{cov}\left(Y_1,Y_2|X\right)\right)+\mathsf{cov}\left(\mathbb{E}\left(Y_1|X\right),\mathbb{E}\left(Y_2|X\right)\right)$ *and* $\mathsf{cov}\left(X,Y\right)=\mathsf{cov}\left(X,\mathbb{E}\left(Y|X\right)\right)$.

Note 8. In class, I wrote in Definition 6 "all **bounded** and measurable functions $h\ldots$". After class, a student asked whether boundedness is necessary for (2). It is not necessary to assume boundedness but it is also not wrong to add "boundedness". Boundedness implies the existence $\mathbb{E}\left[(Y-\mathbb{E}(Y|X))h(X)\right]$ for measurable h by the usual "4-step" strategy in a measure-theoretic proof, that is: first show existence by simple functions i.e. indicator functions, second by linear combination of simple functions, third by taking limit to include all positive measurable functions and fourth by decomposing a measurable function $h=h_+-h_-$ into a positive part h_+ and negative part h_- (note that $h_-\geq 0$). Again, for better reference, you should read Amir Dembo's lecture notes if interested.

Proof.

1. Suppose two different functions $g_1(X)$ and $g_2(X)$ satisfy the conditions given in Definition 6. Then

$$\begin{aligned}&\mathbb{E}\left\{g_1(X)h(X)\right\}=\mathbb{E}\left\{g_2(X)h(X)\right\} \\ \Rightarrow&\mathbb{E}\left\{\left(g_1(X)-g_2(X)\right)h(X)\right\}=0 \\ \overset{\text{Take } h=\mathsf{sign}(g_1-g_2)}{\Rightarrow}&\mathbb{E}\left\{|g_1(X)-g_2(X)|\right\}=0.\end{aligned}$$

Since $|g_1(X)-g_2(X)|\geq 0$, $|g_1(X)-g_2(X)|=0$ almost surely, i.e. $g_1(X)=g_2(X)$ almost surely.

2. Define $h'(X) = h(X)k(X)$. Then

$$\begin{aligned}&\mathbb{E}\left\{\left(k(X)Y - k(X)\mathbb{E}\left(Y|X\right)\right)h(X)\right\}\\ &= \mathbb{E}\left\{\left(Y - \mathbb{E}\left(Y|X\right)\right)k(X)h(X)\right\}\\ &= \mathbb{E}\left\{\left(Y - \mathbb{E}\left(Y|X\right)\right)h'(X)\right\} = 0\end{aligned}$$

   by definition of $\mathbb{E}\left(Y|X\right)$.

3. By Definition 6 with $h(x) \equiv 1$

$$\mathbb{E}\left\{(Y - \mathbb{E}(Y|X))\right\} = 0 \Leftrightarrow \mathbb{E}\left(Y\right) = \mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right).$$

4.

$$\begin{aligned}\mathsf{var}\left(Y\right) &= \mathbb{E}\left(Y^2\right) - \left\{\mathbb{E}\left(Y\right)\right\}^2 = \mathbb{E}\left(\mathbb{E}\left(Y^2|X\right)\right) - \left\{\mathbb{E}\left(Y\right)\right\}^2\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right) + \left\{\mathbb{E}\left(Y|X\right)\right\}^2\right] - \left\{\mathbb{E}\left(Y\right)\right\}^2\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right)\right] + \mathbb{E}\left[\left\{\mathbb{E}\left(Y|X\right)\right\}^2\right] - \left\{\mathbb{E}\left(Y\right)\right\}^2\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right)\right] + \left\{\mathsf{var}\left[\mathbb{E}\left(Y|X\right)\right] + \left[\mathbb{E}\left(\mathbb{E}\left(Y|X\right)\right)\right]^2\right\} - \left\{\mathbb{E}\left(Y\right)\right\}^2\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right)\right] + \left\{\mathsf{var}\left[\mathbb{E}\left(Y|X\right)\right] + \left\{\mathbb{E}\left(Y\right)\right\}^2\right\} - \left\{\mathbb{E}\left(Y\right)\right\}^2\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right)\right] + \mathsf{var}\left[\mathbb{E}\left(Y|X\right)\right].\end{aligned}$$

   From the calculations above, you probably see more easily why I can take $\mathbb{E}Y = 0$ w.l.o.g. in class.

5. Left as exercise.

□

Finally, similar to expectation, we give the following geometric interpretation of conditional expectations.

Proposition 3. *Let X and Y be jointly distributed with $\mathbb{E}Y^2 < \infty$. Then the random variable $g(X)$ that minimizes the mean squared error (MSE) $\mathbb{E}\left(Y - g(X)\right)^2$ of predicting Y is $\mathbb{E}\left(Y|X\right)$.*

Proof. Hint: using Tower law $\mathbb{E}\left(Y - g(X)\right)^2 = \mathbb{E}\left\{\mathbb{E}\left[\left(Y - g(X)\right)^2|X\right]\right\}$. □

MSE is one of the most important quantity in statistics and machine learning because it has the following interpretable decomposition:

$$\begin{aligned}\mathrm{MSE}(g(X), Y) &= \mathbb{E}\left(Y - g(X)\right)^2 = \mathbb{E}\left(Y - \mathbb{E}\left(Y|X\right) + \mathbb{E}\left(Y|X\right) - g(X)\right)^2\\ &= \mathbb{E}\left[\mathbb{E}\left\{\left(Y - \mathbb{E}\left(Y|X\right)\right)^2|X\right\} + \left\{\mathbb{E}\left(Y|X\right) - g(X)\right\}^2\right]\\ &= \mathbb{E}\left[\mathsf{var}\left(Y|X\right) + \left\{\mathbb{E}\left(Y|X\right) - g(X)\right\}^2\right].\end{aligned}$$

Thus MSE = Var + Bias^2. In many regression/supervised machine learning problems, MSE is the target loss function. By minimizing the MSE, we at least try to minimize both variance and bias of using $g(X)$ to predict Y at the same time, and in principle we are hoping that the algorithm will neither overfit (undersmooth) nor under-fit (oversmooth).

Note 9. It is also the squared $L_2(P_{X,Y})$-distance between Y and $g(X)$.

Note 10. Sometimes, in particular when $Y|X = x$ has heavy tails, conditional expectation is no longer of interest because mean does not represent the majority of the population. There we sometimes try to minimize the absolute mean deviation instead: $\mathbb{E}|Y - g(X)|$ and the minimizer is the conditional median of Y given X, which is the so-called median regression. However, compared to MSE, absolute mean deviation is not easy to solve computationally. Therefore computation becomes one of the central topics in statistics or econometrics research on median regression, or quantile regression in general [**?**]. It also has some interesting connection with optimal transport [**?**].

4 (Conditional) independence

Definition 11. $X_1, \ldots, X_n$ are independent iff their joint c.d.f. factorizes into the product of each marginal c.d.f. i.e.

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n)$$

for all $x_1, \ldots, x_n$.

Conditional independence can be defined similarly.

If $X_1 \perp\!\!\!\perp X_2$, then $\mathsf{cov}(X_1, X_2) = 0$; but the reverse direction is not necessarily true. When (X_1, X_2) is jointly normally distributed, then $\mathsf{cov}(X_1, X_2) = 0$ implies $X_1 \perp\!\!\!\perp X_2$. For this reason, people studying statistical inference on graphical models (in which no edge between two vertices means independence/conditional independence) often study Gaussian graphical models.

Similarly, if $X_1 \perp\!\!\!\perp X_2$, $\mathbb{E}(X_1|X_2) = \mathbb{E}(X_1)$.

With independence in mind, we can be creative in our calculations or proofs. In particular, we can rewrite variance formula as follows: for a r.v. $X \in P_X$, create an independent copy X' of X. Then

$$\begin{aligned}\mathsf{var}\left(X\right) &\equiv \mathbb{E}\left(X^2\right) - \{\mathbb{E}\left(X\right)\}^2 \equiv \frac{1}{2}\left\{2\mathbb{E}\left(X^2\right) - 2\{\mathbb{E}\left(X\right)\}^2\right\} \\ &= \frac{1}{2}\left\{\mathbb{E}\left(X^2\right) + \mathbb{E}\left(X'^2\right) - 2\mathbb{E}\left(X\right)\mathbb{E}\left(X'\right)\right\} = \frac{1}{2}\mathbb{E}\left(X - X'\right)^2.\end{aligned}$$

Recall that the sample variance of $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_X$ is

$$\widehat{S} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \text{ with } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

which is an unbiased estimator of $\mathsf{var}\left(X\right)$. With this "new" formula of variance, we can come up with a "different" variance estimator:

$$\widehat{S}' = \frac{1}{2n(n-1)} \sum_{1 \leq i_1 \neq i_2 \leq n} (X_{i_1} - X_{i_2})^2$$

which is a second-order U-statistic. Because $X_{i_1} \perp\!\!\!\perp X_{i_2}$, it is obvious that $\widehat{S}'$ is also an unbiased estimator of $\mathsf{var}(X)$. Actually you can show, after some algebra, that $\widehat{S}' \equiv \widehat{S}$. If you want to compute $\widehat{S}'$ in computer exactly following its U-statistic formula, it takes $O(n^2)$ summations. However, $\widehat{S}$ only takes $O(n)$ summations. The quadratic vs. linear time computation will make a difference in the world of big data.

Remark 12. When sample size is large, the sample variance and $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are almost the same. Division by $n-1$ rather than n to make $\widehat{S}$ an unbiased estimator of $\mathsf{var}(X)$ is due to the correlation between $\bar{X}$ and X_i for $i = 1, \ldots, n$.

Creating an independent copy has very strong statistical flavor, and is called "exchangeable pair" or "replica" trick. This trick is probably the most important statistical contribution to pure mathematics: many novel proofs for central limit theorems (CLT) (i.e. Stein's method) and concentration inequalities are developed using this trick [**?**].

5 Generating functions

We first define the moment generating function (MGF) of a r.v. $X \sim P_X$. MGF is another way of determining distributions, in addition to the c.d.f. It is also powerful for deriving distributions for convolution i.e. $X + Y$ and moments (from its name).

Definition 13. A r.v. X has a MGF if $M(t) = \mathbb{E}e^{tX} < \infty$ for t in an open interval containing 0.

MGF does not always exist. For example, if X is log-normally distributed, then its MGF does not exist, although all of its moments exist. To see this, $X = e^Z$ with $Z \sim N(0,1)$. Then when $t > 0$

$$\mathbb{E}\left(e^{tX}\right) = \int_0^\infty e^{tx}\frac{1}{2\pi x}e^{-(\log(x))^2/2}dx \geq c\int_t^\infty \frac{1}{x}dx = \log(x)|_c^\infty = \infty$$

where given some $c > 0$, t is chosen such that for all $x \geq t$, $e^{tx}\frac{1}{2\pi}e^{-(\log(x))^2/2} > c$.

Fact 2. *For two r.v.s. X_1, X_2, if they have the same MGF $M(t)$ in some open neighborhood of 0, then $X_1 \sim X_2$.*

Fact 3. *MGF, if it exists, is a convex function of t.*

Theorem 14. *If X has an MGF $M(t)$, then $\mathbb{E}(X^m) \equiv M^{(m)}(0)$, where $M^{(m)}(0)$ is the m-th derivative of $M(t)$ with t evaluated at 0.*

Theorem 15. *If $X_1 \perp\!\!\!\perp X_2$, $M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t)$.*

MGF can be viewed as the "Laplace transformation" of the probability distribution. Just like MGF, Laplace transformation does not always exist. But if we expand our field from $\mathbb{R}$ to $\mathbb{C}$ (the field of complex numbers), we have Fourier transformation that always exists, which gives us characteristic function (CF) $C(t) = \mathbb{E}e^{\mathsf{i}tX}$, where $\mathsf{i} = \sqrt{-1}$. CF also uniquely determines probability distributions, following from classical Fourier analysis.

The most fundamental proof idea for CLT is by showing the CF of sample average converges to that of a normal.

6 Important functional inequalities

Proposition 4 (Jensen's inequality)**.** *Let f be a convex (concave) function. Then $\mathbb{E}(f(X)) \geq f(\mathbb{E}(X))$ ($\mathbb{E}(f(X)) \leq f(\mathbb{E}(X))$).*

Proof. "Proof" by picture is good enough in this case and it helps you remember the direction of the inequality.

But if you are interested, a rigorous proof can be done by invoking the "supporting hyperplane theorem" because here we only assume f to be convex rather than continuous or differentiable (if so, we can use Taylor expansion). WLOG, let's consider 1d case. For any convex function f, at any point $(x, g(x))$, there exists a supporting line going through the point $(x, f(x))$ and lying below the graph of f. Let $\mu = \mathbb{E}X$. Denote $L(x) = a + bx$ as the supporting line at $(\mu, f(\mu))$. Thus $f(x) \geq a + bx$. Taking expectation on both sides, we have

$$\mathbb{E}f(X) \geq a + b\mu = L(\mu) = f(\mu) = f(\mathbb{E}(X)).$$

□

Jensen's inequality can be used to prove the following corollary:

Corollary 1. *Arithmetic mean (AM) $\geq$ Geometric mean (GM) $\geq$ Harmonic mean (HM) over positive variables.*

Proof. Take two variables $x, y > 0$ and $p, q \in [0, 1]$ s.t. $p + q = 1$. Then arithmetic mean is $px + qy$, the geometric mean is $x^p y^q$ and the harmonic mean is $\frac{1}{p/x+q/y}$. Thus you can think of a random variable $W = x$ w.p. p and $W = y$ w.p. q.

Notice that $\log(x)$ is a concave function on $[0, \infty)$. Then $\mathbb{E}(\log(W)) \leq \log(\mathbb{E}(W))$ by Jensen, i.e. $p \log(x) + q \log(y) \leq \log(px + qy) \Leftrightarrow x^p y^q \leq px + qy$, hence AM $\geq$ GM.

Again notice that $\log(x)$ is a concave function on $(0, \infty)$. Then $\mathbb{E}(\log(1/W)) \leq \log(\mathbb{E}(1/W))$ by Jensen, i.e. $p \log(1/x) + q \log(1/y) \leq \log(p/x + q/y) \Leftrightarrow x^{-p} y^{-q} \leq \frac{p}{x} + \frac{q}{y} \Leftrightarrow x^p y^q \geq \frac{1}{\frac{p}{x}+\frac{q}{y}}$, hence GM $\geq$ HM. □

Proposition 5 (Cauchy-Schwarz inequality)**.**

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

Note 16. Cauchy-Schwarz inequality is arguably the most important inequality in mathematics. We give a variational proof below.

Proof. Trivially, we have $\mathbb{E}(X - cY)^2 \geq 0$ for any $c \in \mathbb{R}$. Hence $\mathsf{min}_{c\in\mathbb{R}} \mathbb{E}(X - cY)^2 \geq 0$, with $c = \frac{\mathbb{E}(Y^2)}{\mathbb{E}(XY)}$ achieving the minimum. Plugging in c, we have $|\mathbb{E}(XY)|^2 \leq \mathbb{E}(X^2)\mathbb{E}(Y^2)$. □

6.1 Fortuin-Kasteleyn-Ginibre (FKG) Inequality

Proposition 6. *X is an random variable and g, h are monotonically increasing functions. Then we have $\mathsf{cov}(g(X), h(X)) \geq 0$*

Proof. We assume that there are two random variables $X' \perp\!\!\!\perp X$, $X, X' \sim P_x$. (Replica trick)

g, h are monotonically increasing functions$\Rightarrow (g(X) - g(X'))(h(X) - h(X')) \geq 0$.

$$\begin{aligned}
&\mathbb{E}\left\{(g(X) - g(X'))(h(X) - h(X'))\right\} \geq 0 \\
&\Rightarrow \mathbb{E}g(X)h(X) + \mathbb{E}g(X')h(X') - \mathbb{E}g(X)h(X') - \mathbb{E}g(X')h(X) \geq 0 \\
&\Rightarrow 2\mathbb{E}g(X)h(X) - 2\mathbb{E}g(X)\mathbb{E}h(X) \geq 0
\end{aligned}$$

The final inequality holds because of independence and identical distribution of X and X' □

6.2 Norm of random variables

The $L_r(1 \leq r \leq \infty)$ norm of random variables can be defined as following:

$$||X||_r = [\mathbb{E}|X|^r]^{\frac{1}{r}}, \quad 1 \leq r < \infty$$
$$||X||_\infty = \inf \{C \geq 0, P(|X| > C) = 0\}$$

Note 17. It is important to show that triangle inequality holds for $||X||_r$. Minkowski inequality (which will be discussed later) is applied when $1 \leq r < \infty$. When $r = \infty$,we have the following proposition:

Proposition 7. $||X_1 + X_2||_\infty \leq ||X_1||_\infty + ||X_2||_\infty$

Proof. Let $C_1 = ||X_1||_\infty, C_2 = ||X_2||_\infty, C^* = C_1 + C_2$.

$$\begin{aligned} &P(|X_1 + X_2| > C^*) \leq P(|X_1 + |X_2| > C^*) \\ &= P(|X_1 + |X_2| > C^*, |X_2| \leq C_2) + P(|X_1 + |X_2| > C^*, |X_2| > C_2) \\ &\leq P(|X_1| \geq C_1) + P(|X_2| \geq C_2) = 0. \end{aligned}$$

Thus, $C^* \in \{C \geq 0, P(|X_1 + X_2| > C) = 0\} \Rightarrow ||X_1 + X_2||_\infty \leq C^* = ||X_1||_\infty + ||X_2||_\infty$ □

Remark 18. The red formula in the above proof is a very useful ‘event decomposition’ strategy when bounding probabilities. Separate event $\{|X_1 + |X_2| > C^*\}$ into 'bad' event $\{|X_1 + |X_2| > C^*, |X_2| \leq C_2\}$ and 'good' event $\{|X_1 + |X_2| > C^*, |X_2| > C_2\}$. For 'bad' event we can bound the probability by $\{|X_1| \leq C_1\}$ because $X_1 + |X_2| > C^*, |X_2| \leq C_2 \Rightarrow |X_1| > C^* - C_2 = C_1$

6.3 Hölder Inequality

Proposition 8. *If $r, s \geq 1$ are conjugate exponent, that is $\frac{1}{r} + \frac{1}{s} = 1$, then the random variables $X \in L^r(i.e.\mathbb{E}|X|^r < \infty)$ and $Y \in L^s$ satisfy:*

$$|\mathbb{E}XY| \leq ||X||_r ||Y||_s$$

When $r = s = 2$, it is Cauchy Schwarz Inequality.

Proof. Let $p = 1/r, q = 1/s, p + q = 1$, then $|XY| = (|X|^r)^p(|Y|^s)^q \leq p|X|^r + q|Y|^s$ (which is also called ‘Young's Inequality’). Let $X^\dagger = \frac{X}{||X||_r}, Y^\dagger = \frac{Y}{||Y||_s}$. Then from Young's inequality we have $|X^\dagger Y^\dagger| \leq p|X^\dagger|^r + q|Y^\dagger|^s$.

$$\begin{aligned} &\mathbb{E}|X^\dagger Y^\dagger| \leq p\mathbb{E}|X^\dagger|^r + q\mathbb{E}|Y^\dagger|^s = p||X^\dagger||_r^r + q||Y^\dagger||_s^s = p + q = 1 \\ \Rightarrow &\mathbb{E}|X^\dagger Y^\dagger| \leq 1 \Rightarrow \mathbb{E}|XY| \leq ||X||_r ||Y||_s \end{aligned}$$

□

6.4 Minkowski Inequality

Proposition 9. *For any $1 \le r < \infty$ and any random variables $X, Y \in L^r$, we have:*

$$||X+Y||_r \le ||X||_r + ||Y||_r$$

Proof. When $r = 1$, it is obvious that $||X+Y||_1 = \mathbb{E}|X+Y| \le \mathbb{E}|X| + \mathbb{E}|Y|$.

When $r > 1$, we used a proof idea called 'bootstrapping' from simple case $r = 1$ to general case $r \ge 1$:

$$\begin{aligned}
||X+Y||_r^r &= \mathbb{E}|X+Y|^r = \mathbb{E}|X+Y||X+Y|^{r-1} \\
&\le \mathbb{E}|X||X+Y|^{r-1} + \mathbb{E}|Y||X+Y|^{r-1} \\
&= \mathbb{E}|X||Z| + \mathbb{E}|Y||Z| = ||XZ||_1 + ||YZ||_1 \quad (\text{Let} Z = |X+Y|^r) \\
&\le ||X||_r ||Z||_s + ||Y||_r ||Z||_s \\
||Z||_s &= \left\{\mathbb{E}[|X+Y|^{r-1}]^s\right\}^{\frac{1}{s}} = \{\mathbb{E}|X+Y|^r\}^{1-\frac{1}{r}} = ||X+Y||_r^{r-1} \\
\Rightarrow ||X+Y||_r^r &\le ||X||_r ||X+Y||_r^{r-1} + ||Y||_r ||X+Y||_r^{r-1} \\
\Rightarrow ||X+Y||_r &\le ||X||_r + ||Y||_r
\end{aligned}$$

□

Note 19. Considering the situation that $0 < r < 1$, we define a 'pseudo norm': $|||X|||_r = [\mathbb{E}|X|^r]^{\frac{1}{r}}$. Because of the difference between concavity and convexity, pseudo norm does not satisfy triangle inequality. However, we have following properties:

$$\begin{aligned}
&|||X+Y|||_r \ge |||X|||_r + |||Y|||_r \\
&|||X+Y|||_r \le 2^{\frac{1}{r}-1}\left[|||X|||_r + |||Y|||_r\right]
\end{aligned}$$

for X, Y random variables satisfying $P(X \ge 0) = 1$ or $P(X \le 0) = 1$ and $P(Y \ge 0) = 1$ or $P(Y \le 0) = 1$.

Proof. 1. First inequality. Take $\delta \coloneqq |||X|||_r / \left(|||X|||_r + |||Y|||_r\right)$.

$$\begin{aligned}
|||X+Y|||_r^r &= \mathbb{E}|X+Y|^r = \mathbb{E}\left|\delta\frac{X}{\delta} + (1-\delta)\frac{Y}{1-\delta}\right|^r \\
&\ge \mathbb{E}\left\{\delta\left|\frac{X}{\delta}\right|^r + (1-\delta)\left|\frac{Y}{1-\delta}\right|^r\right\} \text{ due to Jensen} \\
&= \delta|||X|||_r^r\delta^{-r} + (1-\delta)|||Y|||_r^r(1-\delta)^{-r} \\
&= \delta\left(|||X|||_r + |||Y|||_r\right)^r + (1-\delta)\left(|||X|||_r + |||Y|||_r\right)^r \\
&= \left(|||X|||_r + |||Y|||_r\right)^r \\
\Rightarrow \ |||X+Y|||_r &\ge |||X|||_r + |||Y|||_r.
\end{aligned}$$

2. Second inequality.

$$\begin{aligned}
|||X+Y|||_r^r &= \mathbb{E}|X+Y|^r = \mathbb{E}\left\{|X+Y|^r\frac{|X|}{|X|+|Y|}+|X+Y|^r\frac{|Y|}{|X|+|Y|}\right\}\\
&\leq \mathbb{E}|X|^r+\mathbb{E}|Y|^r \equiv |||X|||_r^r+|||Y|||_r^r \text{ due to } |X+Y|\leq|X|+|Y|\\
\Rightarrow\ |||X+Y|||_r &\leq \{|||X|||_r^r+|||Y|||_r^r\}^{1/r}\\
&= 2^{1/r}\left\{\frac{1}{2}|||X|||_r^r+\frac{1}{2}|||Y|||_r^r\right\}^{1/r}\\
&\leq 2^{1/r}\frac{1}{2}\left(|||X|||_r+|||Y|||_r\right) \text{ due to Jensen}\\
&= 2^{1/r-1}\left(|||X|||_r+|||Y|||_r\right).
\end{aligned}$$

□

6.5 Other important results

- $||X||_p \leq ||X||_s, \quad 1\leq p\leq s\leq\infty$

  Proof. Let $Y=|X|^p, Z=1$ then $\mathbb{E}|X|^p=\mathbb{E}|YZ|\leq||Y||_{p'}||Z||_{s'}, \quad 1/p'+1/s'=1$. Let $p'=\frac{s}{p}, s'=\frac{1}{1-1/p'}$, then

  $$\begin{aligned}
  ||Y||_{p'} &= \mathbb{E}\left\{[|X|^p]^{\frac{s}{p}}\right\}^{\frac{p}{s}} = \mathbb{E}[|X|^s]^{\frac{p}{s}} = ||X||_s^p\\
  \Rightarrow ||X||_p^p &\leq ||X||_s^p \Rightarrow ||X||_p\leq||X||_s
  \end{aligned}$$

  □

  Note 20. This inequality can be also proven by Jensen's inequality (let $f(x)=x^{\frac{s}{p}}$, which is a convex function when $1\leq p\leq s\leq\infty$).

- Suppose $p\geq1$, then $||\mathbb{E}(Y|X)||_p\leq||Y||_p$. (Informally: 'projections' contracting norm)

  Proof.

  $$\begin{aligned}
  ||\mathbb{E}(Y|X)||_p^p &= \mathbb{E}[|\mathbb{E}(Y|X)|^p]\\
  &\leq \mathbb{E}[\mathbb{E}(|Y|^p|X)] = \mathbb{E}|Y|^p = ||Y||_p^p \quad \text{(Jensen)}
  \end{aligned}$$

  □

- X is a random variable with mean μ, standard deviation σ and median m. Then we have $|\mu-m|\leq\sigma$

  Proof.

  $$\begin{aligned}
  |\mu-m| &= |\mathbb{E}(X-m)|\leq\mathbb{E}|X-m| && \text{(Jensen)}\\
  &\leq \mathbb{E}|X-\mu|\leq\sigma && \text{(Variational interpretation of median)}
  \end{aligned}$$

  □

- Kullback Leibler (KL) divergence: $f(x), g(x)$ are two density functions. $D(f(x)||g(x)) = \int_{\mathbb{R}} \log \frac{f(x)}{g(x)} f(x)dx$, $D(f(x)||g(x)) \geq 0$

  Proof.

  $$\begin{aligned} D(f(x)||g(x)) = \mathbb{E}_f \left(-\log \frac{g(x)}{f(x)} \right) &\geq -\log \mathbb{E} \frac{g(x)}{f(x)} && \text{(Jensen)} \\ &= -\log \left(\int_{\mathbb{R}} \frac{g(x)}{f(x)} f(x)dx \right) = 0 && \text{(property of density)} \end{aligned}$$

  □

One of the most powerful functional inequalities is the Brascamp-Lieb inequality, which subsumes most of the above functional inequalities.

7 Different modes of stochastic convergence

For convergence in distribution (in law), convergence in probability, and convergence almost surely, you can check Larry Wasserman's lecture notes 4, 5, and 6 if unfamiliar with these concepts.

7.1 Exercises on o_p and O_p notation

Pattern: under moment conditions on X, denote $\mu = \mathbb{E}X$ and $\sigma^2 = \mathsf{var}X$. $\bar{X} - \mu = o_p(1) = O_p(n^{-1/2})$ where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\sqrt{n}(\bar{X} - \mu) = O_p(1)$.

Exercises:

(1) $\omega(1)\sqrt{n}(\bar{X} - \mu) \neq O_p(1)$

(2) say $X^{(n)}$ is a random variable indexed by n with mean 0 and variance $1/\log n$. Let $\bar{X}^{(n)} = n^{-1}\sum_{i=1}^n X_i^{(n)}$. $\sqrt{n}(\bar{X}^{(n)} - \mu) =?$

(3)

We start this section by considering several important theorems in real analysis that are adapted to the probability case: Note that my proof might not be super rigorous.

Theorem 21 (Bounded convergence theorem (BCT))**.** *For a sequence of random $\{X_n\}$, $|X_n| \leq C$ for some constant $C > 0$ almost surely. If $X_n \xrightarrow{p} X$ for some random variable X, then $\lim_n \mathbb{E}X_n = \mathbb{E}X$.*

Proof. For convenience, we denote $X = \lim_n X_n$. First, we show that X is also bounded by C almost surely. Take any $\epsilon > 0$,

$$\begin{aligned} &\mathbb{P}(|X| > C + \epsilon) = \mathbb{P}(|X - X_n + X_n| > C + \epsilon) \\ &\leq \mathbb{P}(|X - X_n| + |X_n| > C + \epsilon) \\ &= \mathbb{P}(|X - X_n| + |X_n| > C + \epsilon, |X_n| \leq C) + \mathbb{P}(|X - X_n| + |X_n| > C + \epsilon, |X_n| > C) \\ &\leq \mathbb{P}(|X - X_n| > \epsilon) + \mathbb{P}(|X_n| > C) \\ &= \mathbb{P}(|X - X_n| > \epsilon) + 0 \\ &\to 0 \text{ as } n \to \infty \text{ by } X_n \xrightarrow{p} X. \end{aligned}$$

Given the above result,

$$\begin{aligned}|\mathbb{E}X_n - \mathbb{E}X| &\le \mathbb{E}|X_n - X| \\ &= \mathbb{E}|X_n - X|\mathbb{1}\{|X_n - X| > \epsilon\} + \mathbb{E}|X_n - X|\mathbb{1}\{|X_n - X| \le \epsilon\} \\ &\le 2C\mathbb{P}(|X_n - X| > \epsilon) + \epsilon.\end{aligned}$$

The first term converges to 0 and ϵ can be arbitrarily small. □

Theorem 22 (Monotone convergence theorem (MCT)). *If $X_1 \le X_2 \le \cdots$ and $X_n \xrightarrow{p} X$, then $\lim_n \mathbb{E}X_n = \mathbb{E}X$.*

Proof. We first assume $\mathbb{E}X < \infty$. Since X_n is monotonically non-decreasing, $X_n \uparrow X$ in probability, so $\mathbb{E}X_n \le \mathbb{E}X$ which gives us $\limsup_n \mathbb{E}X_n \le \mathbb{E}X$.

For the other direction, we will use a truncation argument. Take bounded $W \le X$ but $\mathbb{E}W \ge \mathbb{E}X - \epsilon$ for any given $\epsilon > 0$. Now truncate X_n by $W_n = X_n \wedge W \equiv \mathsf{min}\{X_n, W\}$. Then

$$\begin{aligned}\mathbb{P}(|W_n - W| > \epsilon) &= \mathbb{P}(|W_n - W| > \epsilon, X_n \le W) + \mathbb{P}(|W_n - W| > \epsilon, X_n > W) \\ &= \mathbb{P}(|X_n - W| > \epsilon, X_n \le W) + \underbrace{\mathbb{P}(|W - W| > \epsilon, X_n > W)}_{\equiv 0} \\ &= \mathbb{P}(|X_n - W| > \epsilon, X_n \le W) \\ &\le \mathbb{P}(|X_n - X| > \epsilon, X_n \le W) \\ &\le \mathbb{P}(|X_n - X| > \epsilon) \to 0.\end{aligned}$$

So $W_n \xrightarrow{p} W$. Now by BCT, and $X_n \ge W_n$

$$\mathbb{E}[X_n] \ge \mathbb{E}[W_n] \to \mathbb{E}[W] \ge \mathbb{E}[X] - \epsilon$$

so $\liminf_n \mathbb{E}[X_n] \ge \mathbb{E}[X]$ which together with $\limsup_n \mathbb{E}X_n \le \mathbb{E}X$, gives $\lim_n \mathbb{E}X_n = \mathbb{E}X$.

If $\mathbb{E}X = \infty$. Take $W \le X$ but $\mathbb{E}W \ge c$ for any constant $c > 0$. Replacing $\mathbb{E}X$ by c in the above analysis, we get

$$\liminf_n \mathbb{E}[X_n] \ge c.$$

Since c can be taken arbitrarily large, we have $\liminf_n \mathbb{E}[X_n] = \infty$. □

Theorem 23 (Fatou's lemma). *$X_1, X_2 \cdots$ nonnegative random variables. Then*

$$\liminf_n \mathbb{E}[X_n] \ge \mathbb{E}[\liminf_n X_n]$$

Proof. Define $Y = \liminf_n X_n$ and $Y_n = \inf\{X_m : m \ge n\}$. Then $Y_n \xrightarrow{p} Y$ and Y_n monotonically non-decreasing. So by MCT, $\mathbb{E}Y_n \to \mathbb{E}Y$. But we know $\mathbb{E}Y_n \le \mathbb{E}X_n$, so $\mathbb{E}[\liminf_n X_n] \equiv \mathbb{E}Y = \liminf_n \mathbb{E}Y_n \le \liminf_n \mathbb{E}X_n$. □

Theorem 24 (Dominated convergence theorem (DCT)). *A sequence of random variables $\{X_n\}$ such that $X_n \xrightarrow{p} X$ and $|X_n| \le W$ for all n, for some random variable W with $\mathbb{E}W < \infty$. Then $\lim_n \mathbb{E}X_n = \mathbb{E}X$.*

Proof. $|X_n| \leq W$ is equivalent to $-W \leq X_n \leq W$ equivalent to (1) $-X_n + W \geq 0$ and (2) $X_n + W \geq 0$. By (1), we have, using Fatou's lemma,

$$\begin{aligned}\liminf_n \mathbb{E}[-X_n + W] &\geq \mathbb{E}[\underbrace{\liminf_n -X_n}_{-X} + W] \\ \Rightarrow \liminf_n \mathbb{E}[-X_n] &\geq \mathbb{E}[-X] \\ \Rightarrow \limsup_n \mathbb{E}[X_n] &\leq \mathbb{E}[X].\end{aligned}$$

By (2) and Fatou's lemma,

$$\begin{aligned}\liminf_n \mathbb{E}[X_n + W] &\geq \mathbb{E}[\underbrace{\liminf_n X_n}_{X} + W] \\ \Rightarrow \liminf_n \mathbb{E}[X_n] &\geq \mathbb{E}[X].\end{aligned}$$

Combining the above two, we have $\lim_n \mathbb{E}X_n = \mathbb{E}X$. □

Given the above machinery from real analysis, we can answer the question when can we write

$$\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n]?$$

Theorem 25 (Fubini-Tonelli theorem)**.**

(1) Tonelli: If $X_1, X_2, \cdots \geq 0$, then $\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n]$.

(2) Fubini: If $\sum_{n=1}^{\infty} \mathbb{E}|X_n| < \infty$, then $\mathbb{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[X_n]$.

Proof. Define $Y_n = \sum_{i=1}^{n} X_i$ and $Y = \sum_{i=1}^{\infty} X_i$.

(1) Tonelli: By non-negativity of the random variables, $Y_n \uparrow Y$. Using MCT, we have

$$\lim_n \mathbb{E}[Y_n] = \mathbb{E}[Y].$$

(2) Fubini: $|Y_n| \leq \sum_{i=1}^{n} |X_i| \leq \sum_{i=1}^{\infty} |X_i|$. By Tonelli theorem and non-negativity of the absolute values, we have

$$\mathbb{E}\left[\sum_{i=1}^{\infty} |X_i|\right] = \sum_{i=1}^{\infty} \mathbb{E}|X_i| < \infty.$$

So Y_n satisfies the assumptions in DCT, we have $\lim_n \mathbb{E}[Y_n] = \mathbb{E}[Y]$.

□

If you encounter situations violating the assumptions of Fubini-Tonelli, then just be careful in your analysis. But because they are only sufficient conditions, violating these assumptions does not necessarily rule out the possibility of interchanging expectation and infinite sum. So you still have a chance that the analysis turns out to be simple.

I did not cover the following Borel-Cantelli lemma in class, but it can be quite useful when one does analysis. I believe your probability class will also cover this important lemma.

Theorem 26 (Borel-Cantelli lemma)**.**

(1) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}\left(\limsup_n A_n\right) = 0$.

(2) If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $A_1, A_2, \cdots$ are independent, then $\mathbb{P}\left(\limsup_n A_n\right) = 1$.

Proof.

(1) $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \Rightarrow \lim_n \sum_{k=n}^{\infty} \mathbb{P}(A_k) = 0$ because a convergent series must have tail with limit 0. Thus

$$\mathbb{P}\left(\limsup_n A_n\right) = \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k, \forall\, n\right) \overset{\text{Fix some } n}{\leq} \mathbb{P}\left(\bigcup_{k=n}^{\infty} A_k\right)$$

but the latter converges to 0.

Remark 27. The converse is not true. Define $N = \sum_{n=1}^{\infty} \mathbb{1}_{A_n}$. Suppose $\mathbb{P}(N = n) = c/n^2$ for $n > 1$. Then $\mathbb{P}\left(N < \infty\right) = 1$ but $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \mathbb{E}N = \sum_{n=1}^{\infty} \frac{c}{n^2} n = \infty$.

(2) Define $B_{n,N} := \overset{N}{\underset{k=n}{\cup}} A_k$ and $B_n := \overset{\infty}{\underset{k=n}{\cup}} A_k$. Then $B_{n,N}^c = \overset{N}{\underset{k=n}{\cap}} A_k^c$ and $B_{n,N} \subseteq B_n$, which implies $\mathbb{P}(B_{n,N}) \leq \mathbb{P}(B_n)$. Then

$$1 - \mathbb{P}\left(\overset{\infty}{\underset{k=n}{\cup}} A_k\right) = 1 - \mathbb{P}(B_n) \leq 1 - \mathbb{P}(B_{n,N}) = \mathbb{P}(B_{n,N}^c) \overset{\text{independency}}{=} \prod_{k=n}^{N} \mathbb{P}(A_k^c)$$

$$= \prod_{k=n}^{N} (1 - \mathbb{P}(A_k)) \leq \exp\left\{-\sum_{k=n}^{N} \mathbb{P}(A_k)\right\}.$$

Hence

$$1 - \mathbb{P}\left(\overset{\infty}{\underset{k=n}{\cup}} A_k\right) \leq \lim_{N\to\infty} \exp\left\{-\sum_{k=n}^{N} \mathbb{P}(A_k)\right\} = \exp\left\{-\sum_{k=n}^{\infty} \mathbb{P}(A_k)\right\} = 0.$$

Thus for any n, $\mathbb{P}\left(\overset{\infty}{\underset{k=n}{\cup}} A_k\right) = 1$ which further implies $\lim_n \mathbb{P}\left(\overset{\infty}{\underset{k=n}{\cup}} A_k\right) = 1$. Finally, by BCT,

$$\lim_n \mathbb{P}\left(\overset{\infty}{\underset{k=n}{\cup}} A_k\right) = \mathbb{P}\left(\overset{\infty}{\underset{n=1}{\cap}} \overset{\infty}{\underset{k=n}{\cup}} A_k\right) = 1.$$

□

Remark 28. Recall that $\limsup_n A_n := \overset{\infty}{\underset{n=1}{\cap}} \overset{\infty}{\underset{k=n}{\cup}} A_k$ and $\liminf_n A_n := \overset{\infty}{\underset{n=1}{\cup}} \overset{\infty}{\underset{k=n}{\cap}} A_k$. People usually interpret $\limsup_n A_n$ as A_n occurring infinitely often (i.o.) and $\liminf_n A_n$ as A_n occurring for all but finitely many n's. I don't feel these two interpretations are easy for me to understand. I will simply look at the above set-based definitions: for $\limsup_n A_n$, it is "for all n, at least one of $\{A_n, A_{n+1}, \cdots\}$ must happen"; for $\liminf_n A_n$, it is "at least for one n, all of $\{A_n, A_{n+1}, \cdots\}$ must happen".

Borel-Cantelli lemma has an important corollary, which is the celebrated Kolmogorov's 0-1 law.

Corollary 2 (Kolmogorov's 0-1 law)**.** *$\{A_n\}$ a sequence of independent events, then either $\mathbb{P}(\limsup_n A_n) = 0$ or $\mathbb{P}(\limsup_n A_n) = 1$.*

Finally, we introduce the concept of weak convergence of probability measures.

Definition 29. A sequence of probability measures $\{\mathbb{P}_n\}$ is said to weakly converge to a probability measure $\mathbb{P}$ on the sample space $\mathbb{X}$, denoted as $\mathbb{P}_n \Rightarrow \mathbb{P}$, if $\mathbb{P}_n f \to \mathbb{P} f$, as $n \to \infty$, for any $f \in \mathcal{C}_b(\mathbb{X})$, the space of all bounded and continuous functions on $\mathbb{X}$.

There is an important portmanteau lemma providing useful tools of proving weak convergence. Note that portmanteau lemma holds for any metric space $(\mathbb{X}, \|\cdot\|)$.

Lemma 30 (Portmanteau lemma for weak convergence of probability measures)**.** *The following are equivalent.*

1. $\mathbb{P}_n \Rightarrow \mathbb{P}$
2. $\mathbb{P}_n f \to \mathbb{P} f$ *for* f *bounded and Lipschitz continuous*
3. F *a closed subset of* $\mathbb{X}$*, then* $\limsup_n \mathbb{P}_n F \leq \mathbb{P} F$
4. G *an open subset of* $\mathbb{X}$*, then* $\liminf_n \mathbb{P}_n G \geq \mathbb{P} G$
5. *For a subset* A *of* $\mathbb{X}$ *with* $\mathbb{P}(\partial A) = 0$, $\lim_n \mathbb{P}_n A = \mathbb{P} A$

Proof.
1 $\Rightarrow$ 2: obvious because bounded Lipschitz functions are a subset of $\mathcal{C}_b(\mathbb{X})$.

2 $\Rightarrow$ 4: G is open. Define $f_k(x) = k\|x - G^c\| \wedge 1$. If $x \in G$, then as k gets large, $f_k(x) \to 1$; if $x \in G^c$, then $f_k(x) = 0$. It is easy to check that f_k is k-Lipschitz. Moreover, as k gets large, we allow less $x \in G$ to be strictly below 1. So $f_k \uparrow \mathbb{1}_G$. Thus

$$\liminf_n \mathbb{P}_n G \geq \liminf_n \mathbb{P}_n f_k = \mathbb{P} f_k \uparrow \mathbb{P} G \text{ as } k \to \infty.$$

2 $\Rightarrow$ 3: For any $\epsilon > 0$, take a closed set F' such that $F' \supset F$ and $\mathbb{P}(F') - \mathbb{P}(F) = \epsilon$. Define f as follows:

$$f(x) = \begin{cases} 1 & \text{if } x \in F \\ 0 & \text{if } x \notin F' \\ (*)\text{interpolation between 0 and 1 by a bounded Lipschitz function} & \text{if } x \in F' \setminus F \end{cases}$$

Then by 2, we have $\mathbb{P}_n(f) \to \mathbb{P}(f)$. Note that $\mathbb{1}_F(x) \leq f(x) \leq \mathbb{1}_{F'}(x)$, so

$$\begin{aligned} &\mathbb{P}_n(F) \leq \mathbb{P}_n(f) \to \mathbb{P}(f) \leq \mathbb{P}(F') = \mathbb{P}(F) + \epsilon \\ \Rightarrow &\mathbb{P}_n(F) \leq \mathbb{P}(F) + \epsilon \text{ for } n \text{ sufficiently large.} \end{aligned}$$

Thus $\limsup_n \mathbb{P}_n(F) \leq \mathbb{P}(F)$.

3 ⇔ 4 is obvious because one can simply take $G = F^c$.

3/4 ⇒ 5: because $\partial A = \bar{A} \setminus A^o$ where $\bar{A}$ is the closure of A and A^o is the interior of A.

5 ⇒ 1: This one is a little bit technical. Take any $f \in \mathcal{C}_b(\mathbb{X})$ and suppose there exists $B > 0$ such that $\|f\|_\infty \leq B$. Without loss of generality, one can take $B = 1$. For any probability measure $\mathbb{P}$, the cumulative distribution function $\mathbb{P}(f(X) \leq \cdot)$ has at most countably many jump points (discontinuities) because one can easily show an injection from these jump points to rationals, which are countable. Therefore one can do the following: given any precision threshold $\epsilon > 0$, divide $[-B, B]$ into M grids with endpoints $y_0 = -B \leq y_1 \leq y_2 \leq \cdots \leq y_M = B$ satisfying the following:

- $y_j - y_{j-1} \leq \epsilon$ for all $j = 1, \cdots, M$
- for each y_j, define the sets $A_j = \{x \in \mathbb{X} : y_{j-1} < f(x) \leq y_j\}$, so $\partial A_j = \{x \in \mathbb{X} : f(x) = y_j\}$ (a level set) and $\mathbb{P} A_j = 0$.

By 5, $\lim_n \mathbb{P}_n A_j = \mathbb{P} A_j$ which further implies

$$\lim_n \sum_{j=1}^M y_j \mathbb{P}_n A_j = \sum_{j=1}^M y_j \lim_n \mathbb{P}_n A_j = \sum_{j=1}^M y_j \mathbb{P}_{A_j}.$$

Note that by the above construction, we essentially recreate the way people define Lebesgue integration. Then we have

$$\begin{aligned}
&\left| \mathbb{P}(f) - \sum_{j=1}^M \int_{x \in A_j} y_j \mathbb{P}(\mathrm{d}x) \right| \\
= &\left| \sum_{j=1}^M \int_{x \in A_j} f(x) \mathbb{P}(\mathrm{d}x) - \sum_{j=1}^M \int_{x \in A_j} y_j \mathbb{P}(\mathrm{d}x) \right| \\
\leq &\sum_{j=1}^M \int_{x \in A_j} |f(x) - y_j| \mathbb{P}(\mathrm{d}x) \\
\leq &\, \epsilon \sum_{j=1}^M \int_{x \in A_j} \mathbb{P}(\mathrm{d}x) \\
= &\, \epsilon.
\end{aligned}$$

Similarly,

$$\left| \mathbb{P}_n(f) - \sum_{j=1}^M \int_{x \in A_j} y_j \mathbb{P}_n(\mathrm{d}x) \right| \leq \epsilon.$$

Combining the above: we have $|\mathbb{P}_n(f) - \mathbb{P}(f)| \leq 2\epsilon$. Because ϵ is chosen to be arbitrarily small, we have $\lim_n \mathbb{P}_n(f) = \mathbb{P}(f)$. □

Remark 31.

- Aad's proof in [?] is much better and easier to understand by showing $2 \Rightarrow 4$ instead.
- If one insists to show (*). One can choose a sequence k such that when k gets large,

  $$\|x - F\| \gg k^{-1}$$

  when $x \notin F'$. Then set

  $$f_k(x) = \frac{1}{1 + k\|x - F\|}$$

  or simply any other functions similar to it. For $x \in F$, $f_k(x) = 1$; for $x \notin F'$, $f_k(x)$ converges to 0; so we are left to check the Lipschitz-ness of f_k.

  $$\begin{aligned} |f_k(x) - f_k(y)| &= \frac{k|\|x - F\| - \|y - F\||}{1 + k\|x - F\| + k\|y - F\| + k^2\|x - F\|\|y - F\|} \\ &\leq \frac{k}{1 + k\|x - F\| + k\|y - F\| + k^2\|x - F\|\|y - F\|}\|x - y\| \end{aligned}$$

  and it is not hard to see that the numerator in the factor in front of $\|x - y\|$ will not blow up compared to k so f_k is $O(k)$-Lipschitz.
- There are many other equivalent statements in portmanteau lemma that is not covered in this note. You can also check more on page 6 of [?] for Aad's version.

7.2 Representation of probability distributions

7.2.1 Probability integral transform (PIF)

Uniform distribution on $[0, 1]$ is denoted as $\text{Unif}([0, 1])$ with CDF $F_U(u) = u$ if $U \sim \text{Unif}([0, 1])$.

For any continuous random variable X, denote its CDF as F_X, then $F_X(X) \sim \text{Unif}([0, 1])$. This is a very useful result for sampling random variables and the understanding of p-values in applied statistics. When a p-value is correctly constructed, it should be a survival function (i.e. 1 - CDF) under the true data distribution. So if the distribution of p-values is not uniform between 0 and 1, then it indicates there could be some systematic modeling bias. The mis-use of p-values or other types of statistical inference tool might be one major reason for the reproducibility crisis in biomedical research, now also happening in machine learning research. For a recent discussion on this topic, you can read Sander Greenland's essay and similar reproducibility/credibility issues in recent machine learning research, both written by statisticians.

7.2.2 Exponential distribution, Gamma distribution, and Beta distribution

$X \sim \text{Expo}(1)$: $f_X(x) = e^{-x}\mathbb{1}\{x \geq 0\}$. $X = -\log U$ for $U \sim \text{Unif}([0, 1])$.
Gamma distribution $G \sim \text{Gamma}(r)$ if $G = \sum_{j=1}^{r} X_j$ where $X_j \overset{i.i.d.}{\sim} \text{Expo}(1)$
Beta distribution $B \sim \text{Beta}(a, b)$ if $B = \frac{G_a}{G_a + G_b}$ where $G_a \sim \text{Gamma}(a)$ and $G_b \sim \text{Gamma}(b)$ and $G_a \perp\!\!\!\perp G_b$.

Lemma 32. *If $G_a \sim \text{Gamma}(a)$ and $G_b \sim \text{Gamma}(b)$ and $G_a \perp\!\!\!\perp G_b$, then*

$$G_a + G_b \perp\!\!\!\perp \frac{G_a}{G_a + G_b}.$$

Application of Beta distribution: a first encounter of Bayesian inference. Beta distribution is often used as the conjugate prior for Binomial likelihood. In a typical Bayesian analysis, one first sets up a probability model parameterized by some unknown parameter $\theta \in \Theta$ and views the unknown parameter θ as a random variable taking values in Θ, with a given probability measure Π_θ, which is called the "prior" or "prior belief" by Bayesians. Consider the following example: we observe $X \sim \text{Binom}(n, \theta)$, and we want to use (X, n) to estimate θ. An obvious choice is to use $\widehat{\theta} = X/n$, which is both the MLE and the method of moment (MoM) estimator of θ. Both MLE and MoM belong to the so-called frequentist method. As a Bayesian, however, we first need to specify the prior Π_θ with density π_θ and then use data (X, n) to update the prior to form the posterior, using Bayes' rule:

$$\pi_{\theta|X}(\theta|X = x) = \frac{f_{\theta,X}(\theta, x)}{f_X(x)} = \frac{f_{X|\theta}(x|\theta)\pi_\theta(\theta)}{f_X(x)}.$$

For the Binomial problem, with $\text{Beta}(\alpha, \beta)$ prior, we have

$$\begin{aligned}\pi_{\theta|X}(\theta|X = x) &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}\end{aligned}$$

which is again Beta distribution. This is why we call Beta prior is the conjugate prior for Binomial likelihood. Using conjugate prior can dramatically simplify the analysis because we have closed-form formula for many quantities of interest for Beta distribution. Just imagine what would have happened if we consider other types priors for θ, e.g. a normal distribution truncated between $[0, 1]$. When the posterior distribution is too complicated, we need to resort to other more sophisticated methods to obtain the posterior, e.g. by MCMC sampling or variational inference. Even though sampling techniques are quite mature, one can still use conjugate prior in some component of a big model to simplify the computation in practice.

8 Normal, chi-square, and Poisson

The simplest possible high dimensional model is the Gaussian Sequence Model or The Many Normal Mean model:

$$y_k = \mu_k + \epsilon Z_k, \quad Z_k \overset{\text{i.i.d.}}{\sim} N(0, 1).$$

Here each piece of data corresponds to an unknown parameter θ_k. More generally, $y_k = \mu_k + \epsilon p_k Z_k$ or $y_k = \alpha_k\mu_k + \epsilon Z_k$. Gaussian sequence model is a very rich model, and it is now the standard model mathematical statisticians use for theoretical investigation. Iain Johnstone has a whole book [**?**] with almost 500 pages dedicated to this simple model. Iain Johnstone is also the person bringing Random Matrix Theory (RMT) into modern high-dimensional statistics.

8.1 Multi-Variate Normal (MVN)

Definition 33. $Y = (Y_1, \cdots, Y_k)^\top \sim \text{MVN}_k$ if it is of the form:

$$Y_{k\times 1} = A_{k\times m}Z_{m\times 1} + \mu_{k\times 1},$$

here $Z = (Z_1, \cdots Z_m)^\top, Z_i \overset{\text{i.i.d}}{\sim} N(0, 1)$. Write $Y \sim \text{MVN}_k(\mu, V)$ if Y is a MVN of dimension k with mean μ and covariance matrix V, $V = AA^\top$

Remark 34. It is OK for a multivariate normal to be degenerate. If $k > m$, then let's just call this distribution the degenerate multivariate normal. In this case, the covariance matrix of Y has rank at most m and Y is concentrated on a proper subspace of $\mathbb{R}^k$. It also implies that Y does not have a density with respect to the Lebesgue measure.

But I have not encountered a very useful application of degenerate MVN. This might be due to the limitation of my knowledge though.

The benefit of the above definition is an explicit way of constructing multivariate distributions with dependent coordinates from independent product measures. This is a nice philosophy to bear in mind. Similar philosophy was used in copula.

The MVN has the following important properties.

Proposition 10. *Y is MVN(μ, V) if and only if $\forall t \in \mathbb{R}^k, t^\top Y$ is univariate normal.*

Proof. The proof of necessity is trivial. We then utilize MGF to prove the sufficiency.

$$\begin{aligned} \mathbb{E}(t^\top Y) &= t^\top \mu \quad \mathsf{var}(t^\top Y) = t^\top V t \\ M_{t^\top Y}(s) &= \exp\left((t^\top \mu)s + \frac{1}{2}(t^\top V t)s^2\right) = \mathbb{E}(e^{s(t^\top Y)}) \\ &= \mathbb{E}(e^{t'^\top Y}) \quad \text{(regard st as } t'\text{)} \end{aligned}$$

□

Proposition 11. *Assume $Y = \binom{Y_1}{Y_2} \sim MVN_k(\mu, V), V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}$, then $Y_1 \perp\!\!\!\perp Y_2$ if and only if $V_{12} = V_{21} = 0$.*

Note 35. The above proposition is 'within' a MVN: If two univariate normal random variables are uncorrelated, they are not necessarily independent if they are not jointly normal. Here are two counter examples:

1. $Z_1 \sim N(0,1)$, S is a Rademacher variable, $Z_2 = SZ_1$. Then $Z_1 \sim Z_2, \mathsf{cov}(Z_1, Z_2) = 0$, but $Z_1^2 = Z_2^2$ indicates they are dependent. (This example is not so ideal because $Z_1 = SZ_2$ exists only in a null set of $\mathbb{R}^2$)

2. A richer example: $Y_1 = Z_1, Y_2 = \rho Z_1 + \gamma Z_2, 0 \leq \rho \leq 1, \gamma = (1-\rho^2)^{1/2}, Z_1, Z_2 \overset{i.i.d}{\sim} N(0,1), W_1 = SY_1, W_2 = Y_2$. Here $\mathsf{cov}(W_1, W_2) = 0$ but W_1 and W_2 are dependent. (proof hint: $\mathbb{E}(W_2^2 - 1 | W_1^2)$ is a function w.r.t. W_1^2)

Proposition 12 (Rotation invariance of isotropic Gaussian)**.** *$Z \sim MVN_k(0, Id_{k\times k})$, $\Gamma_{k\times k}$ is an orthogonal matrix (i.e. $\Gamma \in \mathcal{O}(k)$, the orthogonal group of dimension k, or $\Gamma\Gamma^\top = Id$), then $\Gamma Z \sim MVN_k(0, Id_{k\times k})$. (Spherically symmetric distribution)*

Proof. Compare the mean vector and covariance matrix. □

Remark 36. This not only holds for multivariate Gaussian, but also holds for spherically symmetric distributions. But if one proves some results under Gaussian using rotation invariance, it generally cannot be generalized to sub-gaussians (see later in the lecture).

Corollary 3. *$Z_1, Z_2, \cdots Z_n \overset{i.i.d}{\sim} N(0,1), \bar{Z} = \frac{1}{n}\sum_{i=1}^n Z_i, s^2 = \frac{1}{n-1}\sum_{i=1}^n (Z_i - \bar{Z})^2$. Then we have $\bar{Z} \perp\!\!\!\perp s^2, \bar{Z} \sim N(0, 1/n), s^2 \sim \frac{1}{n-1}\chi^2_{n-1}$*

Note 37. This above result only holds for normal distribution.

8.2 Lindeberg's telescoping sum technique

We first state the Lindeberg version of the CLT for sum of independent but not necessarily identically distributed random variables:

Theorem 38. *$X_1, \cdots, X_n$ independent, with $\mathbb{E}[X_k] = \mu_k, \mathsf{var}[X_k] = \sigma_k^2$ and $\mathbb{E}|X_k|^3 < \infty$. Define $\mu = \sum_{k=1}^n \mu_k$ and $\sigma^2 = \sum_{k=1}^n \sigma_k^2$ and $Z \sim N(\mu, \sigma^2)$. Then for any $f \in \mathcal{C}_b^3(\mathbb{R})$, we have*

$$\left|\mathbb{E}f\left(\sum_{k=1}^n X_k\right) - \mathbb{E}f(Z)\right| \lesssim ||f'''||_\infty \sum_{k=1}^n \mathbb{E}|X_k|^3. \tag{1}$$

Proof. The proof is the so-called Lindeberg's telescoping sum technique. Create independent $Z_1, \cdots, Z_n$ independent from $X_1, \cdots, X_n$ but $Z_k \sim N(\mu_k, \sigma_k^2)$ for $k = 1, \cdots, n$. Define

$$Y_k = X_1 + \cdots + X_{k-1} + Z_{k+1} + \cdots + Z_n.$$

Then we immediately have the following identity: $Y_k + Z_k = Y_{k-1} + X_{k-1}$. Note that $Y_k \perp\!\!\!\perp X_k \perp\!\!\!\perp Z_k$. Then we can decompose the LHS of eq. (1) as follows:

$$\begin{aligned}
&\left|\mathbb{E}f\left(\sum_{k=1}^n X_k\right) - \mathbb{E}f(Z)\right| \\
=\ &|\mathbb{E}f(Y_n + X_n) - f(Y_n + Z_n) + f(Y_{n-1} + X_{n-1}) - f(Y_{n-1} + Z_{n-1}) + \cdots| \\
\leq\ &\sum_{k=1}^n |\mathbb{E}f(Y_k + X_k) - f(Y_k + Z_k)|.
\end{aligned}$$

We analyze each summand separately by Taylor theorem:

$$\begin{aligned}
&|\mathbb{E}f(Y_k + X_k) - f(Y_k + Z_k)| \\
=\ &\left|\frac{1}{2}\mathbb{E}\int_0^1 f'''(Y_k + tX_k)t^2 X_k^3 \mathrm{d}t - \frac{1}{2}\mathbb{E}\int_0^1 f'''(Y_k + tZ_k)t^2 Z_k^3 \mathrm{d}t\right| \\
\leq\ &\frac{1}{2}\left(\mathbb{E}\int_0^1 |f'''(Y_k + tX_k)|t^2|X_k|^3 \mathrm{d}t + \mathbb{E}\int_0^1 |f'''(Y_k + tZ_k)|t^2|Z_k|^3 \mathrm{d}t\right) \\
\leq\ &\frac{\|f'''\|_\infty}{2}(\mathbb{E}|X_k|^3 + \mathbb{E}|Z_k|^3)\int_0^1 t^2 \mathrm{d}t \\
\lesssim\ &\|f'''\|_\infty \mathbb{E}|X_k|^3.
\end{aligned}$$

Finally we add them all up. □

An immediate corollary of Theorem 38 is the following:

Corollary 4. *$X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} (0, 1)$ (which denotes with mean 0 and variance 1), then for any $f \in \mathcal{C}_b^3(\mathbb{R})$, we have*

$$\left|\mathbb{E}f\left(\frac{1}{\sqrt{n}}\sum_{k=1}^n X_k\right) - \mathbb{E}f(Z)\right| \lesssim ||f'''||_\infty \frac{\mathbb{E}|X|^3}{\sqrt{n}}$$

8.3 Stein's Method

Lemma 39 (Stein's Lemma)**.** *$f\colon \mathbb{R} \to \mathbb{R}$ is a differentiable function, $Z \sim N(0,1), \mathbb{E}[f'(Z) - Zf(Z)] = 0$, which is also called 'Stein's Identity'.*

Proof.

$$\begin{aligned}\mathbb{E}(f'(Z)) &= \int_{-\infty}^{\infty} f'(t)\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}d(f(f))\\ &= \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}f(t)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{2\pi}}de^{-\frac{t^2}{2}}\\ &= -\int_{-\infty}^{\infty} f(t)\frac{1}{\sqrt{2\pi}}(-t)e^{-\frac{t^2}{2}}dt = \mathbb{E}(Zf(Z))\end{aligned}$$

□

Stein's Lemma offers a possible strategy to measure how far a random variable X to a standard normal random variable Z. Intuition indicates: $\mathbb{E}(f'(X) - Xf(X)) \approx 0$ if $X \approx Z$. In order to prove $X \Rightarrow Z$, we can prove $\sup\limits_{h\in Lip} \mathbb{E}h(X) - \mathbb{E}h(Z) \to 0$ (by portmanteau lemma). Here $Lip \coloneqq \{f, |f(x) - f(x')| \le L||x - x'||\}$). Firstly, we introduce Stein's equation, which is important in the proof of a later theorem (Theorem 41).

Lemma 40 (Stein's equation)**.** *$h\colon \mathbb{R} \to \mathbb{R}$ is Lipschitz continuous function. $X \sim P_X, Z \sim N(0,1)$ Then $\mathbb{E}(h(X) - h(Z)) = \mathbb{E}[f_h'(X) - Xf_h(X)]$. Here $f_h(x) = e^{\frac{x^2}{2}}\int_{-\infty}^{x} g(t)e^{\frac{-t^2}{2}}dt, g(x) = h(x) - \mathbb{E}(h(Z))$.*

A remark is: any Lipschitz continuous function has essentially bounded (bounded with Lebesgue measure 1) derivative almost everywhere.

The result is easy to verify when we notice that $f_h'(x) = g(x) + xf_h(x)$, which is a simple ODE. Besides, we have $||f_h''||_\infty \lesssim ||h'||_\infty, ||f_h'||_\infty \lesssim ||h'||_\infty, ||f_h||_\infty \lesssim ||h'||_\infty$. The proof of these bounds can refer to [**?**], page 37-40. Another useful resource is Sourav Chatterjee's lecture notes. Then we utilize this equation to bound $\sup\limits_{h\in Lip}[\mathbb{E}h(X) - \mathbb{E}h(Z)]$.

Theorem 41. *$X_1, X_2 \cdots X_n \overset{i.i.d}{\sim} P_X, \mathbb{E}(X) = 0, \mathbb{E}(X^2) = 1, \mathbb{E}(|X|^3) < \infty, \widetilde{X} = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, then $\sup\limits_{h\in Lip}[\mathbb{E}h(\widetilde{X}) - \mathbb{E}h(Z)] \lesssim \frac{\mathbb{E}|X_i|^3}{\sqrt{n}}$.*

Proof. From Stein's equation, we have:

$$\sup_{h\in Lip}[\mathbb{E}h(\widetilde{X}) - \mathbb{E}h(Z)] = \sup_{h\in Lip}\mathbb{E}[f_h'(\widetilde{X}) - \widetilde{X}f_h(\widetilde{X})]$$

Let $X' = \widetilde{X} - \frac{1}{\sqrt{n}}X_1 = \frac{1}{\sqrt{n}}\sum_{i=2}^n X_i$. Obviously $X' \perp\!\!\!\perp X_1$. Let $f = f_h$.

$$\begin{aligned}
\mathbb{E}[\widetilde{X}f(\widetilde{X})] &= \mathbb{E}[\sqrt{n}X_1 f(\widetilde{X})] && (\forall f \text{differentiable})\\
&= \mathbb{E}[\sqrt{n}X_1(f(X') + \frac{1}{\sqrt{n}}X_1 f'(X'))] + R_1 && (\text{Taylor expansion})\\
&= \mathbb{E}[f'(X')] + R_1 = \mathbb{E}[f'(\widetilde{X})] + R_1 + R_2\\
|R_1| &= |\mathbb{E}\left[\frac{X_1^3}{\sqrt{n}} f''(X'')\right]| && (X' \wedge \widetilde{X}) \leq X'' \leq (X' \vee \widetilde{X})\\
&\leq \mathbb{E}\frac{|X_i|^3}{\sqrt{n}}|f''(X'')| \leq \mathbb{E}\frac{|X_i|^3}{\sqrt{n}}||f''||_\infty \lesssim \frac{\mathbb{E}|X_i|^3}{\sqrt{n}}\\
|R_2| &= |\mathbb{E}(f'(X') - f'(\widetilde{X})| = |\mathbb{E}(f'(X') - f'(X' + \frac{X_1}{\sqrt{n}})|\\
&= |\mathbb{E}(f'(X') - f'(X') - \frac{X_1}{\sqrt{n}}f''(X'''))|\\
&\leq \mathbb{E}|\frac{X_1}{\sqrt{n}}f''(X''')| \leq \frac{\mathbb{E}|X_1|}{\sqrt{n}}||f''||_\infty \lesssim \frac{\mathbb{E}|X_i|^3}{\sqrt{n}}
\end{aligned}$$

In conclusion : $\sup_{h\in Lip}[\mathbb{E}h(\widetilde{X}) - \mathbb{E}h(Z)] = \sup_{h\in Lip}\mathbb{E}[f_h'(\widetilde{X}) - \widetilde{X}f_h(\widetilde{X})] \lesssim \sup_{h\in Lip}\frac{\mathbb{E}|X_i|^3||f''||_\infty}{\sqrt{n}} \lesssim \frac{\mathbb{E}|X_i|^3}{\sqrt{n}}$ □

Remark 42. Stein's identity, Stein's equation, and Stein's method are not only useful in the context of proving normal approximations. In fact, many well known distributions have their own version of Stein's identity, Stein's equation and Stein's method: Poisson distribution in [**?**], Exponential distribution in [**?**], Ising model in [**?**], Diffusion processes in [**?**], Birth-death processes in [**?**], Negative binomial distributions in [**?**], and Markov chains in [**?**]. For a high-level review, you can also look at [**? ? ? ?**]. If you are interested, you can check out the obituary of Charles Stein by Stanford University on his legacy after his passing.

8.4 Gaussian Concentration

Proposition 13 (Chernoff's bound)**.** *Random variable $Z \sim N(0, \sigma^2)$, then we have $P(Z - \mu > t) \leq e^{-\frac{t^2}{2\sigma^2}}$*

Proof.

$$\begin{aligned}
P(Z - \mu > t) &= P(e^{\lambda(Z-\mu)} > e^{\lambda t}) && \forall \lambda > 0\\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(Z-\mu)}] && (\text{Markov Inequality})\\
&= e^{-\lambda(t+\mu)}e^{\lambda\mu + \lambda^2\sigma^2/2} && (*)
\end{aligned}$$

$\lambda_0 = \frac{t}{\sigma^2}$ minimize the above formula (*). Replace λ by λ_0 in (*) we have: $P(Z - \mu > t) \leq e^{-\frac{t^2}{2\sigma^2}}$. □

The tail probability (i.e. $P(|X - \mu| > t)$) dropping at an exponential rate is an ideal property for a distribution. Besides Gaussian, a wider range of distributions such as all bounded distributions has this property. They are called 'Sub-Gaussian' distributions.

Definition 43. Random variable X with mean μ is sub-Gaussian if $\exists \sigma^2 > 0$, s.t. $\mathbb{E}e^{t(X-\mu)} \leq e^{\sigma^2 t^2/2}$. It deduces $P(|X-\mu| > t) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.

Proposition 14. *If $X \sim$ sub-Gaussian$(0, \sigma^2)$ (here σ is not necessarily the standard deviation of X), then $\forall k \geq 1, \mathbb{E}|X|^k \leq (2\sigma^2)^{\frac{k}{2}} k\Gamma(\frac{k}{2})$. [For comparison, $\mathbb{E}|Z|^k = \frac{1}{\sqrt{\pi}}(2\sigma^2)^{\frac{k}{2}}\Gamma(\frac{k}{2}+1) = \frac{1}{\sqrt{\pi}}(2\sigma^2)^{\frac{k}{2}}\frac{k}{2}\Gamma(\frac{k}{2})$ if $Z \sim N(0, \sigma^2)$].*

Proof. Suppose $X \sim$sub-Gaussian$(0, \sigma^2)$. We know from the previous lecture that:

$$\begin{aligned}
\mathbb{E}|X|^k &= \int_0^\infty P(|X|^k > t)dt = \int_0^\infty P(|X| > t^{\frac{1}{k}})dt \\
&\leq 2\int_0^\infty e^{-\frac{t^{2/k}}{2\sigma^2}} dt && \text{(defn. of sub-Gaussian)} \\
&= 2\int_0^\infty e^{-u}\frac{k}{2}(2\sigma^2 u)^{k/2-1} du 2\sigma^2 && (u = \frac{t^{2/k}}{2\sigma^2}) \\
&= k(2\sigma^2)^{\frac{k}{2}} \int_0^\infty e^{-u} u^{\frac{k}{2}-1} du \\
&= (2\sigma^2)^{\frac{k}{2}} k\Gamma(\frac{k}{2})
\end{aligned}$$

□

8.5 Darmois-Skitovič lemma

Lemma 44 (Darmois-Skitovič lemma)**.** *Let $X_1, ..., X_n$ be independent random variables. Given two linear combinations $L_1 = a_1X_1 + ... + a_nX_n$ and $L_2 = b_1X_1 + ... + b_nX_n$, if L_1 and L_2 are independent, then if $a_ib_i \neq 0$ then X_i is normally distributed.*

Proof. Consider the characteristic function of the random vector (L_1, L_2):

$$\begin{aligned}
c_{L_1,L_2}(t_1, t_2) &= \mathbb{E}\left[e^{i(t_1,t_2)^\top (L_1,L_2)}\right] \\
&= \mathbb{E}\left[e^{it_1(a_1X_1+...+a_nX_n)+it_2(b_1X_1+...+b_nX_n)}\right] \\
&= \mathbb{E}\left[e^{i(t_1a_1+t_2b_1)X_1+...+i(t_1a_n+t_2b_n)X_n}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{i(t_1a_i+t_2b_i)X_i}\right] \equiv \prod_{i=1}^n c_{X_i}(t_1a_i + t_2b_i)
\end{aligned}$$

where the last line follows from independencies among $X_1, \cdots, X_n$. Since $L_1 \perp\!\!\!\perp L_2$, we also have

$$c_{L_1,L_2}(t_1, t_2) = c_{L_1}(t_1)c_{L_2}(t_2).$$

Combining the above two, we have:

$$\sum_{i=1}^n \log c_{X_i}(t_1a_i + t_2b_i) = \log c_{L_1}(t_1) + \log c_{L_2}(t_2)$$

which says that the function on the LHS as a function of t_1 and t_2 is separable, which implies that each summand is a polynomials with degree at most n. The only distribution with log characteristic function (cumulant generating function) being a polynomial is Gaussian (a proof can be found in [**?** , Chapter 2.5]). □

A corollary of Lemma 44 is the more famous Kac-Bernstein theorem:

Theorem 45 (Kac-Bernstein theorem)**.** *Given two independent random variables X_1, X_2. If $L_1 = X_1 + X_2$ and $L_2 = X_1 - X_2$ are independent, then X_1 and X_2 must be normally distributed.*

An interesting implication of Darmois-Skitovič lemma in statistical application is the following:

> Suppose there are two random variables X and Y and you know the data generating mechanism is either (1) $Y = \beta X + \epsilon_Y$ where ϵ_Y is a mean zero noise independent of X or (2) $X = \alpha Y + \epsilon_X$ where ϵ_X is a mean zero noise independent of Y. Can you tell the difference between Model (1) and Model (2)? A question that people often ask is whether it is possible to tell whether Model (1) or (2) is the reality, given n i.i.d. pairs $(X_i, Y_i), i = 1, \cdots, n$. A natural idea is to run regressions in the direction of (1) and in the direction of (2) and compare the results. Interestingly, based on Darmois-Skitovič lemma, if Model (1) is true, and X, ϵ_Y are both normally distributed, then Model (2) also holds with Y, ϵ_X normally distributed. So it is impossible to distinguish between Model (1) and (2). However, once either the covariate or the noise is non-Gaussian, running regressions in opposite directions can differentiate (1) from (2) or vice versa. This is the basis of a recent important work [**?**] in a subfield of statistics, machine learning, computer science, and philosophy called "causal discovery". We can first check that there always exist such α and ϵ_Y. Observe that $\mathsf{var}[Y] = \beta^2\mathsf{var}[X] + \mathsf{var}[\epsilon_Y]$ here $\mathsf{var}[X], \beta, \mathsf{var}[\epsilon_Y]$ are all given. Then $\mathsf{var}[X] = \alpha^2\mathsf{var}[Y] + \mathsf{var}[\epsilon_X] = \alpha^2\beta^2\mathsf{var}[X] + \alpha^2\mathsf{var}[\epsilon_Y] + \mathsf{var}[\epsilon_X] \Rightarrow (1 - \alpha^2\beta^2)\mathsf{var}[X] = \alpha^2\mathsf{var}[\epsilon_Y] + \mathsf{var}[\epsilon_X]$ so
>
> $$\mathsf{var}[\epsilon_X] = (1 - \alpha^2\beta^2)\mathsf{var}[X] - \alpha^2\mathsf{var}[\epsilon_Y] > 0.$$
>
> Furthermore, $\mathbb{E}[Y(X - \mathbb{E}[X|Y])] = \mathbb{E}[Y(X - \alpha Y)] = 0$ which gives us
>
> $$\begin{aligned}\mathbb{E}[Y(X - \alpha Y)] &= \mathbb{E}[YX] - \alpha\mathbb{E}[Y^2] = \beta^2\mathbb{E}[X^2] - \alpha\beta^2\mathsf{var}[X] - \alpha\mathsf{var}[\epsilon_Y] \\ &= \beta^2(1-\alpha)\mathsf{var}[X] - \alpha\mathsf{var}[\epsilon_Y] = 0 \Rightarrow \frac{\alpha}{1-\alpha} = \frac{\beta^2\mathsf{var}[X]}{\mathsf{var}[\epsilon_Y]}.\end{aligned}$$
>
> From here, it is easy to see that we can take $\epsilon_X = X - \alpha Y$ which is a Gaussian and hence $\epsilon_X \perp\!\!\!\perp Y$. To prove that this is necessary, we need to use Lemma 44 and consider the following linear combination:
>
> $$\begin{aligned}Y &= \beta X + \epsilon_Y \\ \epsilon_X &= (1 - \alpha\beta)X - \alpha\epsilon_Y\end{aligned}$$

8.6 Gaussian graphical models

We have seen that marginal independency between different (sets of) variables in a multivariate Gaussian distribution $\mathrm{MVN}(0, \Sigma)$ can be read off by whether the corresponding elements in covariance matrix Σ equal zero. How about conditional independency? Actually, it can be studied via "Gaussian graphical models" [**?**].

Here we only consider non-degenerate multivariate Gaussians i.e. $\Sigma \succ 0$ so $\Omega = \Sigma^{-1}$ is well-defined. Given $Z \sim \mathrm{MVN}(0, \Sigma)$ how to determine if $Z_I \perp\!\!\!\perp Z_J | Z_K$ for $I, J, K \subseteq \{1, \cdots, n\} \equiv V$ and $I \cap J = \emptyset$? In fact, one could show the following:

$$Z_i \perp\!\!\!\perp Z_j | Z_{V \setminus \{i,j\}} \Leftrightarrow \Omega_{i,j} = 0.$$

You should try to prove this as an exercise. Given this result, can we devise a straightforward algorithm of deciding if given index subsets I, J, K such that $I \cap J = \emptyset$, to decide if $Z_I \perp\!\!\!\perp Z_J | Z_K$? One natural idea is to attach a probabilistic model to a graph as follows: Given Ω, each dimension corresponds to a vertex of an undirected acyclic graph $\mathcal{G}(V, E)$, where the edge set E excludes edge $i - j$ whenever $\Omega_{i,j} = 0$. Then if $Z_I \perp\!\!\!\perp Z_J | Z_K$ if every path between I and J intersects K.

8.7 Gaussian processes

Now that we have understood finite-dimensional multivariate normals, we briefly discuss its infinite-dimensional extension: Gaussian Processes (GP). Gaussian processes are random functions $f : \mathbb{X} \to \mathbb{R}$ where $\mathbb{X}$ is a normed metric space, say $\mathbb{R}$. f is distributed as Gaussian processes if all finite-dimensional discretization $\{f(x_1), \cdots, f(x_n)\}$ is distributed as a multivariate Gaussian, for any n. We will talk more about GP when it comes to Chapter 3 & Bayesian nonparametrics (if we got time or if you read the papers).

8.8 Chi-squared distribution

Central chi-squared $\chi^2_{d,0} \sim \sum_{i=1}^d Z_i^2$ where $Z_i \overset{iid}{\sim} N(0,1)$;
Noncentral chi-squared $\chi^2_{d,\xi} \sim \sum_{i=1}^d (Z_i + \mu_i)^2$ with $\xi = \sum_{i=1}^d \mu_i^2$.
Density function:

$$X \sim \chi^2_{d,0} : f_X(x) = \frac{2}{\Gamma(d/2)} e^{-x/2} \left(\frac{x}{2}\right)^{d/2-1} \mathbb{1}\{x \geq 0\}$$

where $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x$ is the Gamma function which always satisfies $\Gamma(z+1) = z\Gamma(z)$. When z is an integer, $\Gamma(z) = (z-1)!$.

Theorem 46 (Chi-squared tail bound)**.**

$$\mathbb{P}\left[|\chi^2_d - d| > t\right] \leq 2 \exp\left(-\frac{t^2}{8d}\right)$$

There are many useful results on chi-squared distributions. You can refer to [**?**] when needed.

Related to chi-squared distribution is the sub-exponential distribution, defined as follows:

Definition 47. A random variable X is sub-exponential(μ, σ, α) if $\mathbb{E}X = \mu$ and

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\lambda^2\sigma^2}{2}} \text{ for } |\lambda| < \frac{1}{\alpha}$$

The main difference between sub-Gaussian and sub-exponential is the above upper bound on the MGF does not hold for all $\mathbb{R}$ for sub-exponentials. So sub-exponential can have thicker tail than sub-Gaussian.

Theorem 48. *If $X \sim$ sub-exponential(μ, σ, α), then*

$$\mathbb{P}(|X-\mu| > t) \le 2\exp\left\{-\frac{1}{2}\left(\frac{t}{\alpha} \wedge \frac{t^2}{\sigma^2}\right)\right\}$$

Proof. Everything follows the proof of the Chernoff bound for normal distribution, except when $t > \sigma^2/\alpha$, the unconstrained minimizer $\lambda = t/\sigma^2 > 1/\alpha$. In that range, the minimizer should be $1/\alpha$ instead, so the exponent becomes $-t/\alpha + \sigma^2/2\alpha^2 \le -t/(2\alpha)$. □

A sufficient condition for a random variable to be sub-exponential is the following Bernstein type bound on higher order moments:

Theorem 49. *For X with mean μ and variance σ^2, if $|\mathbb{E}(X-\mu)^k| \le \frac{k!}{2}\sigma^2 b^{k-2}$ for $k = 3, 4, \cdots$, then*

$$\mathbb{E}e^{\lambda(X-\mu)} \le \exp\left\{\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right\}, \quad |\lambda| < \frac{1}{b}.$$

Proof. Taylor expansion:

$$\begin{aligned}
\mathbb{E}e^{\lambda(X-\mu)} &\le 1 + \mathbb{E}\lambda(X-\mu) + \frac{\lambda^2}{2}\mathbb{E}(X-\mu)^2 + \sum_{k=3}^{\infty}\lambda^k \frac{|\mathbb{E}(X-\mu)^k|}{k!} \\
&\le 1 + \frac{\lambda^2\sigma^2}{2} + \frac{\lambda^2\sigma^2}{2}\sum_{k=3}^{\infty}(|\lambda|b)^{k-2} \\
&= 1 + \frac{\lambda^2\sigma^2}{2}(1 + (|\lambda|b) + (|\lambda|b)^2 + \cdots) \\
&= 1 + \frac{\lambda^2\sigma^2}{2} 1/(1-|\lambda|b) \text{ if } |\lambda| < 1/b \\
&\le e^{\frac{\lambda^2\sigma^2}{2(1-|\lambda|b)}}.
\end{aligned}$$

□

One can easily check $\chi_d^2 \sim$ sub-exponential$(d, 2\sqrt{d}, 4)$ by looking at its MGF. So the chi square tail bound follows from Theorem 48.

8.9 Poisson distribution

For $X \sim$ Poisson(λ), $\Pr(X = x) \equiv f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}$ for $x = 0, 1, \cdots$.

Lemma 50. $\chi^2_{d,\xi} \sim \chi^2_{d+2K,0}$ *with $K \sim$ Poisson$(\xi/2)$.*

Lemma 51. *Stein's identity for Poisson: $\mathbb{E}[Kf(K) - \lambda f(K+1)] = 0$ when $K \sim$ Poisson(λ).*

I won't prove the above two results. Check them on your own.

Poisson approximation:

Normal approximation is supported by CLT. Is there any similar result for Poisson approximation? Actually, there is a so-called "law of rare events/law of small numbers":

Theorem 52. *$X_1, \cdots, X_n \overset{i.i.d.}{\sim}$ Bernoulli(p), if as $n \to \infty$, $p \to 0$ but $np \to \lambda$ for some $\lambda > 0$, then $X_1 + \cdots + X_n \xrightarrow{d}$ Poisson(λ).*

Similar to Stein's method, Poisson approximation is also often used in research on random graphs. For examples, for Erdös-Renyi graph $G(n, p)$ with n vertices and edges are independent Bernoulli with probability p, then the total number of edges will be approximately a Poisson distributed random variable. Actually, even the total number of triangles, where the summands are no longer independent, is also approximately Poisson. This can be proved by using Stein's method for Poisson approximations.

Poissonisation/Poissonization trick:

When you are dealing with a multinomial sample of K classes, Multinomial$(n, p_1, \cdots, p_K)$, the number of elements falling in each class is denoted as n_k for $k = 1, \cdots, K$. Each $n_k \sim$ Binomial(n, p_k) but different n_k's are not independent because of the equality constraint $n_1 + \cdots + n_k = n$. It comes in handy if you can somehow treat them as being independent. Poissonisation trick is the tool that you will use in this context. Instead of n being fixed, you can consider a random variable $N \sim$ Poisson(n). Define N_k similarly, then N_k are independent and identically distributed as Poisson(np_k) for $k = 1, \cdots, K$.

Modeling with Poisson:

Poisson can be a useful distribution for modeling count data $(0, 1, 2, 3, \cdots)$. But there is one important caveat – Poisson has equal mean and variance – a very restrictive feature. When modeling count data using Poisson, you always need to check if the sample mean and variance are close (in regression, there is no need to be close marginally, but still need to be close conditionally on the covariates). If Poisson is not a good modeling choice, then negative binomial distribution, Poisson mixtures or others can be tried.