inequalities, uniform bounds, convergence

- the multivariable pdf is such that $P(Z \leq z) = \int \int_A f(x,y)dxdy$. (get cdf and take partial derivatives) → this is also the joint probability if $f(x,y)$ is a joint pdf.
- i.i.d means independent identically distributed. note that this is just an assumption and that isn't always true. whether or not it is a good assumption depends on the problem.
- we should differentiate between a random variable and the parameters. $P(x;\mu,\sigma)$ really specifies a class of distributions, and when we give a value to the parameters it gives us a distribution.
- Gaussian tail inequality: let X ~ $N(0,1)$. Then $P(|X| > \epsilon) \leq \frac{2e^{-\epsilon^2/2}}{\epsilon}$. the proof is pretty simple if you write down the definition of N's pdf, and multiply and divide by x. proof.
- hoeffding's inequality: Let $Y_1 ... Y_n$ be i.i.d observations such that $E(Y_i) = \mu$ and $a \leq Y_i \leq b$. then for any $\epsilon > 0$, $P(|\overline{Y}_n - \mu| \geq \epsilon) \leq 2e^{-2n\epsilon^2/(b-a)^2}$. for the bernoulli case, $a = 0, b = 1$.
- the proof of the above is a bit long and well written here. one idea it uses that comes up often is chernoff's method which says that $P(X > \epsilon) \leq inf_{t \geq 0}\, e^{-t\epsilon}E(e^{tX})$. proof is simple: $P(X > \epsilon)$ → $P(e^X > e^\epsilon)$ → $P(e^{tX} > e^{t\epsilon}) \leq e^{-t\epsilon}E(e^{tX})$ (last step using markov's inequality)
- Kullback Leibler distance: is a measure of the "distance" between two densities p and q.
  $D(p,q) = \int p(x) log \frac{p(x)}{q(x)} dx$. Note that $D(p,p) = 0$. Using Jensen's inequality and writing $-D(p,q)$ as $E(log \frac{q(X)}{p(X)})$, we can show that $D \geq 0$.
  Stare at the definition above and try to understand why if the densities are dissimilar, the number will be bigger.
  The discrete version of this has integral replaced by sums, as usual.
- Using Jensen's inequality, we can also easily prove: Let $X_1 ... X_n$ be random variables. (needn't be iid). Suppose there exists $\sigma > 0$ such that $E(e^{tX_i}) \leq e^{t^2\sigma^2/2}$ for all $t > 0$. then $E(max\ X_i) \leq \sigma\sqrt{2log(n)}$.
- for sequences, $a_n = o(1)$ means that a_n → 0 as $n \to \infty$. $a_n = o(b_n)$ means that $\frac{a_n}{b_n} = o(1)$.
- for sequences, $a_n = O(1)$ means that for all large n (beyond some point), $|a_n| \leq C$. $a_n = O(b_n)$ means that $\frac{a_n}{b_n} = O(1)$.
- we write a_n ~ b_n if $\frac{a_n}{b_n}$ and $\frac{b_n}{a_n}$ are eventually bounded.
- in the probabilistic versions of these, we don't bound the values of the random variables but we instead bound their probabilities.
- so $Y_n = o_P(1)$ means that for every $\epsilon > 0$, $P(|Y_n| > \epsilon)$ → 0. and, $Y_n = o_P(a_n)$ means that $Y_n/a_n = o_P(1)$.
- $Y_n = O_P(1)$ means that for every $\epsilon > 0$, there is a $C > 0$ such that $P(|Y_n| > C) \leq \epsilon$. and, $Y_n = O_P(a_n)$ means that $Y_n/a_n = O_P(1)$.
- one can show easily that:

$$O_P(1)o_P(1) = o_P(1)$$
$$O_P(1)O_P(1) = O_P(1)$$
$$o_P(1) + O_P(1) = O_P(1)$$
$$O_P(a_n)o_P(b_n) = o_P(a_n b_n)$$
$$O_P(a_n)O_P(b_n) = O_P(a_n b_n)$$

- In machine learning, we are interested in the following question:
  If we have an estimator that works by sampling, a natural question is what's the probability that the difference in cdf computed empirically vs the real cdf is within some (small) value?
  Using Hoeffding's the answer is that as $n \to \infty$, the error goes to 0. which basically gives us a theoretical backing for saying if we sample enough, the likelihood that are error is bigger than a very small number is very low!
- Remember, pointwise convergence means that *for each* x and ϵ you can find an N such that (bla bla bla). Here the N is allowed to depend *both* on x and ϵ. In uniform convergence the requirement is strengthened. Here *for each* ϵ you need to be able to find an N such that (bla bla bla) for *all* x in the domain of the function. In other words N can depend on ϵ but not on x.
- we want the above cdf estimator $P(X < t)$ to be good for all t. So we want $P(P(\overline{X} - t) - P(X < t))$ to converge uniformly, not point wise. This is an example of how in statistics and machin learning, we are really interested in uniform convergence and not just point wise convergence.
- let's say A is some event (subset of sample space). Lets say we are trying to estimate $P(A)$ by $P_n(A) = \frac{1}{n}\sum I(X_i \epsilon A)$. We know from Hoeffding's, $P(P_n(A) - P(A) > \epsilon) < 2e^{-2n\epsilon^2}$. Guaranteeing that its small for one thing at a time is not guaranteeing that its always small. that means to construct a good estimator, we need $P(sup_{A \epsilon Q}|P_n(A) - P(A)| > \epsilon) <$ something small. this is called a uniform bound (uniform over a set Q)
- now if set Q contains a finite number of events $A_1, ...A_n$. Let B_j be the event that $|P_n(A_j) - P(A_j)| > \epsilon$. Then, the only way the max difference will exceed ϵ is if one of them exceeds ϵ. That is $P(sup_{A \epsilon Q}|P_n(A) - P(A)| > \epsilon) = P(\cup B_j) \leq 2Ne^{-2n\epsilon^2}$. So if we have a finite set of events, then we can bound the worst case like this.
- but what if we don't have a finite set? introducing shattering... Let A be a class of sets. Let $F = x_1 ... x_n$ be a finite set. Let G be a subset of F. Say that A picks out G if $a \cap F = G$ for some $a \, \epsilon \, A$.
- larry wasserman states some theorem about inequalities, shattering coefficient and VC dimension that he doesn't prove. We won't be stating that here, it will be done in andrew ng's machine learning class.
- convergence is concerned with answering the question: what happens as we gather more and more data? what bounds can we have on our probabilities?
- Let $X_1, ...X_n$ be i.i.d F. A statistic is any function $T_n = g(X_1, ..X_n)$. eg. sample mean. Notice that T_n is a sequence as n grows.

- some definitions:

1. X_n **converges almost surely to** X, written $X_n \overset{a.s.}{\to} X$, if, for every $\epsilon > 0$,

$$\mathbb{P}\Big(\lim_{n\to\infty} |X_n - X| < \epsilon\Big) = 1. \tag{1}$$

X_n **converges almost surely to a constant** c, written $X_n \overset{a.s.}{\to} c$, if, for every $\epsilon > 0$,

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = c\Big) = 1. \tag{2}$$

2. X_n **converges to** X **in probability**, written $X_n \overset{P}{\to} X$, if, for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \tag{3}$$

as $n \to \infty$. **In other words,** $X_n - X = o_P(1)$**.**
X_n **converges to** c **in probability**, written $X_n \overset{P}{\to} c$, if, for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - c| > \epsilon) \to 0 \tag{4}$$

as $n \to \infty$. In other words, $X_n - c = o_P(1)$.

3. X_n **converges to** X **in quadratic mean** (also called convergence in L_2), written $X_n \overset{qm}{\to} X$, if

$$\mathbb{E}(X_n - X)^2 \to 0 \tag{5}$$

as $n \to \infty$.
X_n **converges to** c **in quadratic mean**, written $X_n \overset{qm}{\to} c$, if

$$\mathbb{E}(X_n - c)^2 \to 0 \tag{6}$$

as $n \to \infty$.

4. X_n **converges to** X **in distribution**, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{7}$$

at all t for which F is continuous.
X_n **converges to** c **in distribution**, written $X_n \rightsquigarrow c$, if

$$\lim_{n\to\infty} F_n(t) = \delta_c(t) \tag{8}$$

at all $t \neq c$ where $\delta_c(t) = 0$ if $t < c$ and $\delta_c(t) = 1$ if $t \geq c$.

A couple of proofs of important implications:

$$\begin{array}{ccc} \text{q.m.} & & \\ \downarrow & & \\ \text{prob} & \to & \text{distribution} \end{array}$$

Theorem 5 *The following relationships hold:*

(a) $X_n \xrightarrow{qm} X$ *implies that* $X_n \xrightarrow{P} X$.

(b) $X_n \xrightarrow{P} X$ *implies that* $X_n \rightsquigarrow X$.

(c) If $X_n \rightsquigarrow X$ *and if* $\mathbb{P}(X = c) = 1$ *for some real number* c, *then* $X_n \xrightarrow{P} X$.

(d) $X_n \xrightarrow{as} X$ *implies* $X_n \xrightarrow{P} X$.

In general, none of the reverse implications hold except the special case in (c).

Proof. We start by proving (a). Suppose that $X_n \xrightarrow{qm} X$. Fix $\epsilon > 0$. Then, using Markov's inequality,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^2 > \epsilon^2) \leq \frac{\mathbb{E}|X_n - X|^2}{\epsilon^2} \to 0.$$

Proof of (b). Fix $\epsilon > 0$ and let x be a continuity point of F. Then

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &= F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Also,

$$\begin{aligned} F(x - \epsilon) &= \mathbb{P}(X \leq x - \epsilon) = \mathbb{P}(X \leq x - \epsilon, X_n \leq x) + \mathbb{P}(X \leq x - \epsilon, X_n > x) \\ &\leq F_n(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Hence,

$$F(x - \epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \leq F_n(x) \leq F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Take the limit as $n \to \infty$ to conclude that

$$F(x - \epsilon) \leq \liminf_{n \to \infty} F_n(x) \leq \limsup_{n \to \infty} F_n(x) \leq F(x + \epsilon).$$

This holds for all $\epsilon > 0$. Take the limit as $\epsilon \to 0$ and use the fact that F is continuous at x and conclude that $\lim_n F_n(x) = F(x)$.

Proof of (c). Fix $\epsilon > 0$. Then,

$$\begin{aligned} \mathbb{P}(|X_n - c| > \epsilon) &= \mathbb{P}(X_n < c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \\ &\leq \mathbb{P}(X_n \leq c - \epsilon) + \mathbb{P}(X_n > c + \epsilon) \\ &= F_n(c - \epsilon) + 1 - F_n(c + \epsilon) \\ &\to F(c - \epsilon) + 1 - F(c + \epsilon) \\ &= 0 + 1 - 1 = 0. \end{aligned}$$

- An example where it converges by probability but not in quadratic mean: $X_n = \sqrt{n} I(0 < u < \frac{1}{n})$. you can show that this converges in probability and not in quadratic mean
- as a counter example for converges in distribution but not by probability:
  Consider $X_n = -X$ where $X \sim N(0, 1)$.
- convergence in distribution is about the cdf. convergence in probability is about the random variable.

- the almost surely convergence is not very relevant for us here in this course.
- we can show that if X_n converges in probability to X and Y_n to Y, then $X_n + Y_n$ converges in probability to $X + Y$, and same is true for the products.
- extension of lotus:

  if X_n converges to X (by probability or by distribution), then a continuous function $g(X_n)$ also converges to $g(X)$ (in the same way as $X \rightarrow$ by probability or by distribution)
- the weak law of large numbers (wlln) says that the sample mean converges in probability to the theoretical mean μ for a bunch of i.i.d random variables.
- the strong law of large numbers says that the sample mean converges almost surely, but we will neither use it nor prove that here.
- The central limit theorem states that if $\overline{X}_n$ is the sample mean (here we only consider iid distributions), then $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ converges in distribution to the standard normal.
- we mentioned the proof idea of the above two in the probability course notes.
- thus, the central limit theorem is very useful to approximate these sample means (and other associated functions)
- berry-esseen theorem theorem provides a bound for $P(\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}) < Z - \phi(Z)$ that is proportional to $\frac{1}{\sqrt{n}}$.
- now, it might not always be possible (in fact we rarely do) know the theoretical variance of the underlying distribution. if instead we use the sample variance, does the central limit theorem still hold? yes it does! which makes this incredibly useful. proof: we will use the ideas that if a random variable converges in probability, then any continuous function of it also does converge in probability (slutzky's theorem). this idea permeates the proof → see page 8 and 9 of this.
- remember that the central limit theorem does *not* say that $\overline{X}$ converges to normal. A certain scaled version of this, specifically $\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ or $\sqrt{n}\overline{X}_n$ if already standardized is what converges.
- extending CLT to smooth functions of the random variable:

Theorem 18 (The Delta Method) *Suppose that*

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \rightsquigarrow N(0,1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \rightsquigarrow N(0,1).$$

- Larry Wasserman gives some vague "intution" for the above but doesn't prove it: $g(\overline{X}) \approx g(\mu) + (\overline{X} - \mu)g'(\mu) \rightarrow \sqrt{n}(g(\overline{X}) - g(\mu)) \approx \sqrt{n}(\overline{X} - \mu)g'(\mu)$. Note that $\sqrt{n}(\overline{X} - \mu)$ converges to $N(0, \sigma^2)$. QED (not really).
- notice the usefulness of the above result → it means that we can approximate a wide range of smooth functions of $\overline{X}$.
- Larry talks about the multivariate CLT, but doesn't prove anything about it. i am omitting that here.

- now we've developed all the probability tools we'll need to actually jump in to start talking about statistics/learning.
- the concept of consistency of an estimator:

An estimator $\widehat{\theta}_n = g(X_1, \ldots, X_n)$ is *consistent* for θ if

$$\widehat{\theta}_n \xrightarrow{\text{P}} \theta$$

as $n \to \infty$. In other words, $\widehat{\theta}_n - \theta = o_p(1)$. Here are two common ways to prove that $\widehat{\theta}_n$ consistent.

Method 1: Show that, for all $\varepsilon > 0$,

$$\mathbb{P}(|\widehat{\theta}_n - \theta| \geq \varepsilon) \longrightarrow 0.$$

Method 2. Prove convergence in quadratic mean:

$$\text{MSE}(\widehat{\theta}_n) = \text{Bias}^2(\widehat{\theta}_n) + \text{Var}(\widehat{\theta}_n) \longrightarrow 0.$$

If bias $\to 0$ and var $\to 0$ then $\widehat{\theta}_n \xrightarrow{qm} \theta$ which implies that $\widehat{\theta}_n \xrightarrow{p} \theta$.

- so from the above, we see the significance of bias and variance for estimators.
- Take a moment to think about the significance of proving consistency. we are showing that as $n \to \infty$ the probability that our estimate is off by even a very small number is low. So consistent estimators can have some really good applications!
- when is the MLE consistent? the proof is beyond the scope of this course but there are conditions under which the MLE is consistent. Remember, consistency is an amazing thing.