Introduction to Machine Learning CMU-10701

9. Tail Bounds

Barnabás Póczos

Fourier Transform and Characteristic Function

Fourier Transform

Fourier transform

unitary transf.

$$\mathcal{F}[f](\omega) = \hat{f}(\omega) = \int_{\mathbb{R}^d} f(x) \exp(-2\pi i \langle \omega, x \rangle) dx$$

Inverse Fourier transform

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(2\pi i \langle \omega, x \rangle) d\omega$$

Other conventions: Where to put 2π?

Not preferred: not unitary transf.
Doesn't preserve inner product

$$\hat{f}(\omega) = \int_{\mathbb{R}^n} f(x) \exp(-i \langle \omega, x \rangle)\, dx.$$

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(i \langle \omega, x \rangle)\, d\omega$$

unitary transf.

$$\hat{f}(\omega) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) \exp(-i \langle \omega, x \rangle)\, dx$$

$$f(x) = \mathcal{F}^{-1}[\hat{f}](x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \hat{f}(\omega) \exp(i \langle \omega, x \rangle)\, d\omega$$

Fourier Transform

Fourier transform

$$\mathcal{F}[f](\omega) = \int_{\mathbb{R}^d} f(x) \exp(-2\pi i \langle \omega, x \rangle) dx$$

Inverse Fourier transform

$$\mathcal{F}^{-1}[g](x) = \int_{\mathbb{R}^d} g(\omega) \exp(2\pi i \langle \omega, x \rangle) d\omega$$

Properties:

Inverse is really inverse: $F \circ F^{-1}[g] = g$ $F^{-1} \circ F[f] = f$

and lots of other important ones…

Fourier transformation will be used to define the characteristic function, and represent the distributions in an alternative way.

Characteristic function

How can we describe a random variable?

- cumulative distribution function (cdf)

$$F_X(x) = \Pr(X \leq x) = \mathbb{E}\left[\mathbf{1}_{\{X \leq x\}}\right]$$

- probability density function (pdf)

The Characteristic function provides an alternative way for describing a random variable

Definition:

$$\varphi_X(t) = \mathbb{E}\left[e^{i\langle t,x\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\, dF_X(x) = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} f_X(x)\, dx$$

The Fourier transform of the density

Characteristic function

$$\varphi_X(t) = \mathbb{E}\left[e^{i\langle t,x\rangle}\right] = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle}\,dF_X(x) = \int_{\mathbb{R}^d} e^{i\langle t,x\rangle} f_X(x)\,dx$$

Properties

- $\varphi_X(t)$ of a real-valued random variable X always exists.

  For example, Cauchy doesn't have mean but still has characteristic function.

- Continuous on the entire space, even if *X* is not continuous.
- Bounded, even if *X* is not bounded $|\varphi_X(t)| \leq 1, \ \forall t \in \mathbb{R}^d$.
- Bijection between cdf and characteristic functions: For any two random variables X_1, X_2, $F_{X_1} = F_{X_2} \Leftrightarrow \varphi_{X_1} = \varphi_{X_2}$
- $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$ if $X \perp\!\!\!\perp Y$.
- $\varphi_{\frac{1}{n}X}(t) = \varphi_X(\frac{t}{n})$
- Characteristic function of constant *a*: $\varphi_{\delta_a}(t) = \exp(i\langle t, a\rangle)$
- Levi's: continuity theorem $\varphi_{X_n}(t) \to \varphi_X(t) \quad \forall t \in \mathbb{R} \Rightarrow X_n \xrightarrow{\mathcal{D}} X$

Weak Law of Large Numbers

Proof II: Goal: $\widehat{\mu}_n \xrightarrow{D} \mu.$

Taylor's theorem for complex functions

$$\exp(itx) = 1 + itx + o(t), \quad t \to 0$$

The Characteristic function

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = 1 + it\mu + o(t)$$

Properties of characteristic functions :

$$\varphi_{\frac{1}{n}X}(t) = \varphi_X\left(\tfrac{t}{n}\right) \quad \text{and} \quad \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \quad \text{if } X \perp\!\!\!\perp Y.$$

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$$

$$\Rightarrow \varphi_{\widehat{\mu}_n}(t) = \left[\varphi_X\left(\frac{t}{n}\right)\right]^n = \left[1 + i\mu\frac{t}{n} + o\left(\frac{t}{n}\right)\right]^n \xrightarrow{n\to\infty} e^{it\mu} = 1 + t\mu + \ldots$$

mean

Levi's continuity theorem $\Rightarrow$ Limit is a constant distribution with mean μ

“Convergence rate” for LLN

Gauss-Markov:

$$\Pr(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1 - \frac{\mathbb{E}[|\widehat{\mu}_n - \mu|]}{\varepsilon} = 1 - \delta$$

Doesn’t give rate

Chebyshev:

$$P(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1 - \frac{\sigma^2}{n\varepsilon^2} = 1 - \delta. \Rightarrow |\widehat{\mu}_n - \mu| < \varepsilon = \frac{\sigma}{\sqrt{n\delta}}$$

with probability 1-δ

Can we get smaller, logarithmic error in δ???

$$\sqrt{\log \frac{1}{\delta}} \ll \frac{1}{\sqrt{\delta}} \text{ if } 0 < \delta < 1$$

Further Readings on LLN, Characteristic Functions, etc

- http://en.wikipedia.org/wiki/Levy_continuity_theorem
- http://en.wikipedia.org/wiki/Law_of_large_numbers
- http://en.wikipedia.org/wiki/Characteristic_function_(probability_theory)
- http://en.wikipedia.org/wiki/Fourier_transform

More tail bounds

More useful tools!

Hoeffding's inequality (1963)

$$\left.\begin{aligned} X_1, \ldots, X_n \text{ independent} \\ X_i \in [a_i, b_i] \\ \varepsilon > 0 \end{aligned}\right\} \Rightarrow$$

$$\Rightarrow \begin{cases} \mathbb{P}(|\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i)| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right) \\ \text{two-sided} \\ \\ \mathbb{P}(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > \varepsilon) \leq \exp\left(\frac{-2n\varepsilon^2}{\frac{1}{n}\sum_{i=1}^n (b_i - a_i)^2}\right) \\ \text{one-sided} \end{cases}$$

It only contains the range of the variables,
but not the variances.

"Convergence rate" for LLN from Hoeffding

Hoeffding Let $c^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$

$$\Rightarrow \Pr(|\widehat{\mu}_n - \mu| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\delta = 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

$$\log\frac{\delta}{2} = \frac{-2n\varepsilon^2}{c^2}$$

$$\frac{c^2}{2n}\log\frac{2}{\delta} = \varepsilon^2$$

$$\varepsilon = c\sqrt{\frac{\log 2 - \log\delta}{2n}}$$

$$\Rightarrow |\widehat{\mu}_n - \mu| < \varepsilon = c\sqrt{\frac{1}{2n}\log\frac{2}{\delta}} \qquad \ll \frac{\sigma}{\sqrt{n\delta}}$$

Proof of Hoeffding's Inequality

A few minutes of calculations.

Bernstein's inequality (1946)

$$\left.\begin{array}{r} X_1, ..., X_n \text{ indep.} \\ X_i \in [a,b] \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i) \\ \varepsilon > 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}(|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_i| > \varepsilon) \leq 2\exp\left(\frac{-n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon(b-a)}\right)$$

It contains the variances, too, and can give tighter bounds than Hoeffding.

Benett's inequality (1962)

$$\left.\begin{array}{r} X_1, ..., X_n \text{ indep.} \\ \mathbb{E}X_i = 0 \\ |X_i| \leq a \\ \sigma^2 = \frac{1}{n}\sum_{i=1}^n Var(X_i) \\ h(u) \doteq (1+u)\log(1+u) - u, \quad u \geq 0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \mathbb{P}(\sum_{i=1}^n X_i > t) \leq \exp\left(-\frac{n\sigma^2}{a^2} h\left(\frac{at}{n\sigma^2}\right)\right)$$

Benett's inequality $\Rightarrow$ Bernstein's inequality.

Proof:

$$h(u) \geq \frac{u^2}{2 + 2u/3} \qquad t = n\varepsilon \qquad n\sigma^2 h\left(\frac{n\varepsilon}{n\sigma^2}\right) \geq ... \geq \frac{n\varepsilon^2}{2\sigma^2 + \frac{2}{3}\varepsilon}$$

McDiarmid's Bounded Difference Inequality

Suppose $X_1, X_2, \ldots, X_n$ are independent and assume that

$$\sup_{x_1,x_2,\ldots,x_n,\widehat{x}_i} |f(x_1,x_2,\ldots,x_n) - f(x_1,x_2,\ldots,x_{i-1},\widehat{x}_i,x_{i+1},\ldots,x_n)| \le c_i$$

$$\text{for } 1 \le i \le n$$

(In other words, replacing the i-th coordinate x_i by some other value changes the value of f by at most c_i.)

It follows that

$$\Pr\{f(X_1,X_2,\ldots,X_n) - E[f(X_1,X_2,\ldots,X_n)] \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{E[f(X_1,X_2,\ldots,X_n)] - f(X_1,X_2,\ldots,X_n) \ge \varepsilon\} \le \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right)$$

$$\Pr\{|E[f(X_1,X_2,\ldots,X_n)] - f(X_1,X_2,\ldots,X_n)| \ge \varepsilon\} \le 2\exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Further Readings on Tail bounds

http://en.wikipedia.org/wiki/Hoeffding's_inequality

http://en.wikipedia.org/wiki/Doob_martingale (McDiarmid)

http://en.wikipedia.org/wiki/Bennett%27s_inequality

http://en.wikipedia.org/wiki/Markov%27s_inequality

http://en.wikipedia.org/wiki/Chebyshev%27s_inequality

http://en.wikipedia.org/wiki/Bernstein_inequalities_(probability_theory)

Limit Distribution?

Central Limit Theorem

Let $X_1, \ldots, X_n$ be i.i.d $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$.
LLN: $\frac{X_1 + \ldots + X_n}{n} - \mu \xrightarrow{a.s.} 0$

Lindeberg-Lévi CLT: $X_1, \ldots, X_n$ i.i.d, $E[X_i] = \mu$, and $Var[X_i] = \sigma^2$.

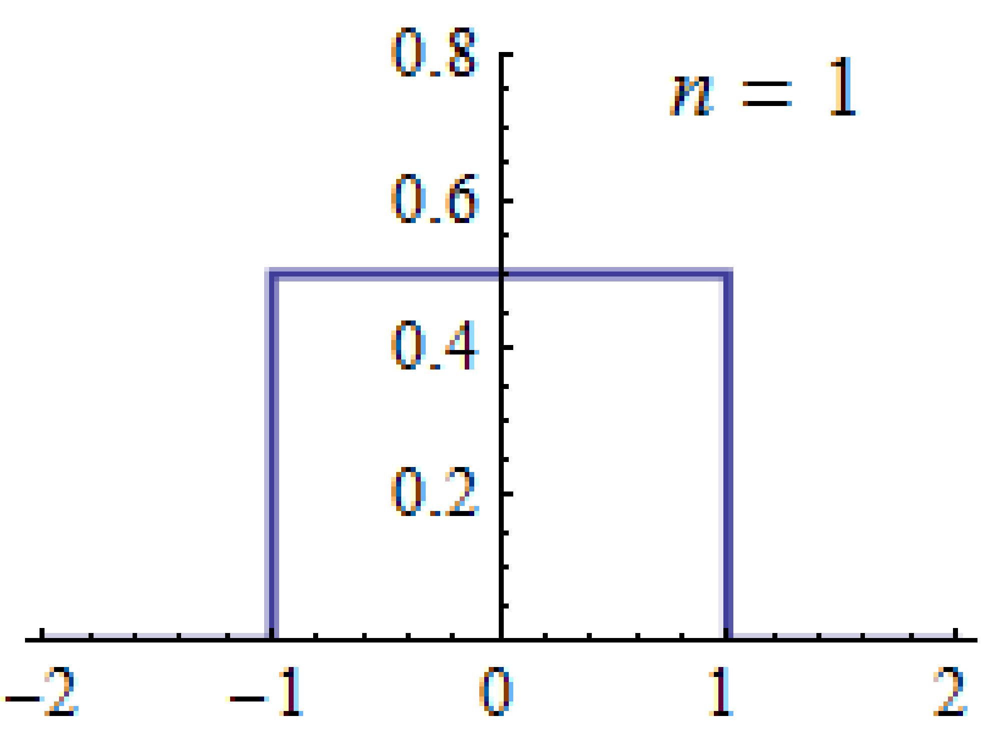
$$\Rightarrow \sqrt{n}\left(\frac{X_1 + \ldots + X_n}{n} - \mu\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$

Lyapunov CLT:

$E[X_i] = \mu_i$, $Var[X_i] = \sigma_i^2$, $s_n^2 = \sum_{i=1}^{n} \sigma_i^2$.

+ some other conditions

$$\Rightarrow \frac{1}{s_n}\left(\sum_{i=1}^{n} X_i - \mu_i\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2)$$



Generalizations: multi dim, time processes

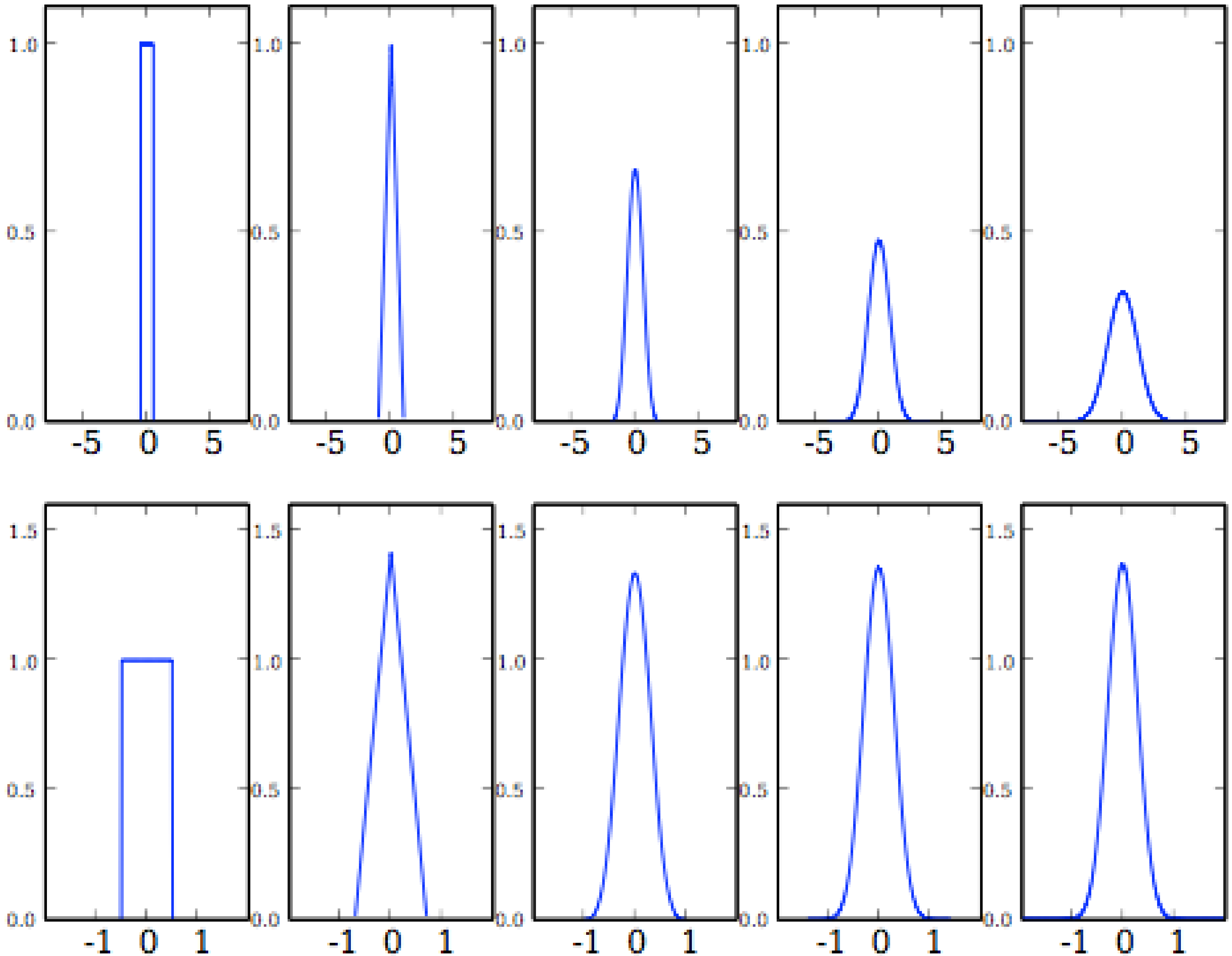
Central Limit Theorem in Practice

unscaled

$$\sum_{i=1}^{n} X_i$$

scaled

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$



Proof of CLT

Let $\mathbb{E}[Y] = 0$, and $Var(Y) = 1$. From Taylor series around 0:

$$\exp(ity) = 1 + ity + \frac{i^2}{2} t^2 y^2 + o(|t|^2)$$

$$\Rightarrow \varphi_Y(t) = \mathbb{E}[\exp(itY)] = 1 - \frac{t^2}{2} + o(t^2), \qquad t \to 0$$

Let $Y_i = \frac{X_i - \mu}{\sigma}$ and let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$

$\mathbb{E}[Y_i] = 0$

$Var(Y_i) = 1$

Properties of characteristic functions :

$$\varphi_{\frac{1}{\sqrt{n}} Z}(t) = \varphi_Z\left(\frac{t}{\sqrt{n}}\right) \quad \text{and} \quad \varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) \quad \text{if } X \perp\!\!\!\perp Y.$$

$$\Rightarrow \varphi_{Z_n}(t) = \prod_{i=1}^{n} \varphi_{Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left[1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right]^n \to e^{-t^2/2}, \qquad n \to \infty$$

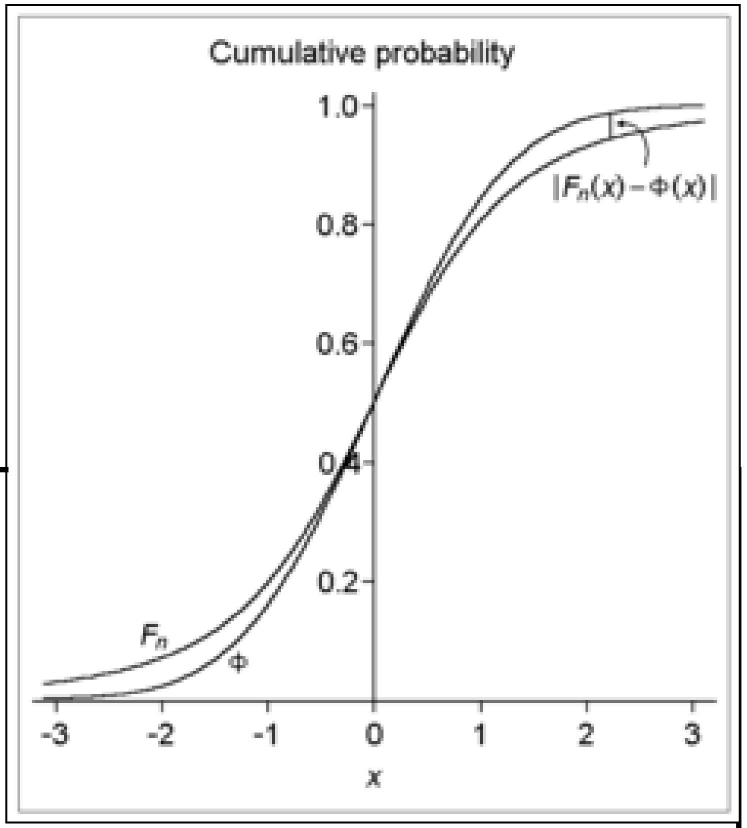
characteristic function of Gauss distribution

Levi's continuity theorem + uniqueness $\Rightarrow$ CLT

How fast do we converge to Gauss distribution?

CLT: $\sqrt{n}\left(\frac{X_1+\ldots+X_n}{n} - \mu\right) \xrightarrow{D} \mathcal{N}(0,\sigma^2)$

It doesn't tell us anything about the convergence rate.



Berry-Esseen Theorem

Let $X_1, \ldots, X_n$ be i.i.d.

$\mathbb{E}[X_1] = \mu$, $\mathbb{E}[X_1^2] = \sigma^2$ $\mathbb{E}[|X_1|^3] = \rho < \infty$

Let $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma}$

F_n is the cdf of Z_n $\Phi(x)$ is the cdf of $\mathcal{N}(0,1)$.

Then $\exists C > 0$ such that for all x and n, $|F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$.

Independently discovered by A. C. Berry (in 1941) and C.-G. Esseen (1942)

Did we answer the questions we asked?

- Do empirical averages converge?
- What do we mean on convergence?
- What is the rate of convergence?
- What is the limit distrib. of “standardized” averages?

Next time we will continue with these questions:

- ❑ How good are the ML algorithms on unknown test sets?
- ❑ How many training samples do we need to achieve small error?
- ❑ What is the smallest possible error we can achieve?

Further Readings on CLT

- http://en.wikipedia.org/wiki/Central_limit_theorem
- http://en.wikipedia.org/wiki/Law_of_the_iterated_logarithm

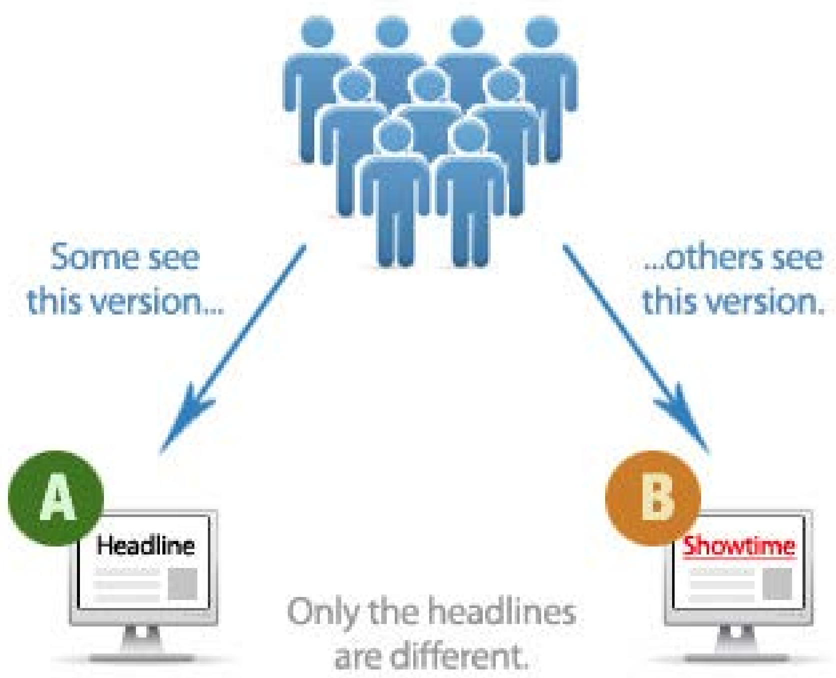
Tail bounds in practice

A/B testing

- Two possible webpage layouts
- Which layout is better?

Experiment

- Some users see A
- The others see design B



How many trials do we need to decide which page attracts more clicks?

A/B testing

Let us simplify this question a bit:

Assume that in group A

$p(click|A) = 0.10$ click and $p(noclick|A) = 0.90$

Assume that in group B

$p(click|B) = 0.11$ click and $p(noclick|B) = 0.89$

Assume also that we *know* these probabilities in group A, but we *don't know* yet them in group B.

We want to estimate $p(click|B)$ with less than 0.01 error

Chebyshev Inequality

$$\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad X_i = \begin{cases} 1 & \text{click} \\ 0 & \text{no click} \end{cases}$$

Chebyshev: $$\Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

- In group B the click probability is μ = 0.11 (we don't know this yet)
- Want failure probability of δ=5%

• If we have no prior knowledge, we can only bound the variance by σ^2 = 0.25 (Uniform distribution hast the largest variance 0.25)

$$\Pr(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} < \delta \Rightarrow \frac{\sigma^2}{\delta\varepsilon^2} < n \Rightarrow \frac{0.25}{0.05 \cdot 0.01^2} = 50,000 < n$$

• If we know that the click probability is < 0.15, then we can bound σ^2 at 0.15 * 0.85 = 0.1275. This requires at least 25,500 users.

Hoeffding's bound

- Hoeffding Let $c^2 = \frac{1}{n}\sum_{i=1}^{n}(b_i - a_i)^2$

$$\Rightarrow \Pr(|\hat{\mu}_n - \mu| > \varepsilon) \leq 2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right)$$

- Random variable has bounded range [0, 1] (click or no click), hence c=1
- Solve Hoeffding's inequality for *n:*

$$2\exp\left(\frac{-2n\varepsilon^2}{c^2}\right) \leq \delta \quad \Rightarrow \left(\frac{-2n\varepsilon^2}{c^2}\right) \leq \log(\delta/2) \Rightarrow -2n\varepsilon^2 \leq c^2\log(\delta/2)$$

$$\Rightarrow n > \frac{c^2\log(2/\delta)}{2\varepsilon^2} = 1 \cdot \frac{\log(2/0.05)}{2 \cdot 0.01^2} = 18,445$$

This is better than Chebyshev.

Thanks for your attention ☺