Lecture notes 16
The Central Limit Theorem

Lecture Outline

- Gaussian Density Revisited
- Standardized Sums
- The CLT

Reading: Bertsekas & Tsitsiklis 7.4

Gaussian Density Revisited

Suppose a random variable Y has a Gaussian pdf $\mathcal{N}(m, \sigma^2)$, i.e.,

$$f_Y(y) = \frac{e^{-\frac{(x-m)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$$

This pdf has several nice properties:

- If $W = aY + b$, then W is also Gaussian, $\mathcal{N}(am + b, a^2\sigma^2)$
- The transform closely resembles the form of the original pdf: $M_Y(s) = e^{sm+s^2\sigma^2/2}$
- If the Y_i are $\mathcal{N}(m_i, \sigma_i^2)$ and mutually independent and $V = \sum_i a_i Y_i$, then V is $\mathcal{N}(\sum_i a_i m_i, \sum_i a_i{}^2 \sigma_i^2)$.

In particular, weighted sums of iid Gaussian variables are also Gaussian.

Will see that certain weighted sums of RVs will be approximately Gaussian even when the original RVs are not Gaussian — a result called the central limit theorem (CLT).

One reason of great popularity of Gaussian models.

Suppose that $\{X_n\}$ is an iid stochastic process with a common distribution F_X described by a pmf or pdf

Assume common mean and variance: $EX_n = m$ and finite variance $\sigma^2_{X_n} = \sigma^2$

Assume also that transform $M_X(s)$ is well behaved for small s

(will make precise later)

Consider the "standardized" or "normalized" sum

$$Z_n = \frac{1}{n^{1/2}} \sum_{k=0}^{n-1} \frac{X_k - m}{\sigma}$$

In the special case $m = 0, \sigma^2 = 1$ this is

$$Z_n = \frac{1}{n^{1/2}} \sum_{k=0}^{n-1} X_k$$

Note the similarity to the sample average S_n considered for the law of large numbers. $Z_n = \sqrt{n} S_n$.

Z_n is called "standardized" because it has zero mean and unit variance:

$$EZ_n = 0 \ , \ \ \sigma^2_{Z_n} = 1 \ .$$

Using transforms and independence:

$$M_{Z_n}(s) = M_{(X-m)/\sigma}\left(\frac{s}{n^{1/2}}\right)^n$$

As $n \to \infty$, $s/\sqrt{n} \to 0$. Provided M_X is suitably smooth, a Taylor series approximation shows that

$$M_{(X-m)/\sigma}(sn^{-1/2}) \approx 1 + \frac{s^2}{2n}$$

Aside: The Taylor series of a function $f(u)$ about the point $u = 0$ has the form

$$\begin{aligned} f(u) &= \sum_{k=0}^{\infty} u^k \frac{f^{(k)}(0)}{k!} \\ &= f(0) + uf^{(1)}(0) + u^2\frac{f^{(2)}(0)}{2} + \text{terms in } u^k \ ; k \geq 3 \ , \quad (1) \end{aligned}$$

where the derivatives

$$f^{(k)}(0) = \frac{d^k}{du^k} f(u)|_{u=0} \ ;$$

are assumed to exist, that is, the function is assumed to be analytic at the origin. Combining the Taylor series expansion with the moment-generating property of transforms yields

$$\begin{aligned} M_Y(s) &= \sum_{k=0}^{\infty} s^k \frac{M_Y^{(k)}(0)}{k!} \\ &= \sum_{k=0}^{\infty} (s)^k \frac{E(Y^k)}{k!} \\ &= 1 + sE(Y) + \tfrac{s^2}{2}E(Y^2) + o(s^2)/2 \ . \end{aligned}$$

where $o(s^2) \to 0$ faster than s^2. Thus replacing Y by $(X-m)/\sigma$ and s by $sn^{-1/2}$,

$$M_{(X-m)/\sigma}(sn^{-1/2}) \approx 1 + \frac{s^2}{2n}$$

as claimed.

Hence

$$\lim_{n\to\infty} M_{Z_n}(s) = \lim_{n\to\infty}[1 + \frac{s^2}{2n}]^n$$

From elementary real analysis:

$$\lim_{n\to\infty} M_{Z_n}(s) = e^{(s^2/2)}$$

So asymptotically the transform of Z_n looks Gaussian $\mathcal{N}(0,1)$

Taking inverse transforms ⇒ the cdf's will also converge to a Gaussian cdf

Theorem (A Central Limit Theorem). Let $\{X_n\}$ be an iid stochastic process with a finite mean m and variance σ^2. Then

$$\lim_{n\to\infty} P(Z_n \leq c) = \Phi(c)$$

where $\Phi(c)$ is the cdf for $\mathcal{N}(0,1)$:

$$\Phi(r) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{r} e^{\frac{-t^2}{2}}$$

Note: X_i might be discrete and not have a pdf, so cannot say pdfs converge.

Intuitively: if sum up a large number of independent random variables and normalize by $n^{-1/2}$, result is approximately Gaussian.

For example, a current meter across a resistor will measure the effects of the sum of millions of electrons randomly moving and colliding with each other. Regardless of the probabilistic description of these micro-events, the global current will appear to be Gaussian.

Thermal noise in radio receivers.

Dust particles suspended on a dish of water and subjected to the random collisions of millions of molecules. Motion of any individual particle in two dimensions will appear to be Gaussian. "Brownian motion."

The Pollster's Problem Revisited

- f is the fraction of the population that intend to vote for Snidely Whiplash against Dudley Doright.
- For the ith person polled

$$X_i = \begin{cases} 1 & \text{prefers Whiplash} \\ 0 & \text{otherwise} \end{cases}$$

- $S_n = (X_1 + X_2 + \cdots X_n)/n$, the sample average of those preferring Whiplash.
- Pollster wants to choose n large enough to ensure that

$$P(|S_n - f| \geq .01) \leq \frac{1}{20}$$

- Before used Chebychev inequality to get

$$P(|S_n - f| \geq .01) \leq \frac{\sigma_X^2}{n(0.01)^2}$$

and conclude $n = 50,000$ suffices

- Now use CLT to approximate probability $P(|S_n - f| \geq .01)$:

$$\begin{aligned} \left|\frac{X_1 + X_2 + \cdots + X_n - nf}{n}\right| &\geq .01 \\ \left|\frac{X_1 + X_2 + \cdots + X_n - nf}{\sqrt{n}\sigma}\right| &\geq \frac{\sqrt{n}.01}{\sigma} \end{aligned}$$

$$0.05 \geq P(|S_n - f| \geq .01)$$

$$= P(|Z_n| \geq \frac{\sqrt{n}.01}{\sigma})$$

Since $\sigma \leq 1/2$, the set $\{|Z_n| \geq \frac{\sqrt{n}.01}{\sigma}\}$ is a subset of the bigger set $\{|Z_n| \geq \sqrt{n}.02\}$ and hence has smaller probability Thus if the bigger probability is less than 0.05, so is the smaller and hence

$$\begin{aligned} 0.05 &\geq P(|Z_n| \geq \sqrt{n}.02) \\ &\geq P(|Z_n| \geq \frac{\sqrt{n}.01}{\sigma}) \\ &= P(|S_n - f| \geq .01) \end{aligned}$$

Using the CLT

$$P(|Z_n| \geq \sqrt{n}.02) \approx 2 \times (1 - \Phi(0.2\sqrt{n}))$$

- Use tables to find n such that this inequality holds. Turns out from tables $n \geq 9604$ does it, a significantly better answer than the previous one.

There are many variations on the CLT with different conditions. The important thing to remember is adding up many iid random variables will *with suitable normalization* yield approximately Gaussian behavior.