CLT FOR DEPENDENT R.V'S

CLT Let $X_1, \dots, X_n$ be mean zero random variables such that each X_i may depend on at most d r.v's in $\{X_1, \dots, X_n\}$. Let $W = X_1 + \dots + X_n$ have $\mathrm{Var}(W) = 1$. Then

$$W_1(W, Z) \le C d^2 \sum_{i=1}^{n} \mathbb{E}|X_i|^3$$

where $Z \sim N(0,1)$.

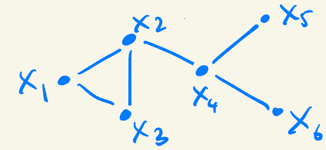
Formally: consider a graph $G = (V, E)$ with $V = \{1, \dots, n\}$. ("Dependency graph")

$\forall i \in V$, let $N_i := \{ j \in V : (i,j) \in E \}$ ("dependency neighborhood")

Assume that:

(a) $\forall i \in V$, $X_i \perp\!\!\!\perp \{X_j : j \notin N_i\}$

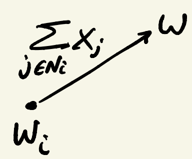
(b) Max. degree of G is $\le d$.



Proof By Stein's lemma (p.133), it is enough to bound

$$|\mathbb{E} W f(W) - \mathbb{E} f'(W)| \quad \text{for any } f \text{ satisfying } \|f'\|_\infty \le 1, \ \|f''\|_\infty \le 1.$$

- $\mathbb{E} W f(W) = \sum_i X_i f(W)$.



$\boxed{W_i := \sum_{j \notin N_i} X_j}$ Taylor approximation about W_i:

$$f(W) = f(W_i) + (W - W_i) f'(W_i) + (W - W_i) \frac{A_i^2}{2!} \quad \text{where} \quad |A_i| \le \|f''\|_\infty \le 1$$

Multiply both sides by X_i, sum over i, take $\mathbb{E}$ $\Rightarrow$

$$\mathbb{E} W f(W) = \underbrace{\sum_i \mathbb{E} X_i f(W_i)}_{\text{independence } = 0} + \underbrace{\sum_i \mathbb{E} \underbrace{X_i \overbrace{(W - W_i)}^{\sum_{j \in N_i} X_j}}_{\leftarrow \text{ independent } \rightarrow} \underbrace{f'(W_i)}}_{= \mathrm{I}} + \underbrace{\sum_i \mathbb{E} X_i \overbrace{(W - W_i)^2}^{\sum_{j \in N_i} X_j} \frac{A_i^2}{2}}_{= \mathrm{II}} \qquad (*)$$

- $\text{I} = \sum_i \sum_{j \in N_i} \mathbb{E}[X_i X_j] \cdot \mathbb{E} f'(W_i)$

$$|f'(W_i) - f'(W)| \le |W_i - W| \cdot \|f''\|_\infty \le |X_i| \Rightarrow \mathbb{E} f'(W_i) = \mathbb{E} f'(W) + B_i \text{ where } |B_i| \le \mathbb{E}|X_i|$$

$$\Rightarrow \text{I} = \underbrace{\sum_i \sum_{j \in N_i} \mathbb{E}[X_i X_j]}_{\sum_{i,j=1}^n \mathbb{E}[X_i X_j] = \mathbb{E}\left(\sum_{i=1}^n X_i\right)^2 = Var(W) = 1} \cdot \mathbb{E} f'(W) + \sum_i \sum_{j \in N_i} \mathbb{E}[X_i X_j] \cdot B_i$$

$a^2 b \le a^3 + b^3$ (Young's inequality)

$$\Rightarrow |\text{I} - \mathbb{E} f'(W)| \le \sum_i \sum_{j \in N_i} \underbrace{\mathbb{E}|X_i X_j| \cdot \mathbb{E}|X_i|}_{\wedge\, CS}$$

$$\|X_i\|_2 \cdot \|X_j\|_2 \cdot \|X_i\|_1 \overset{L_p - L_q}{\le} \|X_i\|_3^2 \cdot \|X_j\|_3 \le \|X_i\|_3^3 + \|X_j\|_3^3$$

$$\le 2d \cdot \sum_{i=1}^n \|X_i\|^3 = 2d \sum_{i=1}^n \mathbb{E}|X_i|^3 .$$

- $$|\text{II}| \underset{\Delta \,\&\, |A_i| \le 1}{\le} \sum_i \mathbb{E}|X_i (W - W_i)^2| = \sum_i \mathbb{E}\Big|X_i \underbrace{\Big(\sum_{j \in N_i} X_j\Big)^2}_{\text{expand}}\Big| \underset{\Delta}{\le} \sum_i \sum_{j,k \in N_i} \underbrace{\mathbb{E}|X_i X_j X_k|}_{\wedge\, \text{am-gm}}$$

$$\frac{\mathbb{E}|X_i|^3 + \mathbb{E}|X_j|^3 + \mathbb{E}|X_k|^3}{3}$$

$$\le d^2 \sum_i \mathbb{E}|X_i|^3 .$$

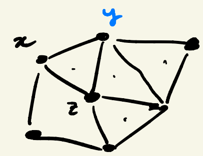
- Put into $(*)$ $\Rightarrow$ $|\mathbb{E} W f(W) - \mathbb{E} f'(W)| \le |\text{I} - \mathbb{E} f'(W)| + |\text{II}| \le 2d^2 \sum_i \mathbb{E}|X_i|^3$. QED

- <u>Example</u>: $W = X_1 X_2 + X_2 X_3 + \cdots + X_{n-1} X_n$ where all X_i are mean 0 indep.
  $d = 3 \Rightarrow W$ satisfies CLT. More generally, time series.

HW: $W = \sum_{i,j=1}^n X_i X_j$ is NOT always approx. normal.

Application: Triangles in Random Graphs

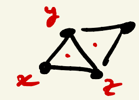
- Consider an Erdös-Rényi random graph $G \sim G(n,p)$ with $p = \text{const}$, $n \to \infty$.



$\Delta = 4$ here

$\Delta := \#(\text{triangles in } G)$ satisfies CLT?

$= \sum_{xyz} \mathbb{1}_{\Delta_{xyz}}$ where $\Delta_{xyz} = \{\text{triple } xyz \text{ is a triangle}\}$



sum over all $\binom{n}{3}$ triples

- $\boxed{d \lesssim n}$ (for a fixed triple xyz, there are $\lesssim n^2$ other triples that share a pair with xyz)

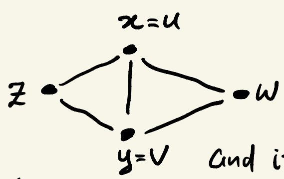
- $\mathbb{P}(\Delta_{xyz}) = p^3 \Rightarrow \boxed{\mathbb{E}\Delta = \binom{n}{3} p^3 \asymp n^3}$

- $\mathrm{Var}(\Delta) = \sum_{xyz,\, uvw} \underbrace{\mathrm{Cov}(\mathbb{1}_{\Delta_{xyz}}, \mathbb{1}_{\Delta_{uvw}})}_{\overset{\shortparallel}{?}}$

(a) All covariances are ≥ 0 [ijk being a Δ may only increase the prob. of lmn being a Δ]

(b) There are $\gtrsim n^4$ covariances that are $\gtrsim 1$

[For such two triples



$\mathrm{cov}(\mathbb{1}_{\Delta_{xyz}}, \mathbb{1}_{\Delta_{uvw}}) > 0$, and its value only depends on $p = \text{const} \Rightarrow \gtrsim 1$

There are $\binom{n}{4} \asymp n^4$ such pairs of triples]

(a) & (b) $\Rightarrow$ $\boxed{\mathrm{Var}(\Delta) \gtrsim n^4}$ $= \sigma^2$

- Use CLT for $W := \frac{\Delta - \mathbb{E}\Delta}{\sqrt{\mathrm{Var}\Delta}} = \sum_{xyz} \frac{\mathbb{1}_{\Delta_{xyz}} - p^3}{\sigma} \Rightarrow$ ($\mathbb{1}_{\Delta_{xyz}}$: Ber$(p^3)$)

$$W_1(W, Z) \lesssim \frac{d^2}{\sigma^3} \underbrace{\sum_{xyz}}_{\lesssim n^3 \text{ terms}} \underbrace{\mathbb{E}\left|\mathbb{1}_{\Delta_{xyz}} - p^3\right|^3}_{\leq 1} \leq \frac{n^2}{n^6} \cdot n^3 \lesssim \frac{1}{n}.$$

$\Rightarrow$ **THM** The # of triangles Δ in Erdös-Rényi graph $G(n,p)$ is approximately normal:

$$d\left(\frac{\Delta - \mathbb{E}\Delta}{\sqrt{\mathrm{Var}\Delta}}, Z\right) \leq \frac{C(p)}{n} \quad \text{where } Z \sim N(0,1)$$

Remark A finer analysis $\Rightarrow$ CLT holds iff $np^3 \to \infty$.

APPLICATION: A Probabilistic Proof of Stirling's Approximation

Thm (Stirling's Approximation) $\quad n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n (1+o(1)) \quad$ as $n\to\infty$

Proof following [Nils Lid Hjort, Emil Aas Stoltenberg, Probability Proofs of Stirling '2024]

$Y_n \sim \text{Poisson}(n)$ can be expressed as a sum of n iid Poisson(1)

$\Rightarrow \quad W := \frac{Y_n - n}{\sqrt{n}}$ is a sum of n iid rv's with mean 0, var 1, third moment $= O(1)$ (check!)

$\Rightarrow$ By Wasserstein CLT,

$$W_1(W, Z) \to 0 \quad \text{where } Z \sim N(0,1)$$

The function $h(x) = \max(x, 0)$ is 1-Lipschitz $\Rightarrow$

$$\Big| \mathbb{E}h(W) - \underbrace{\mathbb{E}h(Z)}_{\frac{1}{\sqrt{2\pi}}\int_0^\infty x e^{-x^2/2}\,dx \;\overset{y=x^2/2}{=\!=}\; \frac{1}{\sqrt{2\pi}}} \Big| \to 0 \quad \text{as } n\to\infty$$

$$\mathbb{E}h(W) = \sum_{k=n}^{\infty} \frac{k-n}{\sqrt{n}} \overbrace{p(k)}^{e^{-n}n^k/k!}$$

$$= \frac{1}{\sqrt{n}}\Big(np(n) - \underbrace{np(n) + (n+1)p(n+1)}_{=\,\text{check!}} - \underbrace{np(n+1) + (n+2)p(n+2)}_{=\,\text{check!}} - \underbrace{np(n+2) + \cdots}_{=} \Big)$$

$$= \frac{np(n)}{\sqrt{n}} = \frac{\sqrt{n}\,e^{-n}n^n}{n!} \longrightarrow \frac{1}{\sqrt{2\pi}} \quad \Rightarrow \text{ Q.E.D.}$$