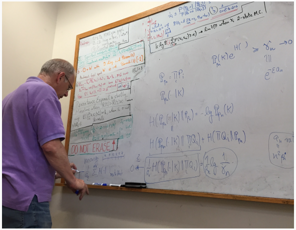
Large deviations of subgraph counts for sparse Erdős–Rényi graphs

Amir Dembo's birthday conference

2018/12/15

Nick Cook, UCLA



Based on joint work with Amir

CLTs and LDPs: Old and new

Classical: $\{X_i\}_{i\geq 1}$ iid, standardized, finite MGF. $S_N = \sum_{i\leq N} X_i$.

- CLT (de Moivre, Laplace...):
  $$\mathbb{P}\left\{\frac{S_N}{\sqrt{N}} \in [a,b)\right\} \to \gamma([a,b)) \;\; \forall\, a < b \qquad \text{(universal)}$$
- LDP (Cramér):
  $$\frac{1}{N}\log \mathbb{P}\left\{\frac{S_N}{N} \in [a,b)\right\} \to -\mathrm{I}(a) \;\; \forall\, 0 < a < b \leq \infty \quad \text{(non-universal)}$$

Extensions to weighted sums $f(X) = \sum_{i\leq N} \alpha_i X_i = \langle \alpha, X\rangle$
(linear functionals on product probability spaces)

Nonlinear functions?

Nonlinear functions:

Example 1: Triangle counts in $G(N,p)$

Let $\boldsymbol{A}$ be $N \times N$ adjacency matrix for $\boldsymbol{G} \sim G(N,p)$.

$f(\boldsymbol{A}) = \operatorname{Tr} \boldsymbol{A}^3 = \sum_{i,j,k} \boldsymbol{A}_{ij}\boldsymbol{A}_{jk}\boldsymbol{A}_{ki} = 6\times$ [# of triangles in $\boldsymbol{G}$].

Cubic polynomial of $\binom{N}{2}$ iid Ber(p) variables.

$\mathbb{E}\, f(\boldsymbol{A}) \sim N^3 p^3$

- CLTs: Ruciński '88, Barbour–Karoński–Ruciński '89 (Stein's method)
- LDPs: (This talk) Chatterjee–Varadhan '11, Chatterjee–Dembo '14, Eldan '16, C.–Dembo '18, Augeri '18, Kozma–Samotij '18... also Lubetzky–Zhao '12, '14, Bhattacharya–Ganguly–L–Z '16

Example 2: k-AP counts in sparse random sets

Let $\boldsymbol{S} \subset \mathbb{Z}/N\mathbb{Z}$ with $\{\mathbf{1}(i \in \boldsymbol{S})\}_i$ iid Ber(p),

$f(\boldsymbol{S})$ the number of 3-term arithmetic progressions in $\boldsymbol{S}$.

Cf. Chatterjee–Dembo '14, Bhattacharya–Ganguly–Shao–Zhao '16.

Nonlinear large deviations (Chatterjee–Dembo '14)

Let $f, h : [0,1]^d \to \mathbb{R} \quad (d \to \infty)$.

Large deviations (with $\boldsymbol{x} \sim \text{Ber}(p)^{\otimes d}$)	Gibbs measure $\nu_h(x) = \frac{1}{Z_h} e^{h(x)}$
$\log \mathbb{P}\left\{f(\boldsymbol{x}) \geq (1+\delta)\,\mathbb{E}\, f(\boldsymbol{x})\right\} \sim ?$	$\log Z_h = \log \sum_{x \in \{0,1\}^d} e^{h(x)} \sim ?$
Conditional on $\boldsymbol{x} \in \{f \geq (1+\delta)\,\mathbb{E}\, f\}$, what does $\boldsymbol{x}$ look like?	What does a typical sample $\boldsymbol{y} \sim \nu_h$ look like?

Well understood for linear functionals. If h has low-complexity gradient,

- (C–D '14) *Naive mean field approximation* is valid:

$$\log Z_h \sim \sup_{x \in [0,1]^d} \{h(x) + H(\mu_x)\}.$$

  (Exact for h linear functional.) Also Yan '17, Augeri '18.
- ν_h is approximately a mixture of $e^{o(d)}$ product measures. ($\Rightarrow$ NMFA) (Eldan '16, Eldan–Gross '17, Austin '18).

Examples: triangle counts

Ex 1. (Eldan) Consider $h : \mathcal{G}_N \cong \{0,1\}^{\binom{N}{2}} \to \mathbb{R}$,

$$h(G) = -\frac{1}{N}\operatorname{Tr} A_G^3 = -\frac{6}{N} \times \#\{\text{ triangles in } G\}.$$

Expect $\boldsymbol{G} \sim \nu_h$ to be approximately a mixture of $2^N = e^{o(N^2)}$ inhomogeneous Erdős–Rényi graphs (product measures) with bipartite structure.

Examples: triangle counts

Ex 2. Let $\boldsymbol{G} \sim G(N, p)$. Conditional on $\boldsymbol{G}$ having extra triangles, i.e.

$$\left\{ \operatorname{Tr} A_{\boldsymbol{G}}^3 \geq N^3 q^3 \right\}, \quad q > p,$$

how are the edges distributed? A few possibilities:

(A) As in $G(N, q)$?

(B) As in $G(N, p)$ with a small planted clique?

(C) As in $G(N, p)$ with a small planted hub?

$*$ For much (but not all!) of $0 < p < q < 1$ fixed, the answer is (A).
(Chatterjee–Varadhan '11 + Lubetzky–Zhao '12).

$*$ Conjecture: For $N^{-1/2} \ll p \ll 1$, $q = (1 + \delta)p$,
Answer is (B) or (C), depending on size of δ.

The upper tail for homomorphism counts: dense case

- For $H = ([n], E_H)$, $G = ([N], E_G)$

$$t(H,G) = \frac{1}{N^n}\,\mathrm{hom}(H,G) = \frac{1}{N^n}\sum_{\varphi:[n]\to[N]}\prod_{\{k,l\}\in E} A_G(\varphi(k),\varphi(l))$$

$$= \mathbb{P}\big\{\text{ uniform random } \varphi : [n]\to[N] \text{ is edge preserving }\big\}.$$

- E.g. $t(C_\ell, G) = \frac{1}{N^\ell}\,\mathrm{Tr}(A_G^\ell)$.
- For p fixed, Chatterjee–Varadhan '11 obtained the LDP for $\{G(N,p)\}_{N\geq 1}$, viewed as measures on the topological space of *graphons* (see Lovasz's book).
- Since $t(H,\cdot)$ are continuous in this topology (the counting lemma), this yields LDPs for $t(H,\boldsymbol{G})$, $\boldsymbol{G}\sim G(N,p)$.

The upper tail for homomorphism counts: sparse case

- Now consider $p = N^{-c}$, $c \in (0,1)$. Graphons are of no help here...
- Chatterjee–Dembo '14: LDP for $t(H, \boldsymbol{G})$ when

$$N^{-\kappa(H)} \ll p \ll 1, \qquad \kappa(H) = \frac{c}{\Delta_H |E_H|}.$$

$(\kappa(C_3) = \frac{1}{41} + \varepsilon)$. Eldan '16: $\kappa(C_3) = \frac{1}{18} + \varepsilon$.

Theorem (C.–Dembo '18)

Fix $H = ([n], E)$ connected of max degree $\Delta \geq 2$. If

$$N^{-\kappa(H)} \ll p \ll 1, \qquad \kappa(H) = \frac{1}{3\Delta - 2},$$

then: $\quad \log \mathbb{P}\left\{t(H, \boldsymbol{G}) \geq (1+\delta)p^{|E|}\right\} \sim -c_H(\delta) N^2 p^{\Delta} \log(1/p).$

The upper tail for homomorphism counts: sparse case

Theorem (C.–Dembo '18)

Fix $H = ([n], E)$ connected of max degree $\Delta \geq 2$. If

$$N^{-\kappa(H)+\varepsilon} \leq p \ll 1, \qquad \kappa(H) = \frac{1}{3\Delta - 2},$$

then: $\quad \log \mathbb{P}\left\{ t(H, \boldsymbol{G}) \geq (1+\delta)p^{|E|} \right\} \sim -c_H(\delta) N^2 p^{\Delta} \log(1/p).$

Remarks:

- Formula for $c_H(\delta)$ was obtained by Bhattacharya, Ganguly, Lubetzky and Zhao '16, valid down to $\kappa(H) = 1/\Delta$. Reflects a phase transition between planted clique and planted hub structures.
- Actually get a better (more complicated) $\kappa(H)$, in particular $\kappa(H) = 1/(2\Delta - 1)$ for H a star.
- In the case of cycles we can sharpen to $\kappa(C_\ell) = 1/2 + \varepsilon$, $\ell \geq 4$, (Augeri '18: $\ell \geq 3$).
- Also get lower tails.

Ideas I: Coverings of events by convex bodies

Want to show for Ber(p) vector $\boldsymbol{a} \in \{0,1\}^d$ and $\mathcal{L} \subseteq [0,1]^d$,

$$\log \mathbb{P}(\boldsymbol{a} \in \mathcal{L}) \leq -\mathrm{I}_p(\mathcal{L}) + \mathit{Error}, \qquad \mathrm{I}_p(\mathcal{L}) := \inf\{\mathrm{I}_p(x) : x \in \mathcal{L}\}.$$

$$I_p(x) = D_{KL}(\mu_x \| \mu_p^{\otimes d}) = \sum_{i=1}^{d} x \log \frac{x}{p} + (1-x) \log \frac{1-x}{1-p}.$$

* (Easy) true (with Error=0) for $\mathcal{L} = \mathcal{H} \cap [0,1]^d$, $\mathcal{H}$ closed half-space.
* **Exercise:** true (with Error=0) for $\mathcal{L}$ compact and convex. [cf. Dembo–Zeitouni Ex. 4.5.5.]
* But for UT problems, we have $\mathcal{L} = \{x : f(x) \geq t\}$, non-convex.
* Idea: Show we can efficiently cover such $\mathcal{L}$ with convex bodies $\{\mathcal{B}_i\}_{i \in \mathcal{I}}$ on which f is essentially constant. Then

$$\begin{aligned} \log \mathbb{P}(\boldsymbol{a} \in \mathcal{L}) \leq \log \sum_{i \in \mathcal{I}} \mathbb{P}(\boldsymbol{a} \in \mathcal{B}_i) &\leq -\min_{i \in \mathcal{I}} \mathrm{I}_p(\mathcal{B}_i) + \log |\mathcal{I}| \\ &= -\mathrm{I}_p(\cup_i \mathcal{B}_i) + \log |\mathcal{I}| \\ &\approx -\mathrm{I}_p(\mathcal{L}) + \log |\mathcal{I}|. \end{aligned}$$

Ideas II: the regularity method

Weighted adjacency matrices $\mathcal{X}_N := \{X = (x_{ij})_{1\le i<j\le N}, x_{ij} \in [0,1]\}$.
Cut norm: $\|X\|_\square = \max_{S,T\subseteq[N]} \left|\sum_{i\in S, j\in T} x_{ij}\right|$.

Weak regularity lemma (compactness):
For any $X \in \mathcal{X}_N$ and $k \in \mathbb{N}$ there exists a partition $\mathcal{P}$ of $[N]$ into k parts and $Y \in \mathcal{X}_N$ constant on $\mathcal{P}$-blocks such that $\|X - Y\|_\square \le \frac{2}{\sqrt{\log k}}$.

Counting lemma (continuity): (recall $t(H,X) = \frac{1}{N^{|V|}} \hom(H,X)$)
For any graph $H = (V,E)$ and $X, Y \in \mathcal{X}_N$,
$|t(H,X) - t(H,Y)| \le |E| \cdot \frac{1}{N^2} \|X - Y\|_\square$.

Weak regularity lemma due to Frieze–Kannan'99
(regularity lemma goes back to Szemerédi '70s).

Taken together: Can cover $\mathcal{X}_N$ with neighborhoods of bdd number of graphons on which $t(H,\cdot)$ functionals are essentially constant. (key for Chatterjee–Varadhan '11.)

Spectral proof of the regularity lemma

(Cf. Frieze–Kannan '99, Szegedy '11, Tao blog '12)

Let $X = \sum_{j=1}^{N} \lambda_j u_j u_j^{\mathsf{T}}$ spectral decomposition for $X \in \mathcal{X}_N$, with $\|X\|_{\mathrm{op}} = \lambda_1(X) \geq |\lambda_2(X)| \geq \cdots \geq |\lambda_N(X)|$.

For any $0 \leq r \leq N-1$,

$$(r+1)|\lambda_{r+1}(X)|^2 \leq \sum_{j=1}^{N} \lambda_j^2 = \sum_{i,j=1}^{N} |X_{ij}|^2 \leq N^2.$$

$\Rightarrow \quad |\lambda_{r+1}(X)| \leq \frac{N}{\sqrt{r+1}}$.

So for r large, X is close in operator norm to a rank-r matrix.
Take parts of $\mathcal{P}$ to be mutual refinement of approximate level sets of $u_1, \ldots, u_r$.

Spectral regularity lemma for random graphs

Proposition

Let $N \in \mathbb{N}, K \geq 1$, $p \in (0,1)$ with $Np \geq \log N$, and $1 \leq r \leq Np$. There exists a partition $\{0,1\}^{\binom{N}{2}} = \bigsqcup_{j=0}^{J} \mathcal{E}_j$ with the following properties:

(a) $\log J \lesssim rN \log(3 + \frac{r}{Kp})$;

(b) $\mathbb{P}\{\boldsymbol{G}_{N,p} \in \mathcal{E}_0\} \lesssim \exp(-cK^2N^2p^2)$;

(c) For each $1 \leq j \leq J$, there exists $Y_j \in \mathcal{X}_N$ of rank at most r such that $\|A_G - Y_j\|_{\mathrm{op}} \lesssim \frac{KNp}{\sqrt{r}}$ for all $G \in \mathcal{E}_j$.

Spectral regularity lemma for random graphs

Proposition

Let $N \in \mathbb{N}, K \geq 1$, $p \in (0,1)$ with $Np \geq \log N$, and $1 \leq r \leq Np$. There exists a partition $\{0,1\}^{\binom{N}{2}} = \bigsqcup_{j=0}^{J} \mathcal{E}_j$ with the following properties:

(a) $\log J \lesssim rN \log(3 + \frac{r}{Kp})$;

(b) $\mathbb{P}\{\boldsymbol{G}_{N,p} \in \mathcal{E}_0\} \lesssim \exp(-cK^2N^2p^2)$;

(c) For each $1 \leq j \leq J$, there exists $Y_j \in \mathcal{X}_N$ of rank at most r such that $\|A_G - Y_j\|_{\mathrm{op}} \lesssim \frac{KNp}{\sqrt{r}}$ for all $G \in \mathcal{E}_j$.

Spectral counting lemma for random graphs

Proposition

Let $H = (V, E)$ with $|V| = n$, $|E| = m$, max degree Δ.
Let $N \in \mathbb{N}$ and $p \in (0, 1)$. For $K \geq 1$ set

$$\mathcal{E}_H(K) = \Big\{X \in \mathcal{X}_N : \exists F \leq H \text{ with } \hom_F(X) > KN^{|V_F|}p^{|E_F|}\Big\}.$$

(a) If $N^{-1/\Delta} < p < 1$, then for any $K \geq 2$,

$$\mathbb{P}\left\{\boldsymbol{G}_{N,p} \in \mathcal{E}_H(K)\right\} \lesssim_H \exp\Big(-c(H)K^{1/n}N^2p^\Delta\Big).$$

(b) For any $X, Y \in \mathcal{X}_N$ with $X \notin \mathcal{E}_H(K)$, for all $F \leq H$,

$$|\hom(F, X) - \hom(F, Y)| \lesssim_H KN^{|V_F|}p^{|E_F|}\frac{\|X - Y\|_{\mathrm{op}}}{Np^{\Delta}}.$$

Open problems and future directions

1. Prove a sharper counting lemma (perhaps using different convex bodies).
2. $p \ll N^{-1/\Delta}$?
3. Other measures besides $G(N, p)$?
4. Partition function / structural decomposition for exponential random graphs (Chatterjee–Diaconis '12, Chatterjee–Dembo '14, Eldan–Gross '17).

Happy birthday, Amir!