Nonlinear large deviations

Sourav Chatterjee

Stanford University

Joint work with Amir Dembo

Large deviations

- Take any smooth $f : [0,1]^n \to \mathbb{R}$.
- Let $Y = (Y_1, \ldots, Y_n)$ be a vector of i.i.d. *Bernoulli*(p) random variables.
- Goal of large deviations theory: Find an approximation for the upper tail probability $\mathbb{P}(f(Y) \geq t)$ when t is much bigger than $\mathbb{E}(f(Y))$.
- Classical large deviations theory well-suited for linear f, in great generality.
- May be quite nontrivial even for very simple nonlinear f. Problems tackled on ad hoc basis.
- For example, the large deviations theory for the number of triangles in a random graph, which is just a polynomial of degree 3, uses Szemerédi's regularity lemma (C. & Varadhan, 2010).

Goal of this work

- For $x = (x_1, \ldots, x_n) \in [0,1]^n$, define

$$I_p(x) := \sum_{i=1}^{n} \left(x_i \log \frac{x_i}{p} + (1 - x_i) \log \frac{1 - x_i}{1 - p} \right).$$

- For each $t \in \mathbb{R}$, define

$$\phi_p(t) := \inf\{I_p(x) \,:\, x \in [0,1]^n \text{ such that } f(x) \geq tn\}.$$

- In many problems, it turns out that

$$\mathbb{P}(f(Y) \geq tn) \approx \exp(-\phi_p(t)). \qquad (\star)$$

- In particular, this is true in great generality for linear functions.
- We give a sufficient condition under which the above approximation is valid for nonlinear maps.

Low complexity gradient condition

Theorem (C. & Dembo, 2014. Very rough statement.)

The approximation ($\star$) *is valid when, in addition to some smoothness conditions on the function* f*, the gradient vector* $\nabla f(x) = (\partial f/\partial x_1, \ldots \partial f/\partial x_n)$ *may be approximately encoded by* $o(n)$ *bits of information.*

- We call this the "low complexity gradient" condition.
- Actual statement of the theorem involves a disgracefully messy error term arising out of the smoothness conditions on f.
- Many notable results on sharp upper and lower bounds for tail probabilities of nonlinear functions (Talagrand, Kim, Vu, Latała, ...) that hold up to constant factors in the exponent, but no results about the precise approximation ($\star$).
- Some preliminary work in C. & Dey (2009).

Example 1: 1D Ising model

- Let
$$f(x) = \sum_{i=1}^{n-1} x_i x_{i+1}\,.$$
- Then, for $2 \le i \le n-1$,
$$\frac{\partial f}{\partial x_i} = x_{i-1} + x_{i+1}\,.$$
- Thus, for this f, the gradient vector cannot be approximately encoded by $o(n)$ many bits. To know $\nabla f(x)$, even approximately, we need to know the values of all the x_i's.
- One can check that the approximation $(\star)$ is not valid for this f.

Example 2: Curie-Weiss model

- Let
$$f(x) = \frac{1}{n} \sum_{1 \le i < j \le n} x_i x_j \, .$$
- For each i,
$$\frac{\partial f}{\partial x_i} = \frac{1}{n} \sum_{j \ne i} x_j = -\frac{x_i}{n} + \frac{1}{n} \sum_{j=1}^{n} x_j \, .$$
- Thus, for this f, the gradient vector is approximately encoded by the single quantity $n^{-1} \sum x_j$.
- The large deviation probabilities for this function satisfy the approximation ($\star$).

Example 3: Subgraph counts in sparse random graphs

- Let T be the number of triangles in an Erdős-Rényi random graph $G(N,p)$.
- Then
$$T = \frac{1}{6}\sum_{i,j,k} Y_{ij}\,Y_{jk}\,Y_{ki}\,,$$
where Y_{ij} is the indicator that edge $\{i,j\}$ is present in the graph.
- Let $n = N(N-1)/2$ and let us agree to denote elements of $\mathbb{R}^n$ as $x = (x_{ij})_{1\le i<j\le N}$, with the convention that $x_{ii} = 0$ and $x_{ji} = x_{ij}$. Define a function $f : \mathbb{R}^n \to \mathbb{R}$ as
$$f(x) = \frac{1}{N}\sum_{i,j,k=1}^{N} x_{ij}x_{jk}x_{ki}\,.$$
- The plan is to apply the main theorem to this f.

Example 3 contd.

- Note that
$$\frac{\partial f}{\partial x_{ij}} = \frac{3}{N}\sum_{k=1}^{N} x_{ik}x_{jk} =: 3a_{ij}(x)\,.$$
- To apply the main theorem, we need to show that $a_{ij}(x)$'s may be approximately encoded by $o(N^2)$ bits.

Example 3 contd.

- Note that for any x and y,

$$\sum_{i,j=1}^{N}(a_{ij}(x)-a_{ij}(y))^2=\frac{1}{N^2}\sum_{i,j,k,l}(x_{ik}x_{jk}-y_{ik}y_{jk})(x_{il}x_{jl}-y_{il}y_{jl})\,.$$

- Consider one pair of terms in the expansion:

$$\frac{1}{N^2}\sum_{i,j,k,l}(x_{ik}x_{jk}x_{il}x_{jl}-x_{ik}x_{jk}y_{il}y_{jl})\,.$$

- This term may be written in a telescoping manner as

$$\frac{1}{N^2}\sum_{i,j,k,l}x_{ik}x_{jk}x_{il}(x_{jl}-y_{jl})+\frac{1}{N^2}\sum_{i,j,k,l}x_{ik}x_{jk}(x_{il}-y_{il})y_{jl}\,.$$

- Let $M(x)$ be the matrix whose (i,j)th entry is x_{ij}.
- Consider the first sum. If i and k are fixed, then the sum in j and l is a quadratic form of the matrix $M(x)-M(y)$.

Example 3 contd.

- This shows that the first sum is bounded by

$$N\|M(x) - M(y)\|_{\text{op}}\,,$$

where $\|M(x) - M(y)\|_{\text{op}}$ is the L^2 operator norm of the matrix $M(x) - M(y)$.

- Similarly bounding other terms, we get

$$\sum_{i,j}(a_{ij}(x) - a_{ij}(y))^2 \le CN\|M(x) - M(y)\|_{\text{op}}\,,$$

where C is a universal constant.

Example 3 contd.

- Fact from linear algebra: For any k, $M(x)$ may be approximated by a rank k matrix with $O(Nk^{-1/2})$ error of approximation in the operator norm.
- Another easy fact: A rank k matrix may be encoded by $O(Nk \log N)$ bits.
- Thus, taking $1 \ll k \ll N/\log N$, and combining with the inequality

$$\frac{1}{N^2}\sum_{i,j}(a_{ij}(x) - a_{ij}(y))^2 \le \frac{C}{N}\|M(x) - M(y)\|_{\mathrm{op}}\,,$$

  it is now easy to see how the $a_{ij}(x)$'s may be approximately encoded by $o(N^2)$ bits.
- This proves the low complexity gradient condition for triangle counts. The proof for general subgraph counts is a messy but straightforward generalization of the above argument.

Example 3 contd.

- What is the precise result for triangle counts?
- For $x = (x_{ij})_{1 \le i < j \le N}$, define

$$I_p(x) := \sum_{1 \le i < j \le N} \left(x_{ij} \log \frac{x_{ij}}{p} + (1 - x_{ij}) \log \frac{1 - x_{ij}}{1 - p} \right)$$

and

$$T(x) := \frac{1}{6} \sum_{i,j,k} x_{ij} x_{jk} x_{ki} \,,$$

where $x_{ij} \in [0, 1]$ and $x_{ji} = x_{ij}$, $x_{ii} = 0$.

- For $u > 1$ define

$$\psi_p(u) := \inf\{I_p(x) : T(x) \ge u\, \mathbb{E}(T)\}\,,$$

where T is the number of triangles in $G(N, p)$.

Example 3 contd.

Theorem (C. & Dembo, 2014.)

For $u > 1$ and N sufficiently large (depending only on u),

$$1 - \frac{c \log N}{N^{1/6} p^2} \le \frac{\psi_p(u)}{-\log \mathbb{P}(T \ge u\,\mathbb{E}(T))} \le 1 + \frac{C(\log N)^{33/29}}{N^{1/29} p^{42/29}},$$

where c and C are constants that depend only on u.

- In particular,
$$\frac{\psi_p(u)}{-\log \mathbb{P}(T \ge u\,\mathbb{E}(T))} \to 1$$
if $N \to \infty$ and $p \to 0$ slower than $N^{-1/42}(\log N)^{11/14}$.
- There is no reason to believe that this should be the optimal threshold for the validity of the approximation, but at least it allows a polynomial rate of decay for p, which is possibly better than anything that may be obtained by a Szemerédi type argument.

Example 4: Three term arithmetic progressions

- Let A be a random subset of $\mathbb{Z}/n\mathbb{Z}$, constructed by keeping each element with probability p, and dropping with probability $1-p$.
- Let X be the number of pairs $(i,j) \in (\mathbb{Z}/n\mathbb{Z})^2$ such that $\{i, i+j, i+2j\} \subseteq A$.
- For $x \in [0,1]^{\mathbb{Z}/n\mathbb{Z}}$, let $I_p(x)$ be defined as before.
- Let

$$\theta_p(u) := \inf\Bigg\{ I_p(x) : x \in [0,1]^{\mathbb{Z}/n\mathbb{Z}} \text{ such that } \sum_{i,j \in \mathbb{Z}/n\mathbb{Z}} x_i x_{i+j} x_{i+2j} \geq u\,\mathbb{E}(X) \Bigg\}.$$

Example 4 contd.

Theorem (C. & Dembo, 2014)

For any $u > 1$,

$$1 - c\, n^{-1/6} p^{-6} \log n \le \frac{\theta_p(u)}{-\log \mathbb{P}(X \ge u\, \mathbb{E}(X))} \le 1 + C n^{-1/29} p^{-162/29} (\log n)^{33/29},$$

where C and c are constants that may depend only on u.

- This theorem gives an approximation for the upper tail of the number of three-term arithmetic progressions in random subsets of $\mathbb{Z}/n\mathbb{Z}$, even when the random subset is allowed to be a little sparse ($p \gg n^{-1/162}(\log n)^{33/162}$).

Example 5: Exponential random graphs

The main theorem gives an error bound in the approximation of normalizing constants in exponential random graph models, which are extensively used in the study of social networks. Previous work of C. & Diaconis (2013), based on Szemerédi's lemma, gave limiting formulas but no error bounds. Will skip this example in this talk.

Why sub-optimal?

- The sub-optimality of the error bounds is probably due to the sub-optimality of the smoothness conditions in the main theorem.
- The proof, in its current form, has scope for improvement. It would be interesting to see if an optimal theorem can be proved along these lines.

Precise statement of the main theorem

- The main theorem follows as a consequence of a more general theorem about normalizing constants, which I state below.
- For any $f : [0,1]^n \to \mathbb{R}$, let $\|f\|$ denote the supremum norm of f.
- Let
$$f_i := \frac{\partial f}{\partial x_i} \quad \text{and} \quad f_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}.$$
- Define
$$a := \|f\|, \quad b_i := \|f_i\| \quad \text{and} \quad c_{ij} := \|f_{ij}\|.$$
- Given $\epsilon > 0$, let $\mathcal{D}(\epsilon)$ be a finite subset of $\mathbb{R}^n$ such that for all $x \in \{0,1\}^n$, there exists $d = (d_1, \ldots, d_n) \in \mathcal{D}(\epsilon)$ such that
$$\sum_{i=1}^n (f_i(x) - d_i)^2 \le n\epsilon^2.$$

Precise statement contd.

- For $x = (x_1, \ldots, x_n) \in [0,1]^n$, let

$$I(x) := \sum_{i=1}^n (x_i \log x_i + (1 - x_i) \log(1 - x_i)) .$$

- Let

$$F := \log \sum_{x \in [0,1]^n} e^{f(x)} .$$

- Given $\epsilon > 0$, define

$$\text{complexity term} := \frac{1}{4}\Big(n \sum_{i=1}^n b_i^2\Big)^{1/2} \epsilon + 3n\epsilon + \log |\mathcal{D}(\epsilon)|, \quad \text{and}$$

$$\text{smoothness term} := 4\Bigg(\sum_{i=1}^n (ac_{ii} + b_i^2) + \frac{1}{4}\sum_{i,j=1}^n \big(ac_{ij}^2 + b_i b_j c_{ij} + 4 b_i c_{ij}\big)\Bigg)^{1/2}$$

$$+ \frac{1}{4}\Big(\sum_{i=1}^n b_i^2\Big)^{1/2}\Big(\sum_{i=1}^n c_{ii}^2\Big)^{1/2} + 3\sum_{i=1}^n c_{ii} + \log 2.$$

Precise statement contd.

Theorem (C. & Dembo, 2014)

For any $\epsilon > 0$,

$$F \le \sup_{x \in [0,1]^n} (f(x) - I(x)) + \text{complexity term} + \text{smoothness term},$$

and

$$F \ge \sup_{x \in [0,1]^n} (f(x) - I(x)) - \frac{1}{2} \sum_{i=1}^{n} c_{ii}.$$

Proof sketch

- Let $X = (X_1, \ldots, X_n)$ be a random vector that has probability density proportional to $e^{f(x)}$ on $\{0,1\}^n$ with respect to the counting measure.
- For each i, define a function $\hat{x}_i : [0,1]^n \to [0,1]$ as

$$\hat{x}_i(x) = \mathbb{E}(X_i \mid X_j = x_j,\ 1 \le j \le n,\ j \ne i).$$

- Let $\hat{x} : [0,1]^n \to [0,1]^n$ be the vector-valued function whose ith coordinate function is $\hat{x}_i$.
- Let $\hat{X} = \hat{x}(X)$.
- The first step in the proof is to show that if the smoothness term is small, then

$$f(X) \approx f(\hat{X}) \quad \text{with high probability.}$$

Proof sketch contd.

- To show this, define

$$h(x) := f(x) - f(\hat{x}(x))\,.$$

- Let $u_i(t,x) := f_i(tx + (1-t)\hat{x}(x))$, so that

$$h(x) = \int_0^1 \sum_{i=1}^n (x_i - \hat{x}_i(x)) u_i(t,x)\, dt.$$

- Thus, if $D := f(X) - f(\hat{X})$, then

$$\mathbb{E}(D^2) = \int_0^1 \sum_{i=1}^n \mathbb{E}((X_i - \hat{X}_i) u_i(t,X) D)\, dt\,. \qquad (\dagger)$$

Proof sketch contd.

- Let $X^{(i)}$ denote the random vector $(X_1, \ldots, X_{i-1}, 0, X_{i+1}, \ldots, X_n)$ and let $D_i := h(X^{(i)})$.
- Then note that $u_i(t, X^{(i)})D_i$ is a function of the random variables $(X_j)_{j \neq i}$ only.
- Therefore since $\hat{X}_i = \mathbb{E}(X_i \mid (X_j)_{j \neq i})$,

$$\mathbb{E}((X_i - \hat{X}_i)u_i(t, X^{(i)})D_i) = 0.$$

- Thus,

$$\begin{aligned}&\mathbb{E}((X_i - \hat{X}_i)u_i(t, X)D) \\ &= \mathbb{E}((X_i - \hat{X}_i)u_i(t, X)D) - \mathbb{E}((X_i - \hat{X}_i)u_i(t, X^{(i)})D_i)\,.\end{aligned}$$

- If the smoothness term is small, then $u_i(t, X) \approx u_i(t, X^{(i)})$ and $D \approx D_i$. Together with the identity (†), this shows that $f(X) \approx f(\hat{X})$ with high probability.

Proof sketch contd.

- Define a function $g : [0,1]^n \times [0,1]^n \to \mathbb{R}$ as

$$g(x,y) := \sum_{i=1}^{n} (x_i \log y_i + (1-x_i)\log(1-y_i))\,.$$

- By a similar argument as above, it is possible to show that if the smoothness term is small, then with high probability,

$$g(X,\hat{X}) \approx g(\hat{X},\hat{X}) = I(\hat{X})\,.$$

- Let A be the set of all x where $f(x) \approx f(\hat{x}(x))$ and $g(x,\hat{x}(x)) \approx I(\hat{x}(x))$.
- Since $X \in A$ with high probability,

$$\frac{\sum_{x\in A} e^{f(x)}}{\sum_{x\in\{0,1\}^n} e^{f(x)}} \approx 1\,.$$

Proof sketch contd.

- Therefore

$$F = \log \sum_{x \in \{0,1\}^n} e^{f(x)} \approx \log \sum_{x \in A} e^{f(x)}$$
$$\approx \log \sum_{x \in A} e^{f(\hat{x}(x)) - I(\hat{x}(x)) + g(x, \hat{x}(x))}.$$

- Now let ϵ be a small positive number.
- Using the set $\mathcal{D}(\epsilon)$, it is easy to produce a set $\mathcal{D}'(\epsilon) \subseteq [0,1]^n$ such that $|\mathcal{D}(\epsilon)| = |\mathcal{D}'(\epsilon)|$, and for each x there exists $p \in \mathcal{D}'(\epsilon)$ such that $\hat{x}(x) \approx p$.
- For each $p \in \mathcal{D}'(\epsilon)$ let $\mathcal{P}(p)$ be the set of all $x \in \{0,1\}^n$ such that $\hat{x}(x) \approx p$.

Proof sketch contd.

- The crucial fact is that for any $p \in [0,1]^n$,
$$\sum_{x \in \{0,1\}^n} e^{g(x,p)} = 1.$$
- Therefore,
$$\log \sum_{x \in A} e^{f(\hat{x}(x)) - I(\hat{x}(x)) + g(x, \hat{x}(x))}$$
$$\leq \log \sum_{p \in \mathcal{D}'(\epsilon)} \sum_{x \in \mathcal{P}(p)} e^{f(\hat{x}(x)) - I(\hat{x}(x)) + g(x, \hat{x}(x))}$$
$$\approx \log \sum_{p \in \mathcal{D}'(\epsilon)} \sum_{x \in \mathcal{P}(p)} e^{f(p) - I(p) + g(x,p)}$$
$$\leq \log \sum_{p \in \mathcal{D}'(\epsilon)} e^{f(p) - I(p)} \leq \log |\mathcal{D}'(\epsilon)| + \sup_{p \in [0,1]^n} (f(p) - I(p)).$$
- This completes the proof sketch for the upper bound. The lower bound is a lot simpler, so I'll skip that.

Summary

- The main result gives an approximation for the tail probabilities of arbitrary smooth functions of independent Bernoulli random variables.
- The approximation holds when the gradient of the function in question has low complexity.
- Applications are given to large deviations for sparse random graphs, three-term arithmetic progressions in random subsets of integers, and exponential random graph models.
- Open problems: Longer arithmetic progressions, optimal levels of sparsity in the subgraph count problem, cleaner version of main theorem, etc.