A general method for lower bounds on fluctuations of random variables

Sourav Chatterjee

The problem of lower bounds

- There are many ways of proving upper bounds on fluctuations of random variables (concentration inequalities), but ...
- ... very few methods for lower bounds.
- In fact, the only ones I know are:
  - Prove distributional convergence.
  - Prove a lower bound on the variance and a matching upper bound on a higher moment, e.g. Aizenman & Wehr (1990).
  - A coupling method of Janson (1994). Only one application.
  - Problem-specific methods, e.g. Pemantle & Peres (1994).
- Many open questions, because there are many modern problems where none of the above approaches work.
- In this talk, I will introduce a new method for lower bounds, that gives new results for:
  - First-passage percolation.
  - Traveling salesman and minimal matching.
  - Random assignment problem.
  - Spin glasses.
  - Random matrices.

Lévy concentration function

- The Lévy concentration function f of a random variable X is defined as
$$f(h) := \sup_{x \in \mathbb{R}} \mathbb{P}(x \leq X \leq x + h).$$
- We will say that a sequence of random variables X_n has fluctuations of order at least δ_n if for some $c > 0$,
$$\limsup_{n \to \infty} f_n(c\delta_n) < 1,$$
where f_n is the Lévy concentration function of X_n.

Main lemma

Lemma (C., 2017)

Let X and Y be two random variables defined on the same probability space. Then for any $-\infty < a \le b < \infty$,

$$\mathbb{P}(a \le X \le b) \le \frac{1}{2}(1 + \mathbb{P}(|X - Y| \le b - a) + d_{\mathrm{TV}}(\mathcal{L}_X, \mathcal{L}_Y)),$$

where $\mathcal{L}_X$ is the law of X, $\mathcal{L}_Y$ is the law of Y, and d_{TV} is total variation distance.

Idea: To show that $\mathbb{P}(X \in I)$ is uniformly bounded away from 1 for all intervals I of length $\le \delta$, construct a random variable Y such that:

- $d_{\mathrm{TV}}(\mathcal{L}_X, \mathcal{L}_Y)$ is small, and
- $\mathbb{P}(|X - Y| \le \delta)$ is small.

Janson (1994): Similar approach, but Janson takes Y such that $\mathcal{L}_X = \mathcal{L}_Y$. The above lemma is more flexible.

Proof

- Let I denote the interval $[a, b]$.
- Then note that

$$\begin{aligned}1 &\ge \mathbb{P}(\{X \in I\} \cup \{Y \in I\}) \\ &= \mathbb{P}(X \in I) + \mathbb{P}(Y \in I) - \mathbb{P}(\{X \in I\} \cap \{Y \in I\}).\end{aligned}$$

- But

$$\mathbb{P}(Y \in I) \ge \mathbb{P}(X \in I) - d_{\mathrm{TV}}(\mathcal{L}_X, \mathcal{L}_Y),$$

and

$$\mathbb{P}(\{X \in I\} \cap \{Y \in I\}) \le \mathbb{P}(|X - Y| \le b - a).$$

- Thus,

$$1 \ge 2\mathbb{P}(X \in I) - d_{\mathrm{TV}}(\mathcal{L}_X, \mathcal{L}_Y) - \mathbb{P}(|X - Y| \le b - a),$$

completing the proof.

A simple example

- Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(1/2) random variables and let $S_n = X_1 + \cdots + X_n$.
- Let $X_i' = 1$ with probability $\alpha n^{-1/2}$, and $X_i' = X_i$ with probability $1 - \alpha n^{-1/2}$. Let $S_n' = X_1' + \cdots + X_n'$.
- Direct calculation shows that

$$d_{\mathrm{TV}}(\mathcal{L}_{S_n}, \mathcal{L}_{S_n'}) \le d_{\mathrm{TV}}(\mathcal{L}_{(X_1,\ldots,X_n)}, \mathcal{L}_{(X_1',\ldots,X_n')}) \le C_1\alpha.$$

- But $S_n' - S_n > C_2 \alpha n^{1/2}$ with high probability.
- Thus, for any interval I of width $\le \delta_n = C_2\alpha n^{1/2}$,

$$\mathbb{P}(S_n \in I) \le \frac{1}{2}(1 + C_1\alpha + \mathbb{P}(|S_n' - S_n| \le \delta_n)).$$

- Choosing α small enough, the right side can be made uniformly bounded away from 1. This shows that S_n has fluctuations of order at least $n^{1/2}$.

Hellinger affinity

- Let us now see how to get an optimal lower bound on the fluctuations of the length of the optimal tour in the traveling salesman problem.
- Let μ and μ' be two probability measures on some space, having densities f and g with respect to some probability measure ν.
- The Hellinger affinity between μ and μ' is defined as

$$\rho(\mu, \mu') := \int \sqrt{fg}\, d\nu.$$

- This quantity does not depend on the choice of ν.

Total variation distance between product measures

- Let $\mu_1, \ldots, \mu_n, \mu'_1, \ldots, \mu'_n$ be probability measures on some space.
- Let $\mu = \mu_1 \times \cdots \times \mu_n$ and $\mu' = \mu'_1 \times \cdots \times \mu'_n$.
- The following bound is well-known and widely used in mathematical statistics:

$$d_{\mathrm{TV}}(\mu, \mu') \le \sqrt{1 - \prod_{i=1}^{n} \rho(\mu_i, \mu'_i)^2}.$$

Multiplicative perturbation

- Let X be a d-dimensional random vector with probability density function $e^{-V(x)}$ on either $\mathbb{R}^d$ or $[0, \infty)^d$.
- Here V is a smooth function satisfying some mild growth conditions (e.g., X may be Gaussian or exponential, but not uniform).
- Take some $\epsilon \in (-1/2, 1/2)$ and let $Y = X/(1 + \epsilon)$.

Hellinger affinity for multiplicative perturbation

Using integration by parts,

$$\rho(\mathcal{L}_X, \mathcal{L}_Y) = \int \sqrt{(1+\epsilon)^d e^{-V(x+\epsilon x) - V(x)}} dx$$
$$= \left(1 + \frac{\epsilon d}{2}\right) \int \left(1 - \frac{\epsilon}{2} x \cdot \nabla V(x)\right) e^{-V(x)} dx + O(\epsilon^2)$$
$$= \left(1 + \frac{\epsilon d}{2}\right)\left(1 - \frac{\epsilon d}{2}\right) + O(\epsilon^2) \geq 1 - C\epsilon^2.$$

Proposition (C., 2017)

If $X_1, \ldots, X_n$ are i.i.d. with density as above and $Y_i = X_i/(1+\epsilon_i)$, then

$$d_{\mathrm{TV}}(\mathcal{L}_{(X_1,\ldots,X_n)}, \mathcal{L}_{(Y_1,\ldots,Y_n)}) \leq C \sqrt{\sum_{i=1}^n \epsilon_i^2},$$

where C depends only on $\mathcal{L}_{X_1}$.

Lower bounds in geometric optimization

- Let $f_n : (\mathbb{R}^d)^n \to \mathbb{R}$ be a function and $r > 0$ be a constant such that for all $\lambda > 0$,

$$f_n(\lambda x_1, \ldots, \lambda x_n) = \lambda^r f_n(x_1, \ldots, x_n).$$

- For example, length of optimal traveling salesman path through $x_1, \ldots, x_n$, length of minimum matching, volume of convex hull, etc.
- Let X_i be as in the previous slide, and let $L_n = f_n(X_1, \ldots, X_n)$. What is a lower bound on the order of fluctuations of L_n?

Theorem (C., 2017)

Let t_n be a sequence of constants such that $\liminf \mathbb{P}(L_n > t_n) > 0$. *Then L_n has fluctuations of order at least $n^{-1/2} t_n$.*

Proof

- Let $Y_i = X_i/(1+\alpha n^{-1/2})$ and $L'_n = f_n(Y_1, \ldots, Y_n)$.
- Then $L'_n = L_n/(1+\alpha n^{-1/2})^r$.
- If $L_n > t_n$, then $L_n - L'_n > C_1 r\alpha n^{-1/2} t_n$.
- On the other hand, by the Proposition,

$$d_{\mathrm{TV}}(\mathcal{L}_{L_n}, \mathcal{L}_{L'_n}) \le d_{\mathrm{TV}}(\mathcal{L}_{(X_1,\ldots,X_n)}, \mathcal{L}_{(Y_1,\ldots,Y_n)}) \le C_2\alpha.$$

- Thus, by the main lemma, for any interval I of length $\le \delta_n := C_1 r\alpha n^{-1/2} t_n$,

$$\begin{aligned}\mathbb{P}(L_n \in I) &\le \frac{1}{2}(1 + d_{\mathrm{TV}}(\mathcal{L}_{L_n}, \mathcal{L}_{L'_n}) + \mathbb{P}(|L_n - L'_n| \le \delta_n))\\ &\le \frac{1}{2}(1 + C_2\alpha + \mathbb{P}(L_n \le t_n)).\end{aligned}$$

- The proof is completed by choosing α small enough.

Applications

- Let L_n = length of optimal traveling salesman path through $X_1, \ldots, X_n$, or L_n = length of minimum matching. $d \geq 2$.
- Well-known: In both cases, size of L_n is of order $n^{1-1/d}$.
- Thus, the fluctuations are of order at least

$$n^{-1/2} n^{1-1/d} = n^{(d-2)/2d}.$$

- For densities with compact support, it is known that the fluctuations are at most of order $n^{(d-2)/2d}$ (Steele, 1997), matching the above lower bound. However, in my theorem, the densities have unbounded support. I have not seen an upper bound for this case.
- The only previous result is due to Rhee (1994), who proved order 1 lower bound for TSP through uniformly distributed points in $[0,1]^2$.

2D First-passage percolation

- Each edge in $\mathbb{Z}^2$ is assigned a random weight. The weights are nonnegative and i.i.d.
- The weight of a path is the sum of edge weights along the path.
- The first-passage time $T(x, y)$ is the minimum over the weights of all paths from x to y.
- Question: What is the order of fluctuations of $T(x, y)$, depending on the distance $|x - y|$ between x and y?
- Best known upper bound: $\sqrt{|x - y|/\log|x - y|}$. Contributors: Kesten (1993), Benjamini-Kalai-Schramm (2003), Benaïm-Rossignol (2008), Damron-Hanson-Sosoe (2015).

Lower bounds in FPP

- Best known lower bound on variance: Newman & Piza (1995) showed that $\mathrm{Var}(T(x,y)) \geq C \log|x-y|$.
- However, this does not prove a lower bound on the order of fluctuations, since the upper bound does not match.
- Pemantle & Peres (1994) proved an actual lower bound of order $\sqrt{\log|x-y|}$, but only if the weights are exponentially distributed.
- The Pemantle-Peres proof uses the memoryless property of the exponential distribution, and does not seem to extend easily to other distributions.

Theorem (C., 2017)

For 2D first-passage percolation, under mild smoothness and decay assumptions on the edge weight distribution, the fluctuations of $T(x,y)$ are at least of order $\sqrt{\log|x-y|}$.

Proof sketch

- Let $n = |x - y|$.
- In a ball of radius $n/2$ around x, replace each edge weight ω_e by $\omega_e/(1+\epsilon_e)$, where

$$\epsilon_e = \frac{\alpha}{\mathrm{dist}(e,x)\sqrt{\log n}}.$$

- Let T and T' be the first-passage times from x to y in the two environments.
- Then, one can show that $T - T' \geq C\alpha\sqrt{\log n}$ with high probability, and $d_{\mathrm{TV}}(\mathcal{L}_T, \mathcal{L}_{T'}) \leq C\alpha$.
- Proof is completed by choosing α sufficiently small and applying the main lemma.

Shape fluctuations in FPP

- Let $B(t)$ be the set of all vertices x such that $T(0,x) \leq t$.
- Cox & Durrett (1981) proved that there exists a symmetric convex set B_0 such that almost surely, for all $\epsilon > 0$,

$$(1-\epsilon)B_0 \subseteq \frac{1}{t}B(t) \subseteq (1+\epsilon)B_0 \quad \text{for all large } t.$$

- B_0 is called the limit shape and fluctuations of $\frac{1}{t}B(t)$ are called shape fluctuations.
- Newman & Piza (1995) defined the natural shape fluctuation exponent

$$\chi' := \inf\{\kappa : (t - t^\kappa)B_0 \subseteq B(t) \subseteq (t + t^\kappa)B_0 \\ \text{for all large } t \text{ a.s.}\}.$$

- It has been an open problem until now to show that $\chi' > 0$ (in any dimension).

A lower bound on the shape fluctuation exponent in 2D

Theorem (C., 2017)

In 2D first-passage percolation, under mild conditions on the edge weight distribution, $\chi' \geq 1/8$.

- The main step is to show that there is a direction in which the first-passage time to a point at distance n has fluctuations of order at least $n^{1/8}$.
- Newman & Piza (1995) proved a corresponding lower bound for the variance, but since that does not imply a lower bound on the order of fluctuations, it could not be used to deduce a lower bound on χ'.
- It is conjectured that $\chi' = 1/3$.

Random assignment problem

- n tasks, to be assigned to n workers.
- a_{ij} is the cost of assigning task j to worker i.
- The minimum cost of assigning tasks is

$$C_n = \min_{\pi \in S_n} \sum_{i=1}^{n} a_{i\pi(i)}.$$

- Random assignment problem: a_{ij} are i.i.d. nonnegative random variables.
- Let f be the probability density function of a_{ij}. If f is continuous and $f(0) = 1$, Aldous (2001) proved that $C_n \to \pi^2/6$ in probability, resolving a conjecture of Mézard & Parisi (1985).
- Later generalized and extended by Linusson-Wästlund (2004), Nair-Prabhakar-Sharma (2005), Wästlund (2005, 2010, 2012).

Fluctuations of the minimum cost

- General cost distributions on $[0, 1]$: Best known upper bound on fluctuations is due to Talagrand (1995), of order $(\log n)^2/(\sqrt{n} \log \log n)$.
- Much more can be done if we assume that a_{ij} are exponentially distributed with mean 1.
- Under this assumption, Alm & Sorkin (2002) showed that $\mathrm{Var}(C_n) \geq c/n$, and Wästlund (2005, 2010) proved that

$$\mathrm{Var}(C_n) = \frac{4\zeta(2) - 4\zeta(3)}{n} + O\left(\frac{1}{n^2}\right).$$

- However, the question about the order of fluctuations is not settled for general cost distributions.

Theorem (C., 2017)

If a_{ij} has a density that is smooth, bounded at zero, and satisfies some mild decay conditions, then the fluctuations of C_n are at least of order $n^{-1/2}$.

A remark about the proof

- A simple multiplicative perturbation does not suffice for this problem.
- Instead, a_{ij} needs to be replaced by a'_{ij}, where a'_{ij} solves

$$a'_{ij} + \alpha n^{-1} \phi_n(a'_{ij}) = a_{ij},$$

  where ϕ_n is the continuous function satisfying $\phi_n(0) = 0$, $\phi'_n(x) = \sqrt{n}$ for $0 < x < 1/n$, and $\phi'_n(x) = 1$ for $x > 1/n$.
- Then the proof proceeds as in other examples.

Determinants of random matrices

Theorem (C., 2017)

Let M be a random square matrix of order N, which is a function of i.i.d. random variables $X_1, \ldots, X_n$. Assume that this function is homogeneous of degree r, and that the law of X_1 has a smooth density satisfying some mild decay conditions. If n and N tend to infinity while r remains fixed, then $\log|\det M|$ has fluctuations of order at least $n^{-1/2}N$.

Example:

- Let X be a $p \times n$ random matrix of with i.i.d. entries, X_0 be the matrix obtained by subtracting off the row mean from each row of X, and let $M = \frac{1}{n}X_0X_0^T$ be the sample covariance matrix for the data matrix X.
- Then the theorem says that $\log|\det M|$ has fluctuations of order at least $(np)^{-1/2}p = \sqrt{p/n}$.
- This matches the order recently obtained by Cai, Liang and Zhou (2015) for the Gaussian case when $p/n \to c \in [0, 1)$.

Free energy of the SK model

- Let $(g_{ij})_{1\le i<j\le n}$ be i.i.d. $N(0,1)$ random variables.
- The free energy of the Sherrington–Kirkpatrick model of spin glasses at inverse temperature $\beta > 0$ is given by

$$F_n(\beta) = \log \sum_{\sigma \in \{-1,1\}^n} \exp\biggl(\frac{\beta}{\sqrt{n}} \sum_{1\le i<j\le n} g_{ij}\sigma_i\sigma_j\biggr).$$

- The best known upper bound on $\mathrm{Var}(F_n(\beta))$ is $O(n/\log n)$ (C., 2009).
- When $\beta < 1$, Aizenman, Lebowitz and Ruelle (1987) proved that $F_n(\beta)$ has fluctuations of order 1 and satisfies a central limit theorem after centering.

Theorem (C., 2017)

For any β, $F_n(\beta)$ has fluctuations of order at least 1 as $n \to \infty$.

Open problems

- Prove a tight lower bound for the fluctuations of the length of the minimum matching when the points are uniformly distributed in $[0, 1]^d$.
- Prove a tight lower bound for fluctuations in the longest common subsequence problem for random words. A solution of this problem would complete the proof of the central limit theorem for longest common subsequences, as shown by Houdré & Işlak (2014).
- Improve the lower bound for the fluctuations of the first-passage time in 2D first-passage percolation.
- Prove any nontrivial lower bound for the fluctuations of the first-passage time in higher dimensions.
- Prove a matching upper bound of order $n^{-1/2}$ for random assignment with general cost distribution.
- Prove a CLT in any of these problems.
- Many other open problems — see paper on arXiv.