Concentration Techniques

Maria Serna

UPC, Barcelona

Slide 1

- Basics
- Tail bounds for a random varuiable.
- Sums of bounded independent variables.
- Indepedent bounded differences method.
- Tail bounds for dependent variables.
- Isoperimetric inequalities.

Slide 2

Basics:

Given a sample space Ω, an event is any subset of Ω. A probability measure $\mathbf{Pr}\colon \mathcal{A} \to \mathbb{R}$ is a function such that: for any $E \in \mathcal{A}$, $0 \leq \mathbf{Pr}[E] \leq 1$; $\mathbf{Pr}[\Omega] = 1$, and $\mathbf{Pr}[\cup_{n\geq 1} E_n] = \sum_{n\geq 1} \mathbf{Pr}[E_n]$ for any sequence $(E_n)_{n\geq 1}$ of events in $\mathcal{A}$ such that $E_n \cap E_m = \emptyset$ for $n \neq m$. Then $(\Omega, \mathbf{Pr})$ is a probability space.

Proposition (Boole's inequalities):

Slide 3 Let $E_1, \dots, E_n$ be a collection of events; then

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}[E_i] \quad \text{and} \quad \mathbf{Pr}\left[\bigcap_{i=1}^{n} E_i\right] \geq 1 - \sum_{i=1}^{n} \mathbf{Pr}\left[\overline{E}_i\right].$$

Given a random variable X, on $(\Omega, \mathbf{Pr})$ its

- its expectation: $\mathbf{E}[X] = \sum_{x\in X(\Omega)} x\,\mathbf{Pr}[X = x]$,
- its variance: $\mathbf{Var}[X] = \mathbf{E}\left[(X - \mathbf{E}[X])^2\right]$,
- its standard deviation: $\sigma_X = \sqrt{\mathbf{Var}[X]}$.
- For $k \in \mathbb{N}$ the k-th moment of X is $\mathbf{E}\left[X^k\right]$.

Proposition (Linearity of the expectation):

Slide 4 For any rv $X_1, \dots, X_n$ and any $c \in \mathbb{R}$,

$$\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i],$$
$$\mathbf{E}[cX_i] = c\,\mathbf{E}[X_i].$$

The Probabilistic Method

The basic idea behind the probabilistic method is to show that a certain property holds for some combinatorial structure. To do so, one constructs an appropriate probability space, and shows that the property has non-zero probability on it.

P. Erdös: "Some remarks on the theory of graphs". *Bull. Ammerican Math. Society*, **53**, 292–294, 1947.

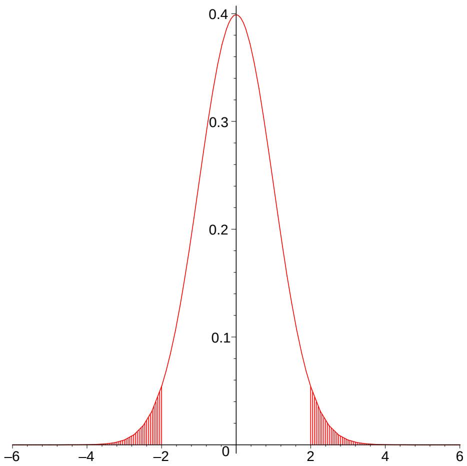
Slide 5

Proposition (First Moment Principle):
If $\mathbf{E}[X_i] \leq t$ then $\mathbf{Pr}[X \leq t] = 0$.

Concentration bounds are a useful tool to prove that the probability of a sequence of events becomes very small.

Basic tail bounds for a random variable

Slide 6



Given a sequence of random variables $X_1, X_2, \ldots$ and a random variable X, X_n converges in distribution to X if $\mathbf{Pr}[X_n < x] \to \mathbf{Pr}[X < x]$ as $n \to \infty$ for all x.

Theorem (Law of large numbers):
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite expectation μ. For each $n \geq 1$, set $S_n = \sum_{i=1}^{n} X_i$. Then, S_n/n converges in distribution to μ.

Slide 7

Theorem (Central limit theorem):
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed random variables with finite expectation μ and finite variance σ^2. For each $n \geq 1$, set $S_n = \sum_{i=1}^{n} X_i$. Then, $(S_n - n\mu)/\sigma\sqrt{n}$ converges in distribution to $N(0,1)$.

Theorem (de Moivre–Laplace limit theorem):
Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed Bernoulli random variables with parameter p. For each $n \geq 1$, set $S_n = \sum_{i=1}^{n} X_i$. Then, for any fixed constants $c_1 < c_2$, as $n \to \infty$,

$$\mathbf{Pr}\left[np + c_1\sqrt{npq} \leq S_n \leq np + c_2\sqrt{npq}\right] \to \Phi(c_2) - \Phi(c_1),$$

where $\Phi(x)$ denotes the distribution function of $N(0,1)$:

Slide 8

$$\Phi(x) = \mathbf{Pr}\left[N(0,1) \leq x\right] = \int_{-\infty}^{x} \tfrac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}u)\mathrm{d}u.$$

Theorem: Let $t > 0$ and let X be a random variable, defined on a probability space Ω. Then, for any integer $k > 0$,

$$\mathbf{Pr}[|X| > t] \leq \frac{\mathbf{E}\left[|X|^k\right]}{t^k}.$$

Proof.

Slide 9

$$\begin{aligned}
\mathbf{E}\left[|X|^k\right] &= \sum_{x \in X(\Omega)} |x|^k \mathbf{Pr}[X = x] \\
&= \sum_{|x| \leq t} |x|^k \mathbf{Pr}[X = x] + \sum_{|x| > t} |x|^k \mathbf{Pr}[X = x] \\
&\geq \sum_{|x| > t} |x|^k \mathbf{Pr}[X = x] \\
&\geq t^k \sum_{|x| > t} \mathbf{Pr}[X = x] \\
&= t^k \mathbf{Pr}[|X| > t].
\end{aligned}$$

□

Corollary (Markov inequality):
Let X be a positive random variable. Then, for any $t > 0$,

$$\mathbf{Pr}[X > t\mathbf{E}[X]] \leq 1/t.$$

Corollary (Chebyshov inequality):
Let X be a random variable with expected value μ and standard deviation σ. Then, for all $t > 0$,

Slide 10

$$\mathbf{Pr}[|X - \mu| > t\sigma] \leq 1/t.$$

Example for basic concentrations

Let us study the concentration of the random variable that counts the number of heads after tossing n times a fair coin. Let $X_1, \dots, X_n$ be the random variables produced by the outcome of each toss. We wish to study the concentration of $S_n = \sum_{i=1}^n X_i$. Recall that $\mathbf{E}[S_n] = n/2$ and $\mathbf{Var}[S_n] = \frac{1}{4}n$.

If we apply Markov, we get

$$\mathbf{Pr}[S_n > (1+\epsilon)\mathbf{E}[S_n]] \leq \frac{1}{1+\epsilon},$$

Slide 11

If we apply Chebyshov inequality, we get

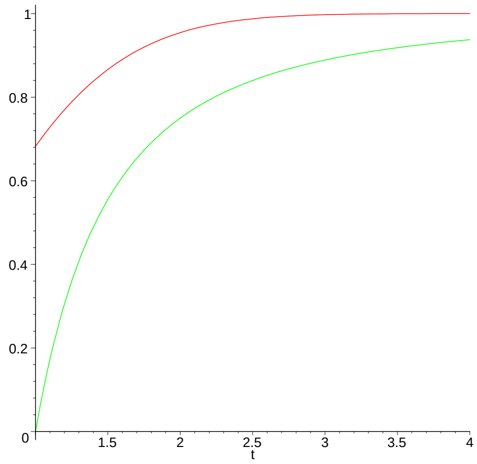
$$\mathbf{Pr}[|S_n - \mathbf{E}[S_n]| > t\sigma] < 1/t^2, \qquad \forall t > 0.$$

If we apply de Moivre–Laplace, then

$$\mathbf{Pr}[|S_n - \mu| \leq t\sigma] \to \Phi(t) - \Phi(-t), \qquad \forall t > 0.$$

Comparing the two bounds, we can observe that, asymptotically, we get a better concentration bound for S_n around its expectation using the de Moivre–Laplace. But Chebyshov inequality is valid for all n, while de Moivre–Laplace Theorem only applies in the limit.

Slide 12



An example for the First Moment Method: k-SAT

Given a set X of n boolean variables and a boolean formula F on those variables, in Conjuctive Normal Form (CNF), where each formula has exactly k literals, F is satisfiable if there is an assignment $T : X \to \{0,1\}$ such that $T(F) = 1$.

$$\begin{aligned} F =&((x_1 \vee x_2 \vee x_3) \wedge (\bar{x}_1 \vee x_2 \vee x_3) \wedge (x_1 \vee \bar{x}_2 \vee x_3) \\ &\wedge (\bar{x}_1 \vee \bar{x}_2 \vee x_3) \wedge (x_1 \vee \bar{x}_2 \vee \bar{x}_3) \wedge (\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3). \end{aligned}$$

Slide 13

F is sat: $T(x_3) = 1, T(x_2) = 0$.

Formula NO sat:

$$\begin{aligned} F =&((x_1 \vee x_2 \vee x_3) \wedge (\bar{x}_1 \vee x_2 \vee x_3) \wedge (x_1 \vee \bar{x}_2 \vee x_3) \wedge (x_1 \vee x_2 \vee \bar{x}_3) \\ &\wedge (\bar{x}_1 \vee \bar{x}_2 \vee x_3) \wedge (\bar{x}_1 \vee x_2 \vee \bar{x}_3) \wedge (x_1 \vee \bar{x}_2 \vee \bar{x}_3) \wedge (\bar{x}_1 \vee \bar{x}_2 \vee \bar{x}_3). \end{aligned}$$

Theorem: Any instance of k-SAT with fewer than 2^k clauses is satisfiable.

We can state the following algorithm for k-SAT

Given a CNF formula F and an $\epsilon > 0$
such that each clause has at least ϵn variables:
Count the number m of clauses
if $m = 2^{\epsilon n}$ then output F is SAT.
  otherwise do exhaustive search of all 2^n possible truth assignments

The previous algorithm works in polynomial time in the size of the input (why?)

Slide 14

Proof of the Theorem

We have the input $F = C_1 \wedge \ldots \wedge C_m$. Consider the random assignment

$$T(x_i) = \begin{cases} 0 & \text{with } \mathbf{Pr} = 1/2 \\ 1 & \text{with } \mathbf{Pr} = 1/2 \end{cases}$$

Let X be a rv that counts the number of clauses not sat. Wish to prove that $\mathbf{Pr}[X < 1] > 0$, which is equivalent to prove $\mathbf{E}[X] < 1$.

Slide 15

For every C_i, define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if } C_i = 0 \\ 0 & \text{otherwise} \end{cases}$$

As $X = \sum_{i=1}^{m} X_i$ by linearity of expected value, $\mathbf{E}[X] = \sum_{i=1}^{m} \mathbf{E}[X_i]$. But $\mathbf{E}[X_i] = \mathbf{Pr}[C_i = 0] = 2^{-k}$, so $\mathbf{E}[X] = m2^{-k} < 2^k 2^{-k} < 1$.

For every $\pi_i \in S_n$ define the indicator random variable:

$$X_{\pi_i} = \begin{cases} 1 & \text{if } X_{\pi_i} \text{ is a Hamilton path} \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{Pr}[X_{\pi_i} = 1] = \frac{1}{2^{n-1}} \Rightarrow \mathbf{E}[X_{\pi_i}] = \frac{1}{2^{n-1}}$$

Therefore the expected number of Hamilton paths is,

Slide 16

$$\mathbf{E}\left[\sum_{i=1}^{n!} X_{\pi_i}\right] = \sum_{i=1}^{n!} \mathbf{E}[X_{\pi_i}] = \frac{n!}{2^{n-1}}$$

Generation of Random graphs

Given a set $V = \{1, 2, \ldots, n\}$ of nodes, do the following experiment:

Repeat m times
  select an edge $e = \{i, j\} \notin E$ uniformly at random
  add e to E
output $G = (V, E)$.

Slide 17

The sample space are all $\binom{\binom{n}{2}}{m}$ G with n nodes and m edges. Each graph is equally likely.

How many edges are needed to make a random graph on n nodes connected ?.
(Which is the smallest value of m s.t. $\mathbf{Pr}[G \text{ connected}] = 1$?)

Direct approach too difficult !

Given a set $V = \{1, 2, \ldots, n\}$ of nodes, do the following alternative generation:

Repeat until $g = (V, E)$ is connected
  select an edge $e = \{i, j\} \notin E$ uniformly at random
  add e to E
output $G = (V, E)$.

Note the sample space are all connected G. Define the rv $X = |E|$. Wish to compute $\mathbf{E}[X]$.

Slide 18

Theorem: The expected number of edges needed to connect a graph with n nodes is $(n-1)[\gamma + \ln(n-1)]$, where $\gamma = .57721...$

Let the rv X_k count the number of edges needed to reduce the number of components from k to $k-1$. $X = \sum_{k=2}^{n} X_k$ and $\mathbf{E}[X] = \sum_{k=2}^{n} \mathbf{E}[X_k]$.

How to compute X_k?

$X_n = 1$, $X_{n-1} = 1$ but X_{n-2} starts to become complicated, it could decrement or not, the number of connected component We compute an upper bound.

Slide 19

If G has k components. Consider any node i. The experiment is equally likely to pick any $\{i,j\} \notin E$. There are $n-1$ edges, al least $k-1$ of them reduce the number of edges. So probability that one of them reduces the number of components (success) $\geq \frac{k-1}{n-1}$ and probability of failure $\leq 1 - \frac{k-1}{n-1}$. Therefore X_k has a geometric distribution,

$$\mathbf{E}[X_k] \leq \frac{1}{\frac{k-1}{n-1}} = \frac{n-1}{k-1}$$

$$\Rightarrow \mathbf{E}[X] = \sum_{k=2}^{n} \mathbf{E}[X_k] \leq (n-1)(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}) = (n-1)(\ln(n-1) + \gamma)$$

Slide 20

What's the probability that we need much more than that ?

Applying Markov's inequality to X, $\mathbf{Pr}[X > c\mu] \leq 1/c$ for any $c \geq 1$. For example, if we add 10μ the probability G connecte $\geq 9/10$.

Modify the last experiment so if G becomes connected before adding m edges, we keep adding random edges until G has exactly m edges. View each point in the random space generated by new experiment: $G = G' + G''$ where G' is exactly the same sample space as in the first experiment. So,

$$\mathbf{Pr}_1[G \text{ connected}] = \mathbf{Pr}_2[G' \text{ connected}] = \mathbf{Pr}_2[X \leq c\mu] \geq 1 - \frac{1}{c}$$

Therefore, A random graph as generated in experiment 1, with $m = c(n-1)(\ln(n-1) + \gamma)$ edges is connected with probability $\geq 1 - \frac{1}{c}$.

Concentration inequalities for the sum of bounded independent random variables.

Assuming that $X_1, \ldots, X_n$ are independent identical distributed Bernoulli random variables, with probability of success p, an application of Chebyshov inequality gives

$$\mathbf{Pr}\left[\sqrt{\frac{n}{p(1-p)}}\left|\frac{1}{n}\sum_{i=1}^{n} X_i - p\right| > t\right] \leq \frac{1}{t^2}.$$

Slide 21 By the central limit theorem, we know that $\sum_{i=1}^{n} X_i$ approaches the normal distribution; so one may expect that

$$\mathbf{Pr}\left[\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - p\right) > t\right] \longrightarrow 1 - \Phi(t) \leq \frac{1}{\sqrt{2\pi}}\frac{e^{-t^2/2}}{t},$$

showing that the bound given by using Chebyshov inequality is far off from what one should see asymptotically, according to the central limit theorem.

Theorem (General Chernoff's bounds):
Let $X_1, \ldots, X_n$ be independent random variables corresponding to Bernoulli experiments, each variable with probability p_i of success and probability $q_i = 1 - p_i$ of failure. Let $X = \sum_{i=1}^{n} X_i$ be a random variable with expectation $\mu = \sum_{i=1}^{n} p_i$. Then, for any $d \in (0, 1)$,

$$\mathbf{Pr}[X < (1-d)\mu] \leq e^{-d^2\mu/2} \quad \text{and} \quad \mathbf{Pr}[X > (1+d)\mu] \leq e^{-d^2\mu/3}.$$

Therefore,

Slide 22

$$\mathbf{Pr}[|X - \mu| > d\mu] \leq 2e^{-d^2\mu/3} \tag{1}$$

Points into boxes

Consider a collection of n points in independently and uniformly distributed in $([0,1]^2)$. Divide the unit square in $n/\log^n n$ squares boxes, each of side $\log n/\sqrt{n}$. Given $\epsilon \in (0,1)$, we say that a the collection of points is ϵ-nice if each box contains at least $(1-\epsilon)\log^2 n$ points and at most $(1+\epsilon)\log^2 n$ points. We would like to prove that, with high probability, the given points are ϵ-nice.

Slide 23

Let X be a random variable counting the number of points in a fixed box. Then X follows a binomial $\mathrm{BIN}(n,p)$ with $p = \log^2 n/n$ so, $\mu = \mathbf{E}[X] = np = \log^2 n$. Using Markov inequality,

$$\mathbf{Pr}[X > (1+\epsilon)\mu] \le \frac{1}{1+\epsilon},$$

which is independent of n. (With that bound we cannot even say that a single fixed box contains less than $(1+\epsilon)\mu$ points with high probability!)

Slide 24

Using Chernoff's bounds, we have

$$\mathbf{Pr}[|X-\mu| > \epsilon\mu] \le 2\exp(-\epsilon^2\mu/3) = 2n^{-\epsilon^2 \log n/3}$$

which tends to 0 as $n \to \infty$. Therefore, we have that, with high probability, a fixed box has at least $(1-\epsilon)\log^2 n$ points and at most $(1+\epsilon)\log^2 n$ points.

Using Boole's inequality, we have now that the probability that some box has less than $(1-\epsilon)\log^2 n$ points or more than $(1+\epsilon)\log^2 n$ points is less than

$$\frac{n}{\log^2 n} 2n^{-\epsilon^2 \log n/3} \le 2n^{1-\epsilon^2 \log n/3}$$

which tends to 0. As a consequence, for any fixed ϵ, with high probability, a random configuration of points in the unit square is ϵ-nice.

Mating lemma [Vishkin, 1984]

Let L be a sequence of n individuals, where each one can be male (M) or female (F) with probability $\frac{1}{2}$ and independently of the sex of the other members. Let us construct the subsequence S of L in the following way: an individual x of L does not belong to S when it is male and its immediate predecessor is female. We wish to prove that, with overwhelming probability, $|S| < \frac{15}{16}n$.

Let $x \in L$ be an element in an even position and let y be its predecessor in L. Then,

Slide 25

$$\mathbf{Pr}[x \notin S] = \mathbf{Pr}[x = M \wedge y = F] = \mathbf{Pr}[x = M]\mathbf{Pr}[y = F] = \tfrac{1}{4}.$$

Therefore, there are $n/2$ candidates to be excluded from S with probability $\frac{1}{4}$.

Let X be the random variable that counts the number of excluded elements. Then, $\mu = \mathbf{E}[X] \geq \frac{1}{8}n$, and using Chernoff bounds with $d = 1/2$, we get

$$\begin{aligned}\mathbf{Pr}\left[X > \tfrac{1}{16}n\right] = \mathbf{Pr}\left[X > (1-1/2)\tfrac{1}{8}n\right] &\leq \mathbf{Pr}\left[X > (1-d)\mu\right] \\ &\leq e^{-d^2\mu/2} \leq e^{-n/64}.\end{aligned}$$

This means

Slide 26

$$\mathbf{Pr}[|S| > \tfrac{15}{16}n] > 1 - e^{-n/64},$$

which implies that, with overwhelming probability, $|S| < \frac{15}{16}n$.

Theorem (Hoeffding's inequalities): Let $X_1, \ldots, X_n$ be independent random variables such that, for all $i \in [n]$, there are values a_i and b_i such that $a_i \leq X_i \leq b_i$. Let $X = \sum_{i=1}^n X_i$ and $\mu = \mathbf{E}[X]$. Then, for any $t \geq 0$,

$$\mathbf{Pr}[X - \mu > t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right) \quad \text{and}$$
$$\mathbf{Pr}[X - \mu < -t] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Slide 27

Therefore

$$\mathbf{Pr}[|X - \mu| > t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

Sum of random numbers

Let $X_1, \ldots, X_n$ be independent random variables, all of them uniformly distributed in the interval $[0, 1]$.

Let $X = \sum_{i=1}^n X_i$. As for all $i \in [n]$, $\mathbf{E}[X_i] = 1/2$, then by linearity of the expectation, $\mathbf{E}[X] = 1/2n$.

We can apply Hoeffding's inequalities with $a_i = 0$ and $b_i = 1$ for all $i \in [n]$:

$$\mathbf{Pr}[|X - \mathbf{E}[X]| > t] \leq 2e^{-2t^2/\sum_{i=1}^n (b_i - a_i)^2} \leq 2e^{-2t^2/n}.$$

Slide 28

This probability goes to zero when $t = b(n)\sqrt{n}$ where $b(n) \to \infty$ as $n \to \infty$.

For a numeric setting, take $n = 200$. Then, the probability that $90 \leq X \leq 110$ is greater than $2/e \approx 0.735$.

The independent bounded differences method

Obtaining bounds on the tails of a function defined on a product space requires that the function satisfies the so-called Lipschitz or independent bounded differences condition.

Definition (Independent bounded differences condition):
A function $f: \mathbb{R}^n \to \mathbb{R}$ satisfies the independent bounded differences condition or is said to be Lipschitz, if there exists a vector $\mathbf{c} = (c_1, \ldots, c_n)$ such that

Slide 29

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leq c_i$$

for all vectors $\mathbf{x}$ and $\mathbf{x}'$ that differ only at the i-th coordinate ($i \in [n]$).

Coin tosses

We want to count the number of heads when tossing n times a fair coin.

The corresponding probability space is $\Omega = \{0,1\}^n$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vector where, for each $i \in [n]$, the probability that the i-th toss is a head is $\mathbf{Pr}[X_i = 1] = \frac{1}{2}$.

Define $f: \Omega \to \mathbb{R}$ assigning to any $\mathbf{x} = (x_1, \ldots, x_n)$ the value $f(\mathbf{x}) = \sum_{i=1}^{n} x_i$, so that f is a random variable counting the number of heads. Then, f is Lipschitz, as if $\mathbf{x}$ and $\mathbf{x}'$ only differ at the i-th coordinate, this means that the number of heads in $\mathbf{x}$ and $\mathbf{x}'$ differs only by one, i.e. $|f(\mathbf{x}) - f(\mathbf{x}')| \leq 1$.

Slide 30

Theorem (McDiarmid's bound): Given a product space defined by $\Omega = \prod_{i=1}^{n} \Omega_i$, let $X_1, \ldots, X_n$ be independent random variables on $\Omega_1, \ldots, \Omega_n$. Assume $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the independent bounded differences condition with $\mathbf{c} = (c_1, \ldots, c_n)$. Then, for all $t > 0$,

$$\mathbf{Pr}\left[f(X_1, \ldots, X_n) - \mathbf{E}\left[f(X_1, \ldots, X_n)\right] > t\right] < \exp\left(-\frac{t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Slide 31 Observe that the inequality in the above theorem is only one-sided, but applying it to the function $-f$ we get

$$\mathbf{Pr}\left[f(X_1, \ldots, X_n) - \mathbf{E}\left[f(X_1, \ldots, X_n)\right] < -t\right] < e^{-t^2/\sum_{i=1}^{n} c_i^2},$$

and therefore,

$$\mathbf{Pr}\left[\left|f(X_1, \ldots, X_n) - \mathbf{E}\left[f(X_1, \ldots, X_n)\right]\right| < t\right] < 2e^{-t^2/\sum_{i=1}^{n} c_i^2}.$$

Coin tosses

In the example of counting the number of heads in n tosses of a fair coin, f is Lipschitz with $\mathbf{c} = \mathbf{1}$. For any $t > 0$, using McDiarmid's bound we get

$$\mathbf{Pr}\left[f(\mathbf{X}) - n/2 > t\right] \leq e^{-t^2/n}.$$

and using Chernoff's bounds (with $t = d\mu$) we get

$$\mathbf{Pr}\left[f(\mathbf{X}) - n/2 > t\right] \leq e^{-\frac{2}{3}t^2/n}.$$

Slide 32 In this case, both bounds are equivalent up to the constants.

Balls and bins [Dubhashi 97]

Consider the experiment: throw independently and uniformly m balls into n bins. We wish to count the number of empty bins.

In this case, $\Omega = [n]^m$. Let Z be a random variable counting the number of empty bins.

Let $X_1, \ldots, X_m$ be a sequence of independent random variables, where each X_i takes values in $[n]$ and indicates the bin where the i-th ball falls, then

Slide 33

$$\mu = \mathbf{E}[Z] = n\left(1 - \frac{1}{n}\right)^m .$$

Z is Lipschitz, if the i-th ball is moved from one bin to another, keeping all other balls in the same position, the number of empty bins can increase or decrease at most by 1. So, applying McDiarmid's bound, we get that, for any $t > 0$,

$$\mathbf{Pr}[|Z - \mu| > t] \leq 2e^{-t^2/m}. \tag{2}$$

A concentration bound tending to zero as $m \to \infty$ is obtained when t grows faster than $\sqrt{m}$.

Bin packing [Karp 89]

The BinPacking problem is another classical **NP**-hard problem:

Given n objects with sizes $x_1, \ldots, x_n$ (such that $0 < x_i < 1$ for all $i \in [n]$) and an unlimited collection of bins, each of size 1, what is the minimal number of bins required to pack the objects?

In the randomized version of this problem, the sizes of the objects are generated uniformly and independently at random.

Slide 34

Given $X_1, X_2, \ldots$, define $Y_n = f(X_1, \ldots, X_n)$ as the minimal number of bins required to pack the first n items.

It has been shown that Y_n grows linearly in n, that is, there exists a constant β such that

$$\frac{\mathbf{E}[Y_n]}{n} \to \beta, \qquad \text{as } n \to \infty.$$

Observe that f is Lipschitz with $\mathbf{c} = \mathbf{1}$. From the independent bounded differences method:

Slide 35

$$\mathbf{Pr}[|Y_n - \mathbf{E}[Y_n]| > t] < 2e^{-t^2/n}.$$

Therefore, with high probability $|Y_n - \mathbf{E}[Y_n]| > b(n)\sqrt{n}$ where $b(n)$ is any function so that $b(n) \to \infty$.

Longest increasing subsequence [Talagrand 95]

A permutation π of $[n]$ induces a sequence with a reordering of the elements $\{1, \ldots, n\}$, an increasing subsequence in π is a subsequence in which each element is smaller than the next.

Let L_π denote the length of the longest increasing subsequence in π.

We are interested in the random variable L_π when the permutation π is random with uniform distribution.

Slide 36

Asymptotically, $\mathbf{E}[L_\pi] = 2\sqrt{n}$. Using the Hoeffding-Azuma inequality, Frieze [Frieze 91] gave a concentration bound for L_π within $(2\sqrt{n})^{2/3}$ of its expected value. His proof is non trivial.

But, changing a bit the setting, we can apply the independent bounded differences method.

Consider n uniformly and independently chosen real random numbers in $[0,1]$, $X_1,\ldots,X_n$.

With probability 1, all of them would be different, therefore the order on the reals induces a permutation π of $[n]$.

Let L_n be the random variable measuring the length of the longest increasing sequence on these reals. Notice that by changing only one number, the length of the longest increasing sequence may be increased or decreased by one, therefore, for any $t \geq 0$

Slide 37

$$\mathbf{Pr}\left[|L_n - \mathbf{E}\left[L_n\right]| > t\right] \leq e^{-2t^2/n}.$$

Better concentration bounds will be given later.

Tail bounds for dependent variables

Martingale theory is a powerful tool to derive concentration results, when the random variables under consideration are not necessarily independent.

The study of martingales is an old topic in probability theory and provides one of the most useful tools to establish limit theorems, because the convergence of martingales is guaranteed.

We will consider only finite martingale sequences, as we are not interested in convergence results.

Slide 38

Definition: A sequence of random variables $Y_0,\ldots,Y_n$ is a martingale with respect to the sequence $X_1,\ldots,X_n$, if for all $i \in [n]$,

$$\mathbf{E}\left[Y_i | X_1,\ldots,X_{i-1}\right] = Y_{i-1}.$$

The name martingale comes from the old game of martingale:

A gambler has a large amount of money, he puts 1 euro in the first bet. While he loses the bet, he doubles the amount in the next bet. The play continues until the gambler wins a bet.

If the gambler arrives to the n-th play, then he bets 2^n euros on the $n+1$-th bet. So, if he wins at the n-th round, there will be a profit of $2^n - \sum_{i=0}^{n-1} 2^i$, that is just 1 euro!

Slide 39

If X_i denotes the amount that the gambler could win or lose at the i-th hand, as at any play round i, the gambler wins X_i with probability $1/2$ and loses X_i with the same probability, then $\mathbf{E}[X_n|X_1,\dots,X_{n-1}] = X_{n-1}$, for any n.

Independent sums

Slide 40

Let $X_1,\dots,X_n$ be a sequence of bounded independent variables with expected value 0. For instance $X_i = 1$ if the toss of a fair coin results in a head and $X_i = -1$ otherwise. For any i, $0 \leq i \leq n$, define $S_i = \sum_{k=1}^{i} X_k$. Notice that in this example the sum is the function that defines S_n. To prove that $S_0,\dots,S_n$ is a martingale with respect to $X_1,\dots,X_n$, we have to show that for any i, $\mathbf{E}[S_i|X_1,\dots,X_{i-1}] = S_{i-1}$.

$$\begin{aligned}\mathbf{E}[S_i|X_1,\dots,X_{i-1}] &= \mathbf{E}[S_{i-1}+X_i|X_1,\dots,X_{i-1}]\\ &= \mathbf{E}[S_{i-1}|X_1,\dots,X_{i-1}] + \mathbf{E}[X_i|X_1,\dots,X_{i-1}].\end{aligned}$$

As X_i is independent from $X_1,\dots,X_{i-1}$, we get

$$\mathbf{E}[S_i|X_1,\dots,X_{i-1}] = \mathbf{E}[S_{i-1}|X_1,\dots,X_{i-1}] + \mathbf{E}[X_i] = S_{i-1} + 0.$$

Often, it will be the case that we have a martingale with respect to itself.

Definition: A sequence of random variables $Y_0, Y_1, \ldots, Y_n$ is a martingale if for all $i \in [n]$,

$$\mathbf{E}\left[Y_i | Y_1, \ldots, Y_{i-1}\right] = Y_{i-1}.$$

Independent sums

Slide 41

It is easy to show that the process $S_0, \ldots, S_n$ defined as the prefix sums of a sequence of bounded independent variables with mean 0, is a martingale with respect to itself. In fact the events $\mathbf{X}_i = \mathbf{x}_i$ and $\mathbf{S}_i = \mathbf{s}_i$ are in a one to one correspondence.

Another useful form of a martingale is presenting the process as a sequence of differences.

Definition: A process $Y_1, \ldots, Y_n$ is a martingale differences sequence with respect to $X_1, \ldots, X_n$ if for every $i \in [n]$, the random variable Y_i verifies $\mathbf{E}\left[Y_i | X_1, \ldots, X_{i-1}\right] = 0$.

Slide 42

The name of differences comes from the fact that any martingale sequence $Y_0, \ldots, Y_n$ with respect to a sequence $X_1, \ldots, X_n$ provides a natural way to define a sequence of differences, $Z_1, Z_2, \ldots, Z_n$, by setting $Z_i = Y_i - Y_{i-1}$ for any $i \in [n]$. Indeed, the sequence $Z_1, \ldots, Z_n$ is a martingale differences:

$$\mathbf{E}\left[Z_i | X_1, \ldots, X_{i-1}\right] = \mathbf{E}\left[Y_i - Y_{i-1} | X_1, \ldots, X_{i-1}\right]$$

by linearity of expectation and the definition of martingale

$$\begin{aligned}\mathbf{E}\left[Z_i | X_1, \ldots, X_{i-1}\right] &= \mathbf{E}\left[Y_i | X_1, \ldots, X_{i-1}\right] - \mathbf{E}\left[Y_{i-1} | X_1, \ldots, X_{i-1}\right] \\ &= Y_{i-1} - Y_{i-1} = 0.\end{aligned}$$

We will refer to such construction as the martingale differences sequence.

Conversely, from any given random variable X_0 and a martingale differences sequence $Y_1, \dots, Y_n$, we obtain a martingale by setting $X_i = X_0 + \sum_{k=1}^{i} Y_k$. Furthermore, observe that any sequence of independent random variables with mean expected value zero is a martingale differences sequence, because $\mathbf{E}[X_{i+1} | X_i, \dots, X_1] = \mathbf{E}[X_{i+1}] = 0$.

Slide 43

The bounded differences condition

In the case of dependent random variables, that form a martingale, in order to get concentration results, we need a guarantee of bounded increments.

Definition (Bounded differences condition): A martingale sequence $Y_0, \dots, Y_n$ is said to satisfy the *bounded differences condition* if for any $i \in [n]$ there exists a vector $\mathbf{c} = (c_1, \dots, c_n)$ such that

$$|Y_i - Y_{i-1}| \leq c_i.$$

Slide 44

The next theorem says that when the bounded differences condition holds, the process itself does not go far away from its starting point.

Theorem (Hoeffding-Azuma's inequality.): Let $Y_0, \ldots, Y_n$ be a martingale with respect to $X_1, \ldots, X_n$ that satisfies the bounded differences condition with $\mathbf{c} = (c_1, \ldots, c_n)$. Then, for any $t > 0$,

$$\mathbf{Pr}\left[|Y_n - Y_0| > t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right).$$

In the Hoeffding-Azuma's inequality we don't need to make any assumption on the independence of the random variables, furthermore the c_i in the bounded differences condition may depend on i.

Slide 45

In the case that all c_i are equal we get

Corollary: Let $Y_0, \ldots, Y_n$ be a martingale satisfying the bounded differences condition such that for a vector $\mathbf{c} = (c, \ldots, c)$. Then for any $t > 0$

$$\mathbf{Pr}\left[|Y_n - Y_0| \geq ct\sqrt{n}\right] \leq 2\exp\left(-\tfrac{1}{2}t^2\right).$$

Coin tosses

Concentration of the random variable $\mathbf{X}$ that counts the number of heads after tossing a fair coin n times.

Let $Y_0, \ldots, Y_n$, be the Doob's martingale with respect to $X_1, \ldots, X_n$, the indicator variables that record the outcome of each toss.

$$Y_0(\omega) = \mathbf{E}\left[\mathbf{X}\right] = n/2,$$
$$Y_i(\omega) = \mathbf{E}\left[f(\mathbf{X})|X_1, \ldots, X_i\right](\omega)$$

Slide 46

Changing the value of X_i can affect the difference $Y_i - Y_{i-1}$ by at most by 1.

Therefore $|Y_i - Y_{i-1}| \leq 1$ and we get

$$\mathbf{Pr}[|Y_n - \mu| \geq t\sqrt{n}] \leq 2e^{-\frac{t^2}{2}}. \tag{3}$$

To compare with Chernoff bound and the bound obtained from the independent bounded condition, take $t = \frac{d\sqrt{n}}{2}$ so, we get,

Slide 47

$$\mathbf{Pr}\left[|S_n - \mathbf{E}\left[S_n\right]| \geq dn/2\right] \leq \begin{cases} 2\exp(-\frac{1}{8}nd^2) & \text{Hoeffding-Azuma's inequality} \\ 2\exp(-\frac{2}{3}nd^2) & \text{Chernoff inequality} \\ 2\exp(-\frac{1}{4}nd^2) & \text{independent bounded differences} \end{cases}$$

The three bounds give the same behavior, but with different constants in the exponent. The better bound in this case is obtained with Chernoff inequality.

Slide 48

In many situations we will have a set of random variables and a function of them:

Definition: A sequence of random variable $X_1, \ldots, X_n$ and a function $f : \mathbb{R}^n \to \mathbb{R}$ satisfy the bounded differences condition if, for any $i \in [n]$, there is a vector $\mathbf{c} = (c_1, \ldots, c_n)$ such that

$$|\mathbf{E}\left[f(\mathbf{X})|\mathbf{X}_i\right] - \mathbf{E}\left[f(\mathbf{X})|\mathbf{X}_{i-1}\right]| < c_i.$$

This leads to a Hoeffding-Azuma's inequality for the Doob's martingale.

Corollary: Let $X_1, \ldots, X_n$ be a sequence of random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function that satisfies the bounded differences condition for $\mathbf{c} = (c_1, \ldots, c_n)$. Then, for any $t > 0$,

$$\mathbf{Pr}\left[|f(\mathbf{X}) - \mathbf{E}\left[f(\mathbf{X})\right]| > t| \leq 2\exp\left(-\frac{t^2}{2c}\right)\right]$$

where $c = \sum_{i=1}^n c_i^2$.

Chromatic number [Shamir, Spencer 87]

Given a graph G its chromatic number $\chi(G)$ is the minimum number of colors that are needed to color the vertices of G, in such a way that no two adjacent vertices have the same color.

Let us consider the $G_{n,p}$ model and for any graph G, define a Doob's vertex exposure martingale for the function $\chi(G)$:

Assume that the graph is generated by exposing its vertices:

Slide 49 Starting from 1, add a new vertex i and the edges that connect i with the previous $i-1$ vertices.

When adding vertex i, the set of possible end-points of its edges is $[i-1]$, this can be represented by a string of $i-1$ bits.

For each $i \in [n]$ define the random variable X_i uniformly distributed on the set $B_i = \{0,1\}^{i-1}$ of numbers represented with $i-1$ bits.

So, any graph G with n vertices will be specified by a sequence of bit strings, one for each addition of a vertex.

Given any graph theoretical function f, we can consider the Doob's vertex exposure martingale by $Y_i = E[f(G)|\mathbf{X}_i]$.

In this case the random variable Y_i, gives the expected value of $\chi(G)$, when the first i vertices and their corresponding internal edges are known.

Observe that a single additional vertex can increase only by one the number of colors needed to color the graph exposed so far. Therefore, the Doob's vertex exposure martingale for chromatic number satisfies the bounded differences condition, with $c_i = 1$ for any $i \in [n]$.

Slide 50 Therefore taking $t = d\sqrt{n}$, we can apply Corollary 27 to say:

$$\mathbf{Pr}\left[|\chi(G) - \mathbf{E}\left[\chi(G)\right]| > d\sqrt{n}\right] \leq 2e^{-d^2/2}$$

Letting d go arbitrary slowly to ∞ we get a sharp concentration result. However there is no no hint on the value of $\mathbf{E}\left[\chi(G)\right]$.

Isoperimetric inequalities

In graph theoretical terms an isoperimetric inequality is an inequality relating the size of a subgraph with the size of its boundary.

To formalize and generalize the above intuition, let (Ω, d) be a metric space with distance d, together with a measure ν. Recall that for any subset $A \subseteq \Omega$ the distance of a point $\mathbf{x}$ to A is defined as

$$d(\mathbf{x}, A) = \inf\{d(\mathbf{x}, \mathbf{y}) \mid \mathbf{y} \in A)\},$$

Slide 51

and the distance between two subsets A and B of Ω is

$$d(A, B) = \inf\{d(\mathbf{x}, B) \mid \mathbf{x} \in A\}.$$

For a given $A \subseteq \Omega$ and any real $t > 0$, define the t-neighborhood A_t of A as

$$A_t = \{\mathbf{x} \in \Omega \mid d(\mathbf{x}, A) \leq t\}.$$

An isoperimetric inequality is a lower bound of the size $\nu(A_t)$ depending on $\nu(A)$ and t, usually presented as an upper bound on the expression $\nu(A)(\nu(\Omega) - \nu(A_t))$.

Let $G = (V, E)$ be a finite graph; for any $u, v \in V$, let $d(u, v)$ be the shortest distance between u and v in G, and for any $A \subseteq V$, let the measure $\nu(A)$ be the number of vertices in A.

For any A, the set A_t consists of the vertices that can be joined to a vertex in A by a path of length at most t.

An isoperimetric inequality would be an inequality $|A_t| \geq f(|A|, t)$, with $f(|A|, t)$ as close to the actual value of $|A_t|$ as possible.

Slide 52

The boundary ∂ of a subgraph are the set of vertices connected by an edge to some vertex in A.

Therefore, $\partial(A) = A_1 \setminus A$.

Isoperimetric inequalities in a product space under the Hamming distance

Let $\Omega_1, \ldots, \Omega_n$ be probability spaces, and let $\Omega = \prod_{i=1}^n \Omega_i$ be the corresponding product space.

The Hamming distance between any two points $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ in Ω, $d_H(\mathbf{x}, \mathbf{y})$, is the number of indices i such that $x_i \neq y_i$.

Given $\mathbf{x}$ and $\mathbf{y}$ as before, for each $i \in [n]$, define the function $\delta_i(\mathbf{x}, \mathbf{y}) = 1$ if $x_i \neq y_i$ and $\delta_i(\mathbf{x}, \mathbf{y}) = 0$ otherwise.

Slide 53

Thus

$$d_H(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \delta_i(\mathbf{x}, \mathbf{y}).$$

The t-neighborhood of a set A under the Hamming distance will be denoted by A_t^H.

For any $A \in \Omega$ consider the measure $\nu(A) = \mathbf{Pr}[\mathbf{X} \in A]$.

In this case, the independent bounded differences inequality can be used to show that when the probability that a point belongs to a set A is not to small, then with probability tending to 1, any random point is close to A.

Theorem (McDiamird): Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of random variables, each X_i defined on Ω_i. Then for any $t > 0$,

$$\mathbf{Pr}[\mathbf{X} \in A] (1 - \mathbf{Pr}\left[\mathbf{X} \in A_t^H\right]) \leq e^{-t^2/4n}.$$

The above Theorem says that for any subset $A \subseteq \Omega$ such that $\mathbf{Pr}[\mathbf{X} \in A]$ is not too small, say at least $1/2$, with high probability, a random point is close to A under the Hamming distance, that is

Slide 54

$$\mathbf{Pr}\left[\mathbf{X} \in A_t^H\right] \geq 1 - 2e^{-t^2/4n}. \tag{4}$$

Isoperimetry on the Hypercube

Let us consider the n-dimensional hypercube Q_n with vertex set $\{0,1\}^n$ and edges between pairs of vertices whose Hamming distance is 1.

Consider the probability space $\Omega = \prod_{i=1}^{n} \Omega_i$ where $\Omega_i = \{0,1\}$ for all $i \in [n]$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a random vertex of Q_n such that
$\mathbf{Pr}[X_i = 0] = \mathbf{Pr}[X_i = 1] = \frac{1}{2}$ for all $i \in [n]$.

Then, for any set A of Ω, $\mathbf{Pr}[X \in A] = |A|/2^n$.

Slide 55 Thus, if we choose $|A| = 2^{n-1}$ so $\mathbf{Pr}[X \in A] \geq 1/2$, we get the isoperimetric inequality for the hypercube: For any $t \geq 0$,

$$|A_t| \geq 2^n(1 - 2e^{-t^2/4n}). \tag{5}$$

Isoperimetric inequalities in a product space under Talagrand's convex distance

The L_2 norm of a vector $\mathbf{a} = (a_1, \ldots, a_n) \geq \mathbf{0}$, is $\|\mathbf{a}\| = \left(\sum_i^n a_i^2\right)^{1/2}$.

Let $\Omega = \prod_{i=1}^{n} \Omega_i$ be a probability product space, let $\mathbf{X}$ be a sequence of random variables $(X_1, \ldots, X_n)$ in the corresponding spaces Ω_i, and let $\mathbf{a} = (a_1, \ldots, a_n)$ be a point in $[0, \infty)^n$, such that not all its components are 0.

For any two points in Ω, $\mathbf{x} = (x_1, \ldots, x_n)$ and any $\mathbf{y} = (y_1, \ldots, y_n)$, define the distance $d_{\mathbf{a}}$ as:

Slide 56

$$d_{\mathbf{a}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} a_i\, \delta_i(\mathbf{x}, \mathbf{y}).$$

If we consider the specific vector $\mathbf{a} = \mathbf{1} = (1, 1, \ldots, 1)$ then,

$$d_{\mathbf{a}}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \delta_i(\mathbf{x}, \mathbf{y}) = d_H(\mathbf{x}, \mathbf{y}).$$

Given $A \subseteq \Omega$ define $d_{\mathbf{a}}(\mathbf{x}, A)$ as usual.

The Talagrand's convex distance $d_T(\mathbf{x}, A)$ from a point $\mathbf{x}$ to a subset A, is the supremum over all possible n-dimensional vectors $\mathbf{a}$ with $||\mathbf{a}|| = 1$ of $d_{\mathbf{a}}(\mathbf{x}, A)$, that is,

$$d_T(\mathbf{x}, A) = \sup_{||\mathbf{a}||=1} \left\{ \inf_{\mathbf{y} \in A} \left\{ \sum_{i=1}^{n} a_i \, \delta_i(\mathbf{x}, \mathbf{y}) \right\} \right\}.$$

The Talagrand's distance is the best convex distance over all choices of non-negative unit vectors.

Slide 57 Given $A \subseteq \Omega$ and $t > 0$, the Talagrand's t-neighborhood of A, A_t^T, is defined in the usual way:

$$A_t^T = \{\mathbf{x} \in \Omega : d_T(\mathbf{x}, A) \leq t\}.$$

Theorem (Talagrand's isoperimetric inequality): Let Ω be a product space, let $\mathbf{X} = (X_1, \ldots, X_n)$ be a sequence of random variables, with each X_i defined on Ω_i. Then, for any $A \subseteq \Omega$ and any $t \geq 0$,

$$\mathbf{Pr}\,[\mathbf{X} \in A]\left(1 - \mathbf{Pr}\left[\mathbf{X} \in A_t^T\right]\right) \leq e^{-t^2/4}.$$

Talagrand's inequality states that under the Talagrand's convex distance, if $A \subseteq \Omega$ is large enough, then A_t^T covers a large portion of Ω.

Slide 58

If we consider the vector $\mathbf{a}' = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$, then $||\mathbf{a}'|| = 1$, and for any $\mathbf{x} \in \Omega$

$$\begin{aligned} d_T(\mathbf{x}, A) &= \sup_{||\mathbf{a}'||=1} \inf_{\mathbf{y}\in A} \left\{ \sum_{i=1}^{n} a_i \delta_i(\mathbf{x}, \mathbf{y}) \right\} \\ &\geq \inf_{\mathbf{y}\in A} \sum_{i=1}^{n} \frac{1}{\sqrt{n}} \delta_i(\mathbf{x}, \mathbf{y}) \\ &= \inf_{\mathbf{y}\in A} d_H(\mathbf{y}, A) = \frac{1}{\sqrt{n}} d_H(\mathbf{x}, A), \end{aligned}$$

Slide 59

therefore, for any $t > 0$, we get

$$A^T_{t/\sqrt{n}} \subseteq A^H_t. \tag{6}$$

Hypercube again

Consider a random vertex $\mathbf{X} \in \Omega = Q_n$. Given a set $A \subseteq \Omega$ with $\mathbf{Pr}\left[\mathbf{X} \in A\right] \geq 1/2$, and using Talagrand's inequality together with (6), we have

$$\mathbf{Pr}\left[\mathbf{X} \in A^H_t\right] \geq \mathbf{Pr}\left[\mathbf{X} \in A^T_{t/\sqrt{n}}\right] \geq 1 - 2e^{t^2/4n}.$$

Therefore, we obtain a better isoperimetric inequality.

Slide 60

Longest increasing subsequence in a random permutation

As we have seen, the problem is equivalent to the problem of generate independently and uniformly n random reals in $[0,1]$ and compute the length of the longest increasing subsequence.

Let $\Omega = [0,1]^n$, and let $\mathbf{X} \in \Omega$, be a random real sequence $\mathbf{X} = (X_1, \dots, X_n)$. We wish to bound the concentration of the random variable $L(\mathbf{X})$ measuring the longest increasing subsequence.

Slide 61

Let $\mathbf{x} = (x_1, \dots, x_n)$ a sequence of reals and let $K_{\mathbf{x}} \subseteq [n]$ be the set of subindex corresponding to the longest increasing subsequence, so $L(\mathbf{x}) = |K_{\mathbf{x}}|$.

For any other sequence $\mathbf{y} \in [0,1]^n$, we have

$$L(\mathbf{y}) \geq L(\mathbf{x}) - |\{i \in K_{\mathbf{x}} \mid x_i \neq y_i\}|.$$

To bound the Talagrand´s distance between any two points $\mathbf{x}$ and $\mathbf{y}$, let us consider the vector $\mathbf{a}' = (a'_1, \dots, a'_n)$ where for every $i \in [n]$, $a'_i = 1/\sqrt{L(\mathbf{x})}$ if $i \in K_{\mathbf{x}}$ and $a'_i = 0$ otherwise. Then $||\mathbf{a}'|| = 1$. Moreover, by definition of $\mathbf{a}'$, $|\{i \in K_{\mathbf{x}} \mid x_i \neq y_i\}| = \sqrt{L(\mathbf{x})} d_{\mathbf{a}'}(\mathbf{x}, \mathbf{y})$ and we get

$$L(\mathbf{y}) \geq L(\mathbf{x}) - \sqrt{L(\mathbf{x})} d_{\mathbf{a}'}(\mathbf{x}, \mathbf{y}). \tag{7}$$

Slide 62

For instance, take

$$\mathbf{x} = (0.7, 0.8, 0.3, 0.1, 0.5, 0.6, 0.2, 0.7, 0.4) \quad \text{and}$$
$$\mathbf{y} = (0.3, 0.7, 0.4, 0.9, 0.5, 0.2, 0.6, 0.8, 0.1).$$

We have $K_{\mathbf{x}} = \{3, 5, 6, 8\}$, $L(\mathbf{x}) = 4$ and $\mathbf{a}' = (0, 0, \frac{1}{2}, 0, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2}, 0)$. Furthermore, $|\{i \in K_{\mathbf{X}} \mid X_i \neq Y_i\}| = 3$ and $5 = L(\mathbf{y}) \geq 4 - 2$.

If we choose $A = \{\mathbf{x} \mid L(\mathbf{x}) \leq m\}$, where $m = \mathbf{M}[L(\mathbf{X})]$, by definition of median and for any random variable $\mathbf{X}$ on Ω,

$$\mathbf{Pr}[\mathbf{X} \in A] = \mathbf{Pr}[L(\mathbf{X}) \leq m] \geq \frac{1}{2}$$

and

$$\mathbf{Pr}[\mathbf{X} \notin A] = \mathbf{Pr}[L(\mathbf{X}) \geq m] < \frac{1}{2}.$$

For any $t \in [n]$ define the usual way A_t^T. Let $\mathbf{y} \in A$, that means $L(\mathbf{y}) \leq m$, and let $\mathbf{x} \notin A_t^T$. From Equation (7), taking the infimum over all possible $\mathbf{a} \in A$ and replacing $L(\mathbf{y})$ by m, we get $L(\mathbf{x}) \leq m + \sqrt{L(\mathbf{x})}\, d_{\mathbf{a}}(\mathbf{x}, A) \leq m + \sqrt{L(\mathbf{x})}\, d_T(\mathbf{x}, A)$.

So,

$$d_T(\mathbf{x}, A) \geq \frac{L(\mathbf{x}) - m}{\sqrt{L(\mathbf{x})}} \geq \frac{t}{\sqrt{t+m}},$$

where this last inequality follows from the fact that the function $\frac{L(\mathbf{x})-m}{\sqrt{L(\mathbf{x})}}$ is increasing for values of $L(\mathbf{x}) > m$. Thus,

Slide 63

$$\begin{aligned} \mathbf{Pr}\left[\mathbf{X} \notin A_t^T\right] &= \mathbf{Pr}\left[L(\mathbf{X}) \geq m+t\right] \\ &\leq \mathbf{Pr}\left[d_T(\mathbf{X}, A) \geq \frac{t}{\sqrt{t+m}}\right] \\ &\leq e^{-\frac{t^2}{4(m+t)}}. \end{aligned} \tag{8}$$

Using Talagrand's inequality, for any $t \in [m]$,

$$\begin{aligned} \mathbf{Pr}\left[L(\mathbf{X}) \leq m\right] \mathbf{Pr}\left[L(\mathbf{X}) \geq m+t\right] &= \mathbf{Pr}\left[\mathbf{X} \in A\right] \mathbf{Pr}\left[\mathbf{X} \notin A_t^T\right] \\ &\leq \mathbf{Pr}\left[\mathbf{X} \in A\right] \mathbf{Pr}\left[d_T(\mathbf{x}, A) \geq \frac{t}{\sqrt{t+m}}\right] \\ &\leq 2e^{-\frac{t^2}{4(m+t)}} \end{aligned}$$

and, as by definition of the set A, we have $\mathbf{Pr}\left[L(\mathbf{x}) \leq m\right] \geq 1/2$, then for any $t \geq 0$

$$\mathbf{Pr}\left[L(\mathbf{X}) \geq m+t\right] \leq 2e^{-\frac{t^2}{4(m+t)}}, \tag{9}$$

and repeating the argument for $m-t$ in the place of m, we get that for any $t \in [m-1]$,

$$\mathbf{Pr}\left[L(\mathbf{X}) \geq m-t\right] \leq 2e^{-\frac{t^2}{4m}}, \tag{10}$$

which gives a good concentration result.

Slide 64