# Concentration of Values beyond Sums

Given r.vs. $X_1, X_2, X_3 \dots X_n$, Chernoff-Hoeffding bounds give concentration for sums. I.e. if $X_i$ are $[0,1]$ bounded, indep, then

$$\Pr\left(\left|\textstyle\sum_i X_i - E\sum_i X_i\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{2\mu+\lambda}\right).$$

But what if we have some (non-sum) functions $f$ and want to understand if $f(\bar{X})$ is close to $Ef(\bar{X})$? Classic Chernoff is not the tool to use. ☹

if $f \neq \sum_i X_i$

—

Example: Consider graph $G(n,p)$

Every edge chosen indep at random w/p $p$.

$m = \binom{n}{2}$ r.v's.

$X_1\, X_2 \dots X_m$

Let $f(\bar{X})$ = size of maximum matching in $G \sim G_{n,p}$.

not the sum of these rvs

Want to say: $\Pr(|f - Ef| \geq \lambda) \leq$ small.

How can we say that?

Today: several techniques

① Martingales and Azuma-Hoeffding

⇒ Concentration for Lipschitz fns. (McDiarmid)

② Talagrand's Inequality

⇒ Conc. of certifiable functions.

# Concentration of Lipschitz functions

$f : \mathbb{R}^n \to \mathbb{R}$

Sufficient to show
if $x$ and $y$ differ on one coord $i$
$\Rightarrow f(x)$ and $f(y)$ differ by $\le c_i$

suppose $\exists\, c_i \geq 0 \;\forall i \in [n]$ such that

$$f(x) \le f(y) + \sum_{i=1}^{n} c_i \cdot \mathbb{1}(x_i \neq y_i)$$

then we say that $f$ is $\vec{c}$-Lipschitz (with respect to the Hamming metric).

(In general, we want that $\underbrace{f(x) - f(y)}$ $\le \alpha \cdot \text{dist}(x, y)$ )

$\alpha$: Lipschitz Constant

dist: metric on $\mathbb{R}^n$

(deviation in function value small if change the input by a little bit)

---

## Thm [Method of Bounded Differences (McDiarmid)]

Sps $f$ is $\vec{c}$-lipschitz for $\bar{c} = (c_1, c_2 \ldots c_n)$

then if $X_1, X_2 \ldots X_n$ are chosen randomly independently,

$$\Pr\left(|f(\bar{X}) - Ef| \ge \lambda\right) \le 2 \cdot \exp\left(-\frac{2\lambda^2}{\sum_i c_i^2}\right) = 2 \cdot \exp\left(-\frac{2\lambda^2}{\|\bar{c}\|_2^2}\right)$$

E.g. $X_i \in \{-1, 1\}$ unbiased (Rademachers) and $f(\bar{X}) = \sum_i X_i$ $\Rightarrow Ef = 0$.

$c_i = 2$

$\Rightarrow \|c\|^2 = 4n$.

And get back concentration for sums of Rademachers

$$\Pr\left(\left|\sum_i X_i\right| \ge \lambda\right) \le 2 \cdot \exp\left(-\frac{\lambda^2}{2n}\right)$$

Lecture #3)
(Just as before)

But much more general:

Ex 2: $\boxed{n \text{ balls } n \text{ bins}}$

$X_i$ = bin in which ball $i$ falls (unif in $\{1, 2 \dots n\}$ indep).

- Let $\boxed{f(\bar{X}) = \# \text{empty bins}}$

$$Ef = n(1-1/n)^n \sim n/e.$$

- $f$ is 1-Lipschitz (#empties can change by 1 per change)

$$\Rightarrow \Pr(|f - Ef| \geq \lambda) \leq 2\exp(-O(\lambda^2/n))$$

$$\Rightarrow \#\text{empty bins} = n/e \pm O(\sqrt{n \lg n}) \text{ whp}.$$

— x —

Ex 3: $G \sim G_{n,p}$ random graph. (edges indep chosen)

$f$ = #isolated vertices or size of max matching or coloring # of the graph.

min # of colors to allow a "proper" coloring of G

(proper = every edge has diff colors on endpoints)

If we change a single edge (present vs absent)

$\Rightarrow$ #isolated vtxs changes by $\leq 2$.
size of max matching changes by $\leq 1$.
coloring # of graph changes by $\leq 1$.

In all these cases, can use Thm 1 (McDiarmid's Ineq)

to say

$$\Pr(|f - Ef| \geq \lambda) \leq 2 \cdot \exp\left(-\frac{2\lambda^2}{m}\right)$$

#edges $m = \binom{n}{2}$

$$\Rightarrow f \in Ef \pm O(\sqrt{m \log m}) \text{ whp}.$$

Hang on! Here $Ef \le n$ (size of matching, #isolated vtxs, etc).

And $O(\sqrt{m \log n})$
$= O(n\sqrt{\log n})$ ☹

So not very tightly concentrated!!

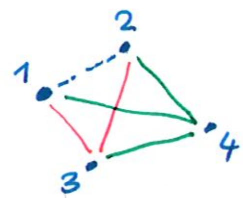
—— x ——

Can do better using a better view/use of the same bound. $\boxed{\text{(for some cases)}}$

Suppose we view the edges in "groups".

Group $Y_i$ = edges going from vertex $i$ to vertices $\{1, 2 \ldots i-1\}$.

⇒ have $n$ groups (actually $n-1$)
of independent r.vs.




$Y_2$ $Y_3$ $Y_4$

$f(Y_1, Y_2 \ldots Y_n)$ is the same thing.

But now: can we show Lipschitzness?

for Matchings: Sure: if we change one group,
the max matching size changes by at most 1.

(In fact, even if change all edges incident to a vertex)

Same for coloring #: even if we change all edges incident to some $i$
can always choose a new color to color it.

(Not so for #isolated vertices, so cannot use "this trick" directly.)

this trick ≡ often called the "vertex exposure martingale" (?)
instead of "edge exposure"
for reasons to be seen soon!

"Exposing edges by vertex groups"

Gives that for matchings / coloring number,

$$f - \mathbb{E}f \in \pm O(\sqrt{n \log n}) \quad \text{whp}.$$

Btw: is this still useful? Is $\mathbb{E}f$ large compared to the deviation $\tilde{O}(\sqrt{n})$?

Well: depends on $p$. Say $p = 1/2$.

$\Rightarrow$ Max matching of size $n/2$

( $G_{n,1/2}$ has perfect matching whp )

$G_{n,p}$ has perf. matching ~~whp~~ wp $1-o(1)$ if $p \geq \frac{\log n}{n}(1+o(1))$

Coloring number $\geq \dfrac{n}{2\log n}$.

Each color class must be an indep set

But know that Max IS in $G_{n,1/2}$ has size $\leq 2\log n$

$\Rightarrow$ these are much larger than the deviation bounds (at least for $p = 1/2$).

[For small $p$, should — and can — do something a bit more ~~s~~ sophisticated — more about this soon].

But first, a glimpse into how to prove McDiarmid's Inequality (Conc. of Lipschitz Functions).

We need "Martingales"...

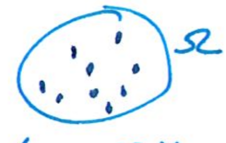
... and conditional expectations...

# Martingales and Concentration

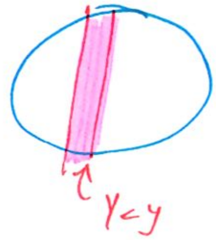
r.v: map from $\Omega$ to reals (say)

Let's recall some basic facts about conditional probability

Sps $X, Y$ are r.vs on a common sample space.

$E(X) =$ 

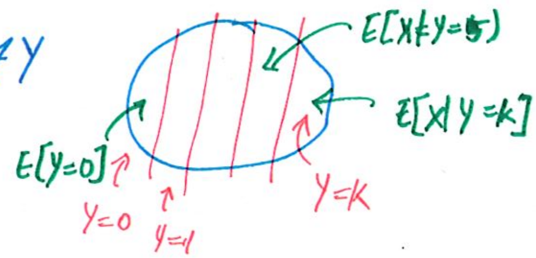

Average of $X$ over all points

Then $E[X | Y = y] =$ 


Average of $X$ over the smaller space $\{Y = y\}$.

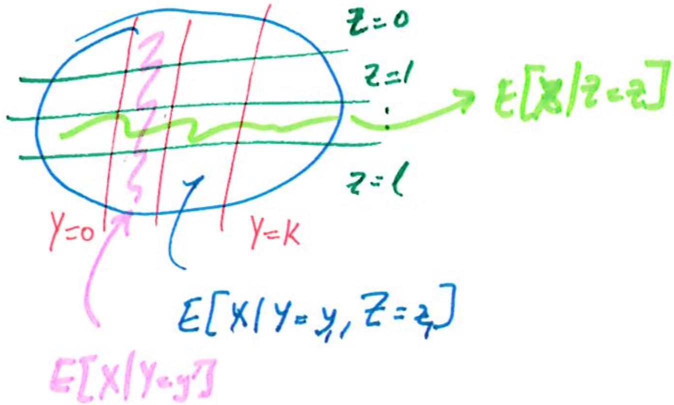
Now $E[X|Y]$ is a function of $Y$

which takes on some constant value in each part.




Fact: $\underbrace{\underset{X}{E}[X]}_{\text{constant}} = \underset{Y}{E}[\underbrace{E[X|Y]}_{\text{r.v.}}] = \sum_i Pr(Y=y) \cdot \underbrace{E[X|Y=y]}_{\text{constant}}$

Fact: $\underset{X}{E}[X|Y] = \underset{Z}{E}[\underset{X}{E}[X|Y,Z] \,|\, \cancel{Y}]$




$E[\cancel{X}] =$ average $X$ over all points in $\Omega$ $=$ scalar (constant)

$E[X|Y] =$ average $X$ over each "levelset" for $Y$. $=$ function of $Y$.

$E[X|Y,Z] =$ average $X$ over each "cell" where $\cancel{Y} =$ something, $Z =$ something

So suppose you have a function $f: \mathbb{R}^n \to \mathbb{R}$

and you have $n$ random variables $Y_1, Y_2 \ldots Y_n$.

and $f(Y_1, Y_2, \ldots Y_n)$ is the (random) value

$f(\bar{Y})$ is a random variable.

"Doob Martingale"

$E[f(Y_1, Y_2 \ldots Y_n)] =$ average.

$\vdots$

$E[f(Y_1, Y_2 \ldots Y_n) \mid Y_1] =$

$\vdots$

$E[f(Y_1 \ldots Y_n) \mid Y_1, \ldots Y_n] = f(Y_1, Y_2 .. Y_n) =$ function itself

go to $f$ from $E[f]$ by "exposing" the variables one at a time

---

OK. All this was discussion of conditional expectations

(Can define things more generally using $\sigma$-algebras and filtrations, see probability text books) useful but not needed today.

Lets get a simple idea called a $\boxed{\text{Martingale.}}$

- Name has a complicated origin, but idea is indeed simple.
- Allows us to model settings where we have
  - correlations between random variables but can still prove concentration
  - get away from merely sums of independent r.Vs.

**Def:** A sequence $(X_t) := X_0, X_1, X_2 \ldots$ of r.vs on a common sample space is a martingale if (a) $E[|X_t|] < \infty$ and.

(b) $E[X_t | X_0, \ldots X_{t-1}] = X_{t-1}$.

**More generally:** a seq $(X_t)$ is a martingale wrt another seq $(Z_t)$ if

(a) $\mathbb{E}[|X_t|] < \infty$

(b) $\exists$ functions $g_t$ st. $X_t = g_t(Z_0, \ldots Z_t)$ and

(c) $\mathbb{E}[X_t | Z_0 \ldots Z_{t-1}] = X_{t-1}$.

—— x ——

It's a bit abstract, let's make this concrete via the most common type of martingales we encounter: $\boxed{\text{Doob Martingales}}$

(Joe Doob, developed the theory of Martingales, @UIUC)

- start with n rvs. $Z_1, Z_2 \ldots Z_n$. over $\Omega$.
- consider a function $f(Z_1, Z_2 \ldots Z_n)$.
- Define $X_t = \mathbb{E}[f | Z_1, \ldots Z_t]$. $\forall t = 0..n$

$$\left.\begin{aligned} \Rightarrow X_0 &= \mathbb{E}[f]. \\ X_1 &= \mathbb{E}[f|Z_1] \\ &\vdots \\ X_n &= E[f|Z_1 \ldots Z_n] = f(Z_1, \cdots Z_n) \end{aligned}\right\}$$

"exposing" the variables one at a time.

—— x ——

**E.g 1:** throw n balls into n bins

$Z_t$ = bin into which $t$th ball goes.

$f(Z_1, \ldots Z_n) =$ number of empty bins

E.g 2: $Z_1, Z_2 \ldots Z_n$ are $n$ iid random points in $[0,1]^2$

$f(Z_1 \ldots Z_n)$ = length of the shortest tour (TSP) over these points.

E.g 3: $Z_1, Z_2 \ldots Z_{\binom{n}{2}}$ are the $n$ edges of a $G_{n,p}$ random graph.

$Z_e = 1$ if the edge $e$ is present in $G \sim G_{n,p}$.

$f(Z_1 \ldots Z_{\binom{n}{2}})$ is the $\boxed{\text{coloring number}}$
or max matching size
or whatever statistic you want } in $G$

— x —

Great. So what?

Concentration for Martingales: (Azuma-Hoeffding)

Suppose $(X_t)$ is a martingale where $|X_t - X_{t-1}| \leq c_t \quad \forall t$.

Then
$$\Pr(|X_n - X_0| > \lambda) \leq 2 \cdot \exp\left(-\frac{\lambda^2}{\sum_{i=1}^{n} c_i^2}\right)$$

— x —

Recall that for Doob martingales, $X_0 = Ef \qquad X_n = f(Z_1 \ldots Z_n)$

$$\Rightarrow \underbrace{\Pr(|f - Ef| > \lambda)}_{} \leq 2\exp\left(-\frac{\lambda^2}{\sum c_i^2}\right)$$

if $c_i \geq |E(f|Z_1 \ldots Z_i] - E(f|Z_1 \ldots Z_{i-1}]|$

Probability that function deviates from its mean!

— x —

Exercise: Prove McDiarmid's Inequality from Azuma-Hoeffding!!

Exist Extensions of these Inequalities

if want instance specific deviations

to situations where want deviations to not ~~always~~ be $O(\sqrt{n \log \frac{1}{\delta}})$

but to be much smaller (depending on the mean)

---

As an aside, $\boxed{\text{Freedman's Inequality}}$ is one such:

Sps. $Y_i := \mathbb{E}[f(\bar{x}) \mid X_1 \ldots X_i]$    if $f$ is $c$-Lipschitz $\Rightarrow |Y_i - Y_{i-1}| \le c$.

Now define $W_n = \sum_{i=1}^{n} \mathbb{E}[(Y_i - Y_{i-1})^2 \mid X_1 \ldots X_{i-1}]$

"Observed variation".

if indep $\Rightarrow \sigma^2$ = Variance of differences.

Then for any $\lambda, t \geq 0$

$$\Pr\left(|f - Ef| \geq \lambda \text{ and } \underbrace{W_n \le t^2}\right) \le 2\exp\left(-\frac{\lambda^2}{t^2 + 2c\lambda}\right)$$

typically choose $t$ st $W_n \le t^2$ whp.

$\Rightarrow$ LHS $\approx \Pr(|f - Ef| \geq \lambda)$

——— x ———

But today lets consider (a special case) of another famous Inequality:

$\boxed{\text{Talagrand's Concentration Inequality.}}$

And see how it gives good concentration for (say) size of matching in $G_{n,p}$ even when $p$ is very small.

__In general__: for all functions that are Lipschitz and have "small certificates" for their value.

$\Rightarrow$ this bound can give concentration

Sps $X_1, X_2 \dots X_m$ independent.

$f(\bar{X})$. where $f$ is c-Lipschitz (say)

$|f(x_1, \dots x_n) - f(y_1, \dots y_n)| \le c \cdot$ (#coords where $\bar{x}, \bar{y}$ differ).

**Def**: $f$ is h-certifiable if

when $f(x) \geq s$ $\exists \le h(s)$ coordinates s.t.

$f(y) \geq s$ whenever $y$ & $x$ agree on those coords.

e.g. $f(X_1, \dots X_m)$ = max matching in a graph

$X_i \in \{0,1\}$
whether $i^{th}$ edge exists or not

$\Rightarrow$ if $\exists$ matching of size $\geq s$. Choose those $s$ vars, clearly if those $s$ vars $= 1$ $\Rightarrow$ max matching has size $\geq s$.

$\Rightarrow$ (Application of) Talagrand's Concentration Inequality

If $f$ is c-Lipschitz and h-certifiable:-

$$\Pr(|f - M_f| \geq \lambda) \leq 4 \cdot \exp\left(-\frac{\lambda^2}{4c^2 \cdot h(M_f + \lambda)}\right)$$

"Median" of $f$
$\equiv$ value s.t. $\Pr(f \geq M_f) \geq 1/2$
$\Pr(f \leq M_f) \geq 1/2$

"Concentration around the median"

rather than around the mean.

$f$ is bounded and
But if Prob of deviation is small
$\Rightarrow$ both are very close

For matching example: $h(s) = s$. and $c = 1$.

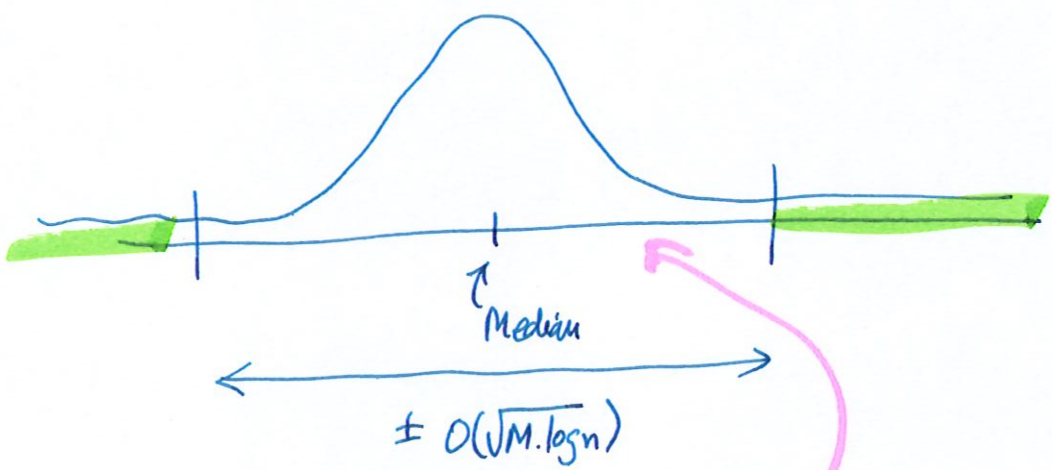
$$\Rightarrow \Pr\left(|f - M_f| \geq \lambda\right) \leq 4 \exp\left(-\frac{\lambda^2}{4(M_f+\lambda)}\right)$$

if set $\lambda = 10\sqrt{M_f \cdot \log n}$ $\quad \leq 4 \exp\left(-\frac{10 \, M_f \log n}{4 \cdot 2 M_f}\right)$

say $M_f \geq 100 \log n$ as well $\quad \leq 1/\mathrm{poly}(n)$.

$\Rightarrow \lambda \leq M_f$

$\Rightarrow$ have concentration around the median matching value, whp.




$\Rightarrow$ mean must be here as well!

$\Rightarrow$ concentration around the mean too.

General Talagrand's Inequality proves more, but

less easy to directly use — see Alon Spencer book, etc

———— x ————

Side note:

median

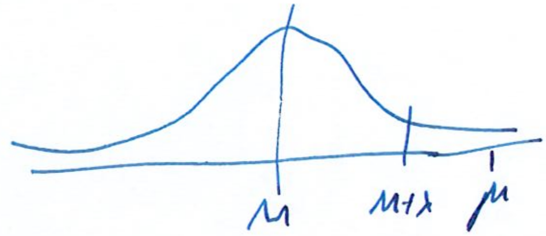
or any other statistic, really

Sps. $\Pr(|f-M| \geq \lambda) \leq 1/\mathrm{poly}(n)$ for some $\lambda$

and $|f| \leq n$ say

$\Rightarrow$ $|M-\mu| \leq \lambda + o(1)$.

mean




SPS: sps $\mu \geq M+\lambda+c$

by def, $\mu = E[f \mid f \geq M+\lambda] \Pr[f \geq M+\lambda]$
$\quad + E[f \mid f \leq M+\lambda] \Pr[f \leq M+\lambda]$

$\leq n \cdot 1/\mathrm{poly}(n) + (M+\lambda) \cdot 1$

not possible unless $c \leq n/\mathrm{poly}(n)$. Similarly for $\mu \leq M-\lambda-c$.

$\Rightarrow$ so $|M-\mu| \leq \lambda + o(1)$

$\Rightarrow$ $\Pr(|f-\mu| \geq 2\lambda + o(1)) \leq 1/\mathrm{poly}(n)$.

Wrap up:

Saw: McDiarmid's Inequality (for Lipschitz fns of Ind. r.vs)

(special case of) Talagrand's Concentration Inequality

- A quick glimpse of Freedman's Inequality
  (which is more "Chernoff like", depends on the mean
  and not just on # of r.vs)

- Martingales and Doob Martingales
  - important devices when dealing with dependent choices
    - Fair Matching (Bikstad+)
    - Next lecture in Discrepancy Algos.

- Did not get to:
  - Stopping times and Optional Stopping Theorem
  - Wald's Inequality
    - have some notes on these, completely optional
      (adding them in)