Concentration of Measure

A. Winter

1. Elementary beginnings
2. Concentration functions
3. $l^p$-products; bounded differences
4. Gaussian measure; spherical concentration
5. Random subspaces; counterexamples to a certain norm-multiplicativity conjecture
1. Elementary beginnings

(i) **Markov inequality**

Let $X \geq 0$ be a random variable s.t. $EX = \sum_x P_x(x) x$ exists (IP$_x$: prob. dist. of $X$).

Then $EX \geq \sum_{x \geq t} P_x(x) x \geq t \sum_{x \geq t} P_x(x) = t \cdot P_x \{ X \geq t \}$

$\Rightarrow P\{ X > t \} < \frac{EX}{t}$ \hspace{1cm} (ii)
(ii) Potentially better estimate:

\[
P(\{X > t\}) = P(\{|X^2 - t^2| < \frac{EX^2}{t^2}\}) < \frac{EX^2}{t^2} \quad (if \ EX^2 < \infty)
\]

**Chebyshev's inequality:**

Let \( X \) be RV s.t. \( EX, \ EX^2 \) exist. Then

\[
P(\{|X - EX| > t\}) = P(\{|X - EX|^2 > t^2\}) < \frac{\text{Var} X}{t^2}
\]

with \( \text{Var} X = \text{E}(X - EX)^2 \)

\[= \text{EX}^2 - (\text{EX})^2 \]
(iii) We can of course apply the squaring trick more generally: for \( x > 0 \)
\[ P_X(X > t^2) \leq \frac{\text{EX}^m}{t^m} \]

Indeed: any monotone function of on the range of \( X \) will do...

\[ P_X(X > t^2) = P_X(f(x) > f(t)^2) < \frac{\text{E} f(x)}{f(t)} \]
(iv) (Weak) Law of Large Numbers (LLN)

[simplify by assuming existence of \( \mathbb{E}X, \mathbb{E}X^2 \)]

... e.g. if \( X \) is bounded

Let \( X_1, X_2, \ldots, X_n, \ldots \sim \mathcal{IP}_X \) (pairwise) independent

Then for \( S_n = \frac{1}{n} \sum_{i=1}^{n} X_i \)

\[ P \left( | S_n - \mathbb{E}X | > t \right) < \frac{1}{n} \frac{\mathbb{Var}X}{t^2} \rightarrow 0 \text{ as } n \rightarrow \infty \]

Proof: Wlog. \( \mathbb{E}X = 0 \), so \( \mathbb{E}S_n = 0 \) & \( \mathbb{Var}X = \mathbb{E}X^2 \)

Now:

\[ \mathbb{Var} S_n = \mathbb{E}S_n^2 = \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^{n} X_i X_j \right) = \frac{1}{n} \mathbb{E}X^2 \]

Apply Chebyshev's...
Exploiting independence & higher moments of other functions, we get much stronger bounds:

(v) Assume $X$ is bounded, say $0 \leq X \leq 1$

Then for every $s > 0$

$$\Pr\{X > t\} = \Pr\{e^{sX} > e^{st}\} \leq e^{-st} \mathbb{E} e^{sX}$$

with $X_i \sim P_X$ (i.i.d) and $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$:

$$\Pr\{S_n > t\} < e^{-nst} \mathbb{E} e^{\sum_{i=1}^{n} X_i} = (e^{-st} \mathbb{E} e^{sX})^n$$

Optimise $s$ to get this $< 1$!
**Chernoff-inequality:**

\[ P\{ S_n > t \} < e^{-n \Delta^*(t)} \]

with \[ \Delta^*(t) = \sup_{s \geq 0} s - \ln \mathbb{E} e^{sX} \]

\[ = \Lambda(s) \]

\[ \mathbb{E} e^{sX} \] is called the moment generating function.

(similar for \( P\{ S_n < t \} \) and so on...)

Ex. \( \Delta^*(t) \) minimal for Bernoulli variable \( X \), i.e. \( P(X=1) = 1 \) \( \Rightarrow \mathbb{E}X = 1 - P(X=0) \)

Ex. \( \Delta^*(t) \geq D(t \mid \mid 1 - \mathbb{E}X) \succ 2(t - \mathbb{E}X)^2 \)
2nd scenario gives

(vi) Hoeffding's inequality:

\[ P \left\{ |S_n - \mu| > \epsilon \right\} \leq 2e^{-2n\epsilon^2} \]

(4)

So, in many cases a RV (e.g. Sums of iid) is "concentrated" near one value — its mean.
(vii) **Strong LLN:** (Assuming $X$ bounded, \\
\text{w.r.t. } 0 \leq X \leq 1)

For independent $X_i \sim P_X$, $S_n = \frac{1}{n} \sum_{i=1}^{n} X_i$

$S_n \to \mu X$ with prob. 1

**Proof:** Let $\epsilon_n \to 0$ s.t. $\sum_{n=1}^{\infty} n \epsilon_n^2 \to \infty$ fastenough s.t. $\sum_{n=1}^{\infty} 2 e^{-2n \epsilon_n^2} < \infty$ [\text{e.g. } \epsilon_n = n^{-1/4}]

\[\Pr (|X| > \epsilon_n) \leq \frac{\sum_{n=1}^{\infty} P(|X_n - \mu X| > \epsilon_n)}{n=1} \to 0\]
Now we can use the **Borel-Cantelli Lemma**:

If $\sum P(E_n) < \infty$ for events $E_n$, then with prob 1 only finitely many $E_n$ occur. I.e., letting $T = \sum 1_{E_n}$ count how many occur, then $P\{T = \infty\} = 0$.

Why? $E_T = \sum P(E_n) < \infty$ & use Markov.

I.e., $P\{S_n \to \infty\} \leq P\{|S_n - TEx| > \epsilon, \textrm{ infinitely often}\} = 0$.  

\[\quad\]
This is not all about sums of iid RVs & how they converge to the expectation (Laplace/central theorem).

There is something deeper, geometric, going on:

**Geometric Measure Concentration**

is the phenomenon that on some metric spaces with probability distribution, if a set $A$ has "some" mass, then a neighborhood of $A$ has "most" of the mass.
2. Concentration functions

We'll consider metric measure spaces

\( X \) — metric d

prob. distrib. \( \mu \)

**Definition.** The concentration function \( c_X \) is:

\[
    c_X(\tau) = \sup \{ 1 - \mu(A_\tau) : \mu(A) \geq \frac{1}{2} \}
\]

where \( A_\tau = \{ x \in X : d(x, A) \leq \tau \} \)

is the \( \tau \)-neighborhood of \( A \)

\[
    \mu(A) \geq \frac{1}{2} \Rightarrow \mu(A^c \cap A_\tau) \leq c_X(\tau)
\]
Concentration fact: tells how much one has to "blow up" a fat set to get an even fatter set.

Says also something about how much we need to get a "small" set fat:

**Lemma:** Let \( A \) st. \( \mu(A) \geq \epsilon \), and \( r_0 \) st. \( x(r_0) \leq \epsilon \)
Then \( \mu(A_{r_0}) \geq \frac{1}{2} \); consequently \( \mu(A_{r_0 + r}) \geq x(x(r)) \)

**Proof:** Let \( B = A_{r_0} \), so \( A \subset B_{r_0} \). If \( \mu(B) \leq \frac{1}{2} \) then \( \mu(A) \leq 1 - \mu(B_{r_0}) \leq x(r_0) \leq \epsilon \), contradicting.
Hence \( \mu(A_{r_0}) \geq \frac{1}{2} \). \( \Box \)
What it's good for:

Consider 1-Lipschitz function $F$ on $\mathbb{X}$ (that's a random variable $\mathcal{F}$)

i.e. $|F(x) - F(y)| \leq d(x,y)$

Then:

$$\Pr \{ F > m_F + t \} \leq \frac{d_X(t)}{X}$$

for any median $m_F$ [i.e. $\Pr \{ F > m_F \} \geq \frac{1}{2}$]

$\Pr \{ F \leq m_F \} \leq \frac{1}{2}$

So, if $d_X(t)$ decays "fast", $F$ will be concentrated near $m_F$ (hence $\mathbb{E} F$)...
Conversely, if \( \| \mathbf{F} \| \leq c \) holds for any function \( F \):

\[
Pr \{ |F| > m_F + t \} \leq \alpha(t) \quad \forall 1\text{-Lipschitz} \quad F
\]

Then: \( d_X(r) \leq d(t) \)

Consider \( F(x) = d(x, A) \) for a set \( A \) s.t. \( \mu(A) = \frac{1}{2} \).

\( m_F = 0 \) is a median for it, hence

\[
\mu(A_+) = \mathbb{P} \{ F \leq r^2 \} \geq 1 - \alpha(t)
\]

Consequence:

\[
d_X(r) = \sup_{1\text{-Lipschitz}} Pr \{ F > m_F + r \}
\]
Can also characterize/bound $\alpha_x$ in terms of deviation from the mean:

Let $\beta_x(r) = \sup_{F \text{ $1$-Lipschitz (bounded)}} P[F > EF + r]$

Then for any set $A$:

$$1 - \mu(A_+) \leq \beta_x(\mu(A) r)$$

... in particular:

$$\alpha_x(r) = \beta_x(\frac{r}{2})$$

Consider $F(x) = \min \{ d(x; A) \}, r \geq 0$. Observe $\text{IEF} \leq r(1 - \mu(A))$

Hence $1 - \mu(A_+) = P[F > r] \leq P[F > \text{IEF} + r\mu(A)] \leq \beta_x(\mu(A) r)$

... perhaps it's time for some more concrete examples...
For $\mathbb{R}$ with the usual metric $d(x,y) = |x - y|$, let $X$ be distributed according to $\mu$.

**Chebyshev:** If $E X^2 < \infty$, then $\beta_X(r) \leq \frac{\text{Var} X}{r^2}$

** Chernoff:** If $X$ is bounded, then $\beta_X(r) \leq e^{-\Delta_X(E X + r)}$

**Hoeffding:** If $X \in [0,1]$, then $\beta_X(r) \leq e^{-2r^2}$
3. $\ell^\prime$-products

To exploit independence: consider product of metric measure spaces $X_i$ (metric $d_i$, p.d. $\mu_i$)

$$X = X_1 \times X_2 \times \ldots \times X_n$$

$$\mu = \mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_n$$

$$\lambda = \sum_{i=1}^{\infty} \lambda_i$$, \text{i.e.} \hspace{1em} \Delta(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{\infty} \lambda_i \Delta_i(x_i, y_i)$$

Let's assume $\Delta_i = \dim(X_i) = \sup_{x,y \in X_i} d_i(x,y) < \infty$
For discussing the Brenshel trick:

Consider \( \operatorname{Exp}_x(\lambda) := \sup_{F \text{ 1-lipschitz}} \int d\mu(x) e^{\lambda F(x)} \) bounded, \( \lambda F = 0 \)

\[ \text{Observe: } \operatorname{Exp}_{xy} = \operatorname{Exp}_x \cdot \operatorname{Exp}_y \]

\[ \text{Proof:} \]
- \( \text{easy - Look at } F(x) + G(y) \)
- \( \leq \text{ for given } F, \text{ consider } G(y) := \int d\mu(x) F(x,y) \)
- \( F^g(x) := F(x,y) - G(y) \)

\( \Rightarrow \) both 0-mean, 1-lipschitz

Now: \( \int e^{\lambda F(x,y)} d\mu(x) d\nu(y) = \int d\mu(x) e^{\lambda F^g(x)} \leq \operatorname{Exp}_x(\lambda) \)

\( \leq \operatorname{Exp}_y(\lambda) \cdot \operatorname{Exp}_x(\lambda) \)
Need only one simple lemma:

For metric measure space $X$ of diameter $D = \max_{x \in X} d(x) < \infty$,

$$\exp_x(A) \leq e^{\lambda^2 D^2/2}$$

**Proof.**

$$\int d\mu(x) \, e^{\lambda \Xi(x)} \leq \int \int d\mu(x) d\mu(y) \, e^{\lambda \Xi(x) - \lambda \Xi(y)}$$

Converting to $e^{is}$

$$\leq \sum_{i=0}^{\infty} \frac{(\lambda D)^{2i}}{(2i)!} = e^{\lambda^2 D^2/2}$$

(Handwritten notes...)

$\blacksquare$
For $X = X_1 \times \ldots \times X_n$, $\dim(X_i) = D_i$:

$\operatorname{Exp}_X(X) \leq e^{\frac{1}{2} D^2}$ with $D^2 = \frac{1}{n} \sum D_i^2$

$P\{F > t^2\} \leq e^{\frac{-X^2}{2D^2}}$

If $F \sim \chi^2$ with $k$ degrees of freedom,

$P\{F > t^2\} \leq e^{\frac{-t^2}{2D^2}}$

Applications: Any function of $n$ independent variables which does not depend too much on any one of them is strongly concentrated.
Random Bin Packing:

Let $A_1, A_2, \ldots, A_n \sim \text{iid random}$
sizes of "items".

Task: distribute them into a number of "bins" each of unit size. Count the number of bins used, according to some method: $B(A_1, \ldots, A_n)$.

$B_{opt}^n : = \text{minimal number of bins}$

($n$-hard to compute!)

$B_{FF}^n : = \text{number used in "first-fit"}$

(go through the list and place next item into

first available bin)

$B_{SFF}^n : = \text{number of bins in "sort & first-fit"}$ ($\ldots$)
All of the above satisfy

(i) \( B_{n+m} \leq B_n + B_m \)

implies the existence of \( \lim_{n \to \infty} \frac{\text{EB}_n}{n} =: \gamma \) (Ex.)

(ii) \( B_n \) is 1-Lipschitz

Note: \( B_n \geq \sum_{i=1}^{\infty} A_i \), \( \gamma \geq \mathbb{E}A_1 \)

Gives already:

\[ P\{ |B_n - \text{EB}_n| > n \epsilon^2 \} \leq 2 e^{-n \epsilon^2 / 2} \]

...and by the Borel-Cantelli-trick from before:

\[ \lim_{n \to \infty} \frac{1}{n} B_n = \gamma \text{ w. prob. 1} \]
Don't even always need independence.

Ex. Look again at the proof on p74 to show

(i) Let F be any function on $X = X_1 \times \cdots \times X_n$
    endowed with any measure s.t.
    $\forall i, \text{ increment } \Delta_i = E_{x_1 \cdots \hat{x}_i \cdots x_n} F - E_{x_1 \cdots x_n} F$

    $\mid \Delta_i \mid \leq D_i$

    Then $\ell E_{x_i} \Delta_i = 0$ and
    \[ \sum_{i=1}^{n} \rho(x_1, \cdots, x_i) e^{x_i} \leq e^{\frac{x^2 D^2}{2}} \quad \text{ and } \quad D^2 = \sum_{i=1}^{n} D_i^2 \]

(ii) Infer Azuma's inequality
    \[ \mathbb{E} F \geq \ell F + t \frac{1}{2} \leq e^{-t^2 / 2D^2} \quad (A) \]
This can be applied in the analysis of randomized algorithms.

E.g. QUICKSORT ... has expected time $T(n) \approx n \log n$.

Median/Haywood shows:

$$\Pr\{\left| \frac{q_n}{\E q_n} - 1 \right| > \epsilon \} \leq n^{-2\epsilon \log \log n}$$