Concentration Inequalities

August 17, 2012

Outline

- What are concentration inequalities ?
- Different methods
  - Moment method
  - Exponential Moment method
  - Martingale methods
  - Entropy Methods
  - Talagrand's Inequality (Induction methods)

Introduction

- Some simple but very important statements:
  - In a long sequence of tossing a fair coin, it is likely that head will come up nearly half of the time
  - A random variable that depends (in a smooth way) on the influence of many independent variables (but not too much on any of them) is essentially constant
  - A random variable that depends (in a smooth way) on the influence of many independent variables satisfies Chernoff-type bounds
- Concentration inequalities make the above statements precise
- These inequalities are in general a manifestation of the phenomena of **measure concentration** (especially on product spaces)

Basic Setting

- We'll consider random variables in product spaces
- $X_1, X_2, \ldots, X_n$ be n independent RVs
- What can be said about $S_n = \sum_{i=1}^{n} X_i$?
  - If each of the X_i is of $O(1)$, what is the typical size of S_n ?
  - Linear processing of RVs (noise), projections etc.
- What can be said about some non-linear "well-behaved" $F(X_1, X_2, \ldots, X_n)$?
  - Norms, Output of (say) a convex optimization program ?

Moment method: We already know this ...

- For linear combinations moment method is very natural and useful
- First moment method (Use Markov Inequality)

$$\mathbb{P}(|S_n| \geq t) \leq \frac{1}{t} \sum_{i=1}^{n} \mathbb{E}|X_i| \tag{1}$$

- Second moment method (Use Chebyshev Inequality)

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq t) \leq \frac{1}{t^2} \sum_{i=1}^{n} Var(X_i) \tag{2}$$

Moment method (Contd.)

- Second moment: some remarks ...
  - Informally, size of $S_n = \mathbb{E}S_n + O(\sqrt{\sum_{i=1}^{n} Var(X_i)})$
  - Do not need full independence, just pairwise uncorrelated will suffice
  - Instead of S_n if we have some other function, then Chebyshev bound still applies as long as we can estimate (or upper bound) mean and the variance
  - Clearly this is way off the mark . . . (Why ? / Why not ?)
- Using Markov's inequality and some book-keeping this can be extended to k moments with k even

$$\mathbb{P}(|S_n| \geq t) \leq 2\left(\frac{\sqrt{enk/2}}{t}\right)^k \tag{3}$$

Moment method (Contd.)

- k^{th} moment: some remarks ...
    - Informally, S_n grows as $O(\sqrt{nk})$
    - For higher k we get higher decay rate (still polynomial though) ...
- What can full independence give us ?
    - We can use the above equation for any k. Thus by optimizing in k we are able to get exponential quadratic decay.

$$\mathbb{P}(|S_n| \geq t) \leq Cexp(-ct^2/n) \tag{4}$$

- But there are better ways to see this ...

Chernoff Bound: We know this also ..

- Chernoff Bounding method

$$\mathbb{P}(|X| \geq t) \leq \min_{s>0} \frac{\mathbb{E}e^{sX_i}}{e^{st}} \tag{5}$$

- Chernoff bound is well suited to tackle sums of independent RVs (Why ?)
- First estimate $\mathbb{E}e^{sX_i}$ and then optimize over s
- For bounded random variables, Hoeffding's Lemma is one of the best known results

Hoeffding's Lemma

- Let X be a bounded scalar random variable taking values in $[a, b]$. Then for any $t > 0$:

$$\mathbb{E}e^{tX} \leq e^{t\mathbb{E}X}\{1 + O(t^2\mathbf{Var}(X)exp(O(t(b-a))))\} \tag{6}$$

- In particular, if $\mathbb{E}X = 0$ then,

$$\mathbb{E}e^{tX} \leq e^{s^2(b-a)^2/8} \tag{7}$$

- Proof ?

Hoeffding's Inequality and cousins . . .

- (Hoeffding's Inequality): Let $X_i, i = 1, 2, \ldots, n$ be independent RVs taking values in an interval $[a_i \; b_i]$, respectively. Then there exist constants $C, c > 0$ such that

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq t) \leq Cexp(-ct^2/\sigma^2) \text{ where } \sigma^2 = \sum_{i=1}^{n}(b_i - a_i)^2$$

  - If $\mathbf{Var}(X_i)$ is known then the above bound is little conservative
  - Bernstein, Bennet, Chernoff's Inequality remedy that fact by using variance in the upper bound and also tightening the bounds further
  - But boundedness of all the RVs is *still* assumed

Chernoff Method: Norm of a Gaussian RV

- $X \sim N(0, \sigma^2 \mathbf{I})$. Concentration inequality for $\|X\|_2^2$?
- $\mathbb{E}e^{s\|X\|_2^2} = (1 - 2s\sigma^2)^{-n/2}$ (Completion of squares)
- $\mu \triangleq \mathbb{E}(\|X\|_2^2) = n\sigma^2$
- Chernoff:

$$\mathbb{P}(\|X\|_2^2 \geq (1+t)\mu) \leq \min_{s>0} (1 - 2s\sigma^2)^{-n/2} e^{-s(1+t)n\sigma^2}$$

- Optimize in $s = t/(2(1+t)\sigma^2)$, and after some calculus

$$\mathbb{P}(\|X\|_2^2 \geq (1+t)\mu) \leq e^{-t^2 n/6} \quad \text{for} \quad 0 < t < 1/2$$

- This forms the basis of one of the proofs of JL-Lemma

Truncation Methods

- What happens if the RVs are not bounded ? (e.g. Gaussian, exponential etc.)
- Sometimes above results can be extended if the tails of RVs decay sufficiently fast
- Spirit of the method: Divide and conquer
    - Divide: $X = X_{\leq N} + X_{>N}$
    - $X_{\leq N}$ is bounded; For $X_{>N}$ the hope is that if X has good decay properties then we can use simple (Union bound or First moment method) to control $\mathbb{P}(X_{>N} > t)$
    - Classical examples: Weak LLN; Strong LLN;

Hoeffding's Inequality for Sub-Gaussian RVs

- Sub-Gaussian RV:
  $\mathbb{P}(|X| > t) \leq Cexp(-ct^2) \iff \mathbb{E}e^{tX} \leq e^{ct^2}$(zero mean)
- Let X_i be zero-mean, independent sub-gaussian RV. Then

$$\mathbb{P}(|\sum_{i=1}^{n} a_i X_i| \geq t) \leq Cexp(-ct^2/\|a\|^2)$$

- Let X_i be iid sub-gaussian RV. Then for sufficently large A (independent of n) we have:

$$\mathbb{P}(|S_n - \mathbb{E}S_n| \geq An) \leq C_A exp(-c_A n)$$

Furthermore, c_A grows linearly in A as $A \to \infty$

Basics

- Let $\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \ldots \subset \mathcal{F}_n = \mathcal{F}$ be a finite filtration of sub-fields of $(\Omega, \mathcal{F}, \mathbb{P})$
- A sequence Y_i is martingale if $\mathbb{E}(Y_{i+1}|\mathcal{F}_i) = Y_i$
- Basic results of conditional expectations
  - $X \in \mathcal{F}_1$, then $\mathbb{E}(X|\mathcal{F}_1) = X$
  - $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_i)) = \mathbb{E}(X)$ - Iterated expectations
  - $X \in \mathcal{F}_i$, $\mathbb{E}(XY) = \mathbb{E}(\mathbb{E}(XY|\mathcal{F}_i)) = \mathbb{E}(X\mathbb{E}(Y|\mathcal{F}_i))$
  - $\mathcal{F}_1 \subset \mathcal{F}_2$, $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$
  - $\mathcal{F}_2 \subset \mathcal{F}_1$, $\mathbb{E}(\mathbb{E}(X|\mathcal{F}_1)|\mathcal{F}_2) = \mathbb{E}(X|\mathcal{F}_1)$. Smaller sub-field always wins !!
  - $\mathbb{E}(X|\mathcal{F}_n) = X$ and $\mathbb{E}(X|\mathcal{F}_0) = \mathbb{E}(X)$

Concentration using Martingales

- Let X be a RV, then $X_i \triangleq \mathbb{E}(X|\mathcal{F}_i)$ is a martingale
- $d_i \triangleq X_i - X_{i-1}$. Or $d_i = (\mathbb{E}^{\mathcal{F}_i - \mathcal{F}_{i-1}})(X)$
  - $\mathbb{E}(d_i|\mathcal{F}_{i-1}) = 0$
  - $X - \mathbb{E}(X) = \sum_{i=1}^{n} d_i$
- Main result: For every $t \geq 0$,

$$\mathbb{P}(\sum_{i=1}^{n} d_i \geq t) \leq e^{-t^2/2D^2} \quad \text{where} \quad D^2 \geq \sum_{i=1}^{n} \|d_i\|_\infty^2$$

- Key is to come up with decomposition such that d_i of a given function can be controlled

Proof idea

- Chernoff scheme is not useful since independence is no longer available; Iterated expectations come to rescue
- For $-1 \leq u \leq 1$, $e^{su} \leq \frac{1+u}{2} e^{s} + \frac{1-u}{2} e^{-s}$
- $\mathbb{E}(e^{sd_i} | \mathcal{F}_{i-1}) \leq cosh(s \|d_i\|_\infty) \leq e^{s^2 \|d_i\|_\infty^2 / 2}$
- $\mathbb{E}(e^{s \sum_{i=1}^{n} d_i}) = \mathbb{E}(e^{s \sum_{i=1}^{n-1} d_i} \mathbb{E}(e^{sd_n} | \mathcal{F}_{n-1})) \leq e^{s^2 \|d_n\|_\infty^2 / 2} \mathbb{E}(e^{s \sum_{i=1}^{n-1} d_i})$
- Iterate over i and then optimize over s

Concentration of functions with bounded difference

- Let $f : \mathbb{R}^n \to \mathbb{R}$ has a bounded difference property

$$|f(x_1, \ldots, x_i, \ldots, x_m) - f(x_1, \ldots, x_i', \ldots, x_m)| \leq c_i$$

for all $x_1, x_2, \ldots, x_n, x_i'$. Let X_i for $i = 1, 2, \ldots, n$ be independent RVs. Then

$$\mathbb{P}(f - \mathbb{E}(f) \geq t) \leq e^{-2t^2 / \sum_{i=1}^n c_i^2}$$

  - Choose $\mathcal{F}_i = \sigma(X_0, X_1, \ldots, X_i)$; Let $d_i = (\mathbb{E}^{\mathcal{F}_i - \mathcal{F}_{i-1}})(f)$
  - We can prove that $d_i | \mathcal{F}_{i-1}$ is bounded by c_i and then use Hoeffding inequality to bound $\sum_{i=1}^n d_i$

Simple bound of variance of a function ..

- Let $Z = f(X_1, \ldots, X_i, \ldots, X_n)$, where X_i are independent
- $\mathbb{E}_i(Z) \triangleq \mathbb{E}(Z|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$
- $Var(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[(Z - \mathbb{E}_i Z)^2\right]$
- (Efron-Stein Inequality). Let $X_1', X_2', \ldots, X_n'$ be an independent copy of above RVs. Let $Z_i' = f(X_1, \ldots, X_i', \ldots, X_n)$. Then

$$Var(Z) \leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\left[(Z - Z_i')^2\right] \tag{8}$$

  - Proof using Martingale difference sequence
  - Once variance is bounded, we can use Chebyshev's inequality
  - Another way to prove McDiarmid's inequality

Martingale method: another example

- Norm of sum of independent RVs
- Let $\mathcal{F}_i$ be generated by $Y_1, Y-2, \ldots, Y_i$

$$|d_i| = |(\mathbb{E}^{\mathcal{F}_i - \mathcal{F}_{i-1}})(\|S\|)| \tag{9}$$
$$= |(\mathbb{E}^{\mathcal{F}_i - \mathcal{F}_{i-1}})(\|S\| - \|S - Y_i\|)| \tag{10}$$
$$\leq \|Y_i\| + \mathbb{E}(\|Y_i\|) \tag{11}$$

- Thus if Y_i are independent, bounded RVs and let $S = \sum_{i=1}^n Y_i$. Then:

$$\mathbb{P}(|\ \|S\| - \mathbb{E}(\|S\|)\ | > t) \leq 2e^{-t^2/2D^2} \tag{12}$$

where $D^2 \geq \sum_{i=1}^n \|Y_i\|_\infty^2$.

Summary

THANK YOU