A Gentle Introduction to Concentration Inequalities

Karthik Sridharan

Abstract

This notes is ment to be a review of some basic inequalities and bounds on Random variables. A basic understanding of probability theory and set algebra might be required of the reader. This document is aimed to provide clear and complete proof for some inequalities. For readers familiar with the topics, many of the steps might seem trivial. None the less they are provided to simplify the proofs for readers new to the topic. This notes also provides to the best of my knowledge, the most generalized statement and proof for Symmetrization lemma. I also provides the less famous but geralized proof for Jensen's inequality and logarithmic sobolev inequality. Refer [2] for a more detailed review of many of these inequalities with examples demonstrating their uses.

1 Preliminary

Throughout this notes we shall consider a probability space $(\Omega, \mathcal{E}, P)$ where Ω is the sample space, $\mathcal{E}$ is the event class which is a $\sigma - algebra$ on Ω and P is a probability measure. Further, we shall assume that there exists a borel measurable function mapping every point $\omega \in \Omega$ to a real number uniquely called a random variable. We shall call the space of random variables X (note that $X \subseteq R$). Further, we shall assume that all the functions and sets defined in the notes are measurable under the probability measure P.

2 Chebychev's Inequality

Let us start this notes by proving what is refered to as Chebychev's Inequality in [3]. Note, often by Chebychev's inequality an inequality derived from the below proved theorem is used. However [3] refers to this inequality as Tchebycheff's Inequality and the same is followed in this notes.

Theorem 1 *For some $a \in X$ where $X \subseteq R$, let f be a non-negative function such that $\{f(x) \geq b | \forall x \geq a\}$, where $b \in Y$ where $Y \subseteq R$. Then the following inequality holds,*

$$P(x \geq a) \leq \frac{E\{f(x)\}}{b}$$

Proof Let set $X_1 = \{x : x \geq a \& x \in X\}$ therefore we have,

$$X_1 \subseteq X$$

Since f is a non-negative function, taking the lebesgue integral of the function over sets X_1 and X we have,

$$\int_X f dP \geq \int_{X_1} f dP \geq b \int_{X_1} dP$$

where $\int_{X_1} dP = P(X_1)$. However the lebague integral over probability measure of a function is its expectation. Hence we have,

$$E\{f(x : x \in X)\} \geq bP(X_1)$$

$$\Rightarrow E\{f(x)\} \geq bP(x \geq a)$$

and hence

$$P(x \geq a) \leq \frac{E\{f(x)\}}{b} \tag{1}$$

■

Now let us suppose that the function f is monotonically increasing. Therefore for every $x \geq a$, $f(x) \geq f(a)$. In (1) use $b = f(a)$. Therefore we get

$$P(x \geq a) \leq \frac{E\{f(x)\}}{f(a)}$$

From this we can get the well known inequalities like

$$P(x \geq a) \leq \frac{E\{x\}}{a}$$

called Markov inequality which holds for $a > 0$ and nonnegetive x,

$$P(|x - E(x)| \geq a) \leq \frac{E\{|x - E(x)|^2\}}{a^2} = \frac{Var\{x\}}{a^2} \tag{2}$$

often called the Chebychev's Inequality[1] and the Chernoff's bound

$$P(x \geq a) \leq \frac{Ee^{sx}}{e^{sa}} \tag{3}$$

[1]In this note, $Var\{\}$ and σ^2 are used interchangeably for variance

3 Information Theoretic Bounds

3.1 Jensen's Inequality

Here we shall state and prove a generalized, measure theoretic proof for Jensen's inequality. In general, in probability theory, a more specific form of Jensen's inequality is famous. But before that we shall first define a convex function.

Definition A function $\phi(x)$ is defined to be convex in interval (a, b) if for every point x' in the interval (a, b) there exists an m such that

$$\phi(x) \geq m(x - x') + \phi(x') \tag{4}$$

for any $x \in (a, b)$ $\blacksquare$

Note that this definition can be proved to be equivallent to the definition of convex function as one in which the value of the function for any point in the interval of convexity is always below the line segment joining the end points of any subinterval of the convex interval containing that point. We chose this particular definition for simplyfying the proof of Jensen's inequality. Now without further a due, let us move to stating and proving Jensen's Inequality. (Note: Refer [4] for a similar generalized proof for Jensen's Inequality.)

Theorem 2 *Let f and μ be measurable functions of x which are finite a.e. on $A \subseteq R^n$. Now let $f\mu$ and μ be integrable on A and $\mu \geq 0$. If ϕ is a function which is convex in interval (a, b) which is the range of function f and $\int_A \phi(f)\mu$ exists then,*

$$\phi\left(\frac{\int_A f\mu}{\int_A \mu}\right) \leq \frac{\int_A \phi(f)\mu}{\int_A \mu} \tag{5}$$

Proof From our assumptions, the range of f is (a, b) which is the interval in which $\phi(x)$ is convex. Hence, consider the number,

$$x' = \frac{\int_A f\mu}{\int_A \mu}$$

Clearly it is within the interval (a, b). Further, from Equation (4) we have for almost every x,

$$\phi(f(x)) \geq m(f(x) - x') + \phi(x')$$

Multiplying by μ and integrating both sides we get,

$$\int_A \phi(f)\mu \geq m(\int_A f\mu - x' \int_A \mu) + \phi(x') \int_A \mu$$

Now see that $\int_A f\mu - x' \int_A \mu = 0$ and hence we have,

$$\int_A \phi(f)\mu \geq \phi(x') \int_A \mu$$

hence we get the result,

$$\int_A \phi(f)\mu \geq \phi(\frac{\int_A f\mu}{\int_A \mu}) \int_A \mu$$

■

Now note that if μ is a probability measure then, $\int_A \mu = 1$ and since expected value is simply lebesgue integral of function w.r.t. probability measure, we have

$$E[\phi(f(x))] \geq \phi(E[f(x)]) \tag{6}$$

for any function ϕ convex for the range of function $f(x)$.

Now a function $\phi(x)$ is convex if $\phi'(x)$ exists and is monotonically increasing and if second derivative exists and is nonnegetive. Therefore we can conclude that the function $-log\ x$ is a convex function. Therefore by Jensen's inequality, we have

$$E[-log\ f(x)] \geq log\ E[f(x)]$$

Now if we take function $f(x)$ to be the probability result we get the result that Entropy $H(P)$ is always greater than or equal to 0. If we make $f(x)$ the ratio of two probability measures dP and dQ, we get the result that relative entropy or KL divergence of two distributions is always non negetive. That is

$$D(P||Q) = E_P[log(\frac{dQ(x)}{dP(x)})] \geq log(E_P[\frac{dQ(x)}{dP(x)}]) = log(E_Q[dQ(x)]) = 0$$

Therefore,

$$D(P||Q) \geq 0$$

3.2 Han's Inequality

We shall first prove the Han's Inequality for entropy and then usingthe result, we shall prove the Han's Inequality for relative entropy.

Theorem 3 *Let $x_1, x_2, ..., x_n$ be discrete random variables from sample space X. Then*

$$H(x_1, ..., x_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) \tag{7}$$

Proof Note that $D(P_{X,Y}||P_X \times P_Y) = H(X) - H(X|Y)$. Since we already proved that Relative entropy is non-negetive, we have, $H(X) \geq H(X|Y)$. This in a vague way means that information (about some variable Y) can only reduce entropy or uncertainity ($H(X|Y)$), which makes intutive sense. Now consider the entropy,

$$H(x_1, ..., x_n) = H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) + H(x_i|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$$

Since we have already seen that $H(X) \geq H(X|Y)$, applying this we have,

$$H(x_1, ..., x_n) \leq H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) + H(x_i|x_1, ..., x_{i-1})$$

Summing both sides upto n we get

$$nH(x_1, ..., x_n) \leq \sum_{i=1}^{n} H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) + H(x_i|x_1, ..., x_{i-1})$$

Now note that by definition of conditional entropy, $H(X|Y) = H(X,Y) - H(Y)$. Therefore extending this to many variables we get the chain rule of entropy as,

$$H(x_1, ..., x_n) = \sum_{i=1}^{n} H(x_i|x_1, ..., x_{i-1})$$

Therefore using this chain rule,

$$nH(x_1, ..., x_n) \leq \sum_{i=1}^{n} H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) + H(x_1, ..., x_n)$$

Therefore,

$$H(x_1, ..., x_n) \leq \frac{1}{n-1} \sum_{i=1}^{n} H(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$$

■

Now we shall prove the Han's inequality for relative entropies. Let $x_1^n = x_1, x_2, ..., x_n$ be discrete random variables from sample space X just like in the previous case. Now let P and Q be probability distributions in the product space X^n and let P be a distribution such that $\frac{P_{X^n}(x_1,...,x_n}{dx_1^n}) = \frac{dP_1(x_1)}{dx_1}\frac{dP_2(x_2)}{dx_2}...\frac{dP_n(x_n)}{dx_n}$. That is distribution P assumes independence of the variables $(x_1, ..., x_n)$ with the probability density function of each variable x_i as $\frac{dP_i(x)}{dx_i}$. Let $x^{(i)} = (x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)$. Now with this setting we shall state and prove Han's relative entropy inequlity.

Theorem 4 *Given any distribution Q on product space X^n and a distribution P on X^n which assumes independence of variables $(x_1, ..., x_n)$*

$$D(Q||P) \leq \frac{1}{n-1}\sum_{i=1}^{n} D(Q^{(i)}||P^{(i)}) \tag{8}$$

where

$$Q^{(i)}(x^{(i)}) = \int_X \frac{dQ(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)}{dx_1^n}dx_i$$

and

$$P^{(i)}(x^{(i)}) = \int_X \frac{dP(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)}{dx_1^n}dx_i$$

Proof By definition of relative entropy,

$$D(Q||P) = \int_{X^n} dQlog(\frac{dQ}{dP}) = \int_{X^n} \frac{dQ}{dx_1^n}log(dQ) - \frac{dQ}{dx_1^n}log(\frac{dP}{dx_1^n})dx_1^n \tag{9}$$

In the above equation, consider the term $\int_{X^n} \frac{dQ}{dx_1^n}log(\frac{dP}{dx_1^n}))dx_1^n$. From our assumption about P we know that

$$\frac{dP(x_1^n)}{dx_1^n} = \frac{dP_1(x_1)}{dx_1}\frac{dP_2(x)}{dx_2}...\frac{dP_n(x_n)}{dx_n}$$

Now $P^{(i)}(x^{(i)}) = \int_X \frac{dP(x_1,...,x_{i-1},x_i,x_{i+1},...,x_n)}{dx_1^n}dx_i$, therefore,

$$\frac{dP(x_1^n)}{dx_1^n} = \frac{dP_i(x_i)}{dx_i}\frac{dP^{(i)}(x^{(i)})}{dx^{(i)}}$$

Therefore using this we get

$$\int_{X^n} \frac{dQ}{dx_1^n}log(\frac{dP(x_1^n)}{dx_1^n})dx_1^n = \frac{1}{n}(\sum_{i=1}^{n}\int_{X^n} \frac{dQ}{dx_1^n}(log(\frac{dP_i(x_i)}{dx_i}) + log(\frac{dP^{(i)}(x^{(i)})}{dx^{(i)}}))dx_1^n)$$

$$= \frac{1}{n}(\sum_{i=1}^{n}\int_{X^{n-1}}\frac{dQ^{(i)}}{dx^{(i)}}(log(\frac{dP^{(i)}(x^{(i)})}{dx^{(i)}}))dx^{(i)}) + \frac{1}{n}\int_{X^n}\frac{dQ}{dx_1^n}log(\frac{dP(x_1^n)}{dx_1^n})dx_1^n$$

Rearranging the terms we get,

$$\int_{X^n}\frac{dQ}{dx_1^n}log(\frac{dP(x_1^n)}{dx_1^n})dx_1^n = \frac{1}{n-1}(\sum_{i=1}^{n}\int_{X^{n-1}}\frac{dQ^{(i)}}{dx^{(i)}}(log(\frac{dP^{(i)}(x^{(i)})}{dx^{(i)}}))dx^{(i)})$$

Now also note that by Han's inequlity for entropy,

$$\int_{X^n}\frac{dQ}{dx_1^n}log(\frac{dQ(x_1^n)}{dx_1^n})dx_1^n \geq \sum_{i=1}^{n}\int_{X^n}\frac{dQ^{(i)}}{dx^{(i)}}log(\frac{dQ^{(i)}(x^{(i)})}{dx^{(i)}})dx^{(i)}$$

Therefore when we consider relative entropy given by Equation (9) we get,

$$D(Q||P)$$

$$\geq \frac{1}{n-1}\int_{X^{n-1}}\sum_{i=1}^{n}\frac{dQ^{(i)}}{dx^{(i)}}log(\frac{dQ^{(i)}(x^{(i)})}{dx^{(i)}})dx^{(i)} - \frac{1}{n-1}\int_{X^{n-1}}\sum_{i=1}^{n}\frac{dQ^{(i)}}{dx^{(i)}}log(\frac{dP^{(i)}(x^{(i)})}{dx^{(i)}})dx^{(i)}$$

Thus finally simplifying we get the required result as

$$D(Q||P) \leq \frac{1}{n-1}\sum_{i=1}^{n}D(Q^{(i)}||P^{(i)})$$

■

4 Inequalities of Sums of Random Variables

4.1 Hoeffding's Inequality

Theorem 5 *Let be independent bounded random variables such that the random variable x_i falls in the interval $[p_i, q_i]$. Then for any $a > 0$ we have*

$$P(\sum_{i=1}^{n}x_i - E(\sum_{i=1}^{n}x_i) \geq a) \leq e^{-\frac{2a^2}{\sum_{i=1}^{n}(q_i-p_i)^2}}$$

Proof Form the Chernoff's bound given by (3) we get,

$$P(x - E(x) \geq a) \leq \frac{Ee^{s(x-E(x))}}{e^{sa}}$$

Let $S_n = \sum_{i=1}^{n} x_i$. Therefore we have,

$$P(S_n - E(S_n) \geq a) \leq \frac{Ee^{s(S_n - E(S_n))}}{e^{sa}}$$

$$\leq e^{-sa} \prod_{i=1}^{n} Ee^{s(x_i - E(x_i))} \tag{10}$$

Now Let y be any random variable such that $p \leq y \leq q$ and $Ey = 0$. Then for any $s > 0$ due to convexity of exponential function we have

$$e^{sy} \leq \frac{y-p}{q-p} e^{sq} + \frac{q-y}{q-p} e^{sp}$$

Taking expectation on both sides we get,

$$Ee^{sy} \leq \frac{-p}{q-p} e^{sq} + \frac{q}{q-p} e^{sp}$$

Now let $\alpha = \frac{-p}{q-p}$. Therefore,

$$Ee^{sy} \leq (\alpha e^{s(q-p)} + (1-\alpha))e^{-s\alpha(q-p)}$$

$$\Rightarrow Ee^{sy} \leq e^{log(\alpha e^{s(q-p)} + (1-\alpha)) - s\alpha(q-p)}$$

$$\Rightarrow Ee^{sy} \leq e^{\phi(u)} \tag{11}$$

Where the function $\phi(u) = log(\alpha e^u + (1-\alpha)) - u\alpha$ and $u = s(q-p)$. Now using Taylor's theorem, we have for some η,

$$\phi(x) = \phi(0) + x\phi'(0) + \frac{x^2}{2}\phi''(\eta) \tag{12}$$

But we have $\phi(0) = 0$ and $\phi'(x) = \frac{\alpha e^x}{\alpha e^x + (1-\alpha)} - \alpha$. Therefore $\phi'(0) = 0$ Now,

$$\phi''(x) = \frac{\alpha(1-\alpha)e^x}{(1-\alpha+\alpha e^x)^2}$$

If we consider $\phi''(x)$ we see that the function is maximum when

$$\phi'''(x) = \frac{\alpha(1-\alpha)e^x}{(1-\alpha+\alpha e^x)^2} - \frac{2\alpha^2(1-\alpha)e^{2x}}{(1-\alpha+\alpha e^x)^3} = 0$$

$$\Rightarrow e^x = \frac{1-\alpha}{\alpha}$$

Therefore, for any x

$$\phi''(x) \leq \frac{((1-\alpha)^2)}{(2-2\alpha)^2} = \frac{1}{4}$$

Therefore from (12) and (11) we have

$$Ee^{sy} \leq e^{\frac{u^2}{8}}$$

Therefore for any $p \leq y \leq q$

$$Ee^{sy} \leq e^{\frac{s^2(q-p)^2}{8}} \tag{13}$$

Using this in (10) we get,

$$P(S_n - E(S_n) \geq a) \leq e^{-sa} e^{s^2 \frac{\sum_{i=1}^{n}(q_i-p_i)^2}{8}}$$

Now we find the best bound by minimizing the L.H.S. of the above equation w.r.t s. Therefore we have

$$\frac{de^{s^2 \frac{\sum_{i=1}^{n}(q_i-p_i)^2}{8} - sa}}{ds} = e^{s^2 \frac{\sum_{i=1}^{n}(q_i-p_i)^2}{8} - sa}(2s\frac{\sum_{i=1}^{n}(q_i-p_i)^2}{8} - a) = 0$$

Therefore for the best bound we have

$$s = \frac{4a}{\sum_{i}^{n}(q_i-p_i)^2}$$

and correspondingly we get

$$P(S_n - E(S_n) \geq a) \leq e^{-\frac{2a^2}{\sum_{i=1}^{n}(q_i-p_i)^2}} \tag{14}$$

■

Now an interesting result from the Hoeffding's inequality often used in learning theory is to bound not the differnce in sum and its corresponding expectation but the emperical average of loss function and its expectation. This is done by using (14) as,

$$P(S_n - E(S_n) \geq na) \leq e^{-\frac{2(an)^2}{\sum_{i=1}^{n}(q_i-p_i)^2}}$$

$$\Rightarrow P(\frac{S_n}{n} - \frac{E(S_n)}{n} \geq a) \leq e^{-\frac{2(aN)^2}{\sum_{i=1}^{n}(q_i-p_i)^2}}$$

Now $E_n x = \frac{S_n}{n}$ denotes the emperical average of x and $\frac{E(S_n)}{n} = Ex$. Therefore we have

$$P(E_n(x) - E(x) \geq a) \leq e^{-\frac{2n^2a^2}{\sum_{i=1}^{n}(q_i-p_i)^2}} \tag{15}$$

4.2 Bernstein's Inequality

Hoeffding's inequality does not use any knowledge about the distribution of variables. The Bernstein's inequality [7] uses the variance of the distribution to get a tighter bound.

Theorem 6 *Let $x_1, x_2, ..., x_n$ be independent bounded random variables such that $Ex_i = 0$ and $|x_i| \leq \varsigma$ with probability 1 and let $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var\{x_i\}$ Then for any $a > 0$ we have*

$$P(\frac{1}{n}\sum_{i=1}^{n} x_i \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2 + 2\varsigma\epsilon/3}}$$

Proof We need to re-estimate the new bound starting from (10). Let

$$F_i = \sum_{r=2}^{\infty} \frac{s^{r-2}E(x_i^r)}{r!\sigma_i^2}$$

where $\sigma_i^2 = Ex_i^2$. Now $e^x = 1 + x + \sum_{r=2}^{\infty} \frac{x^r}{r!}$. Therefore,

$$Ee^{sx_i} = 1 + sEx_i + \sum_{r=2}^{\infty} \frac{E(x_i^r)}{r!}$$

Since $Ex_i = 0$ we have,

$$\begin{aligned} Ee^{sx_i} &= 1 + F_i s^2 \sigma_i^2 \\ &\leq e^{F_i s^2 \sigma_i^2} \end{aligned}$$

Consider the term Ex_i^r. Since expectation of a function is just the Lebesgue integral of the function with respect to probability measure, we have $Ex_i^r = \int_P x_i^{r-1} x_i$. Using Schwarz's inequality we get,

$$Ex_i^r = \int_P x_i^{r-1} x_i \leq (\int_P |x_i^{r-1}|^2)^{\frac{1}{2}} (\int_P |x_i|^2)^{\frac{1}{2}}$$

$$\Rightarrow Ex_i^r \leq \sigma_i (\int_P |x_i^{r-1}|^2)^{\frac{1}{2}}$$

Proceeding to use the Schwarz's inequality recursively n times we get

$$Ex_i^r \leq \sigma_i^{1+\frac{1}{2}+\frac{1}{2}^2+...+\frac{1}{2}^{n-1}} (\int_P |x_i^{(2^n r - 2^{n+1} - 1)}|)^{\frac{1}{2^n}}$$

$$\{\sigma_i^{2(1-\frac{1}{2}^n)}(\int_P |x_i^{(2^n r-2^{n+1}-1)}|)^{\frac{1}{2^n}}\}$$

Now we know that $|x_i| \leq \varsigma$. Therefore

$$(\int_P |x_i^{(2^n r-2^{n+1}-1)}|)^{\frac{1}{2^n}} \leq (\varsigma^{(2^n r-2^{n+1}-1)})^{\frac{1}{2^n}}$$

Hence we get

$$Ex_i^r \leq \{\sigma_i^{2(1-\frac{1}{2}^n)}\varsigma^{(r-2-\frac{1}{2^n})}\}$$

Taking limit n to infinity we get

$$Ex_i^r \leq lim_{n\to\infty}\{\sigma_i^{2(1-\frac{1}{2}^n)}\varsigma^{(r-2-\frac{1}{2^n})}\}$$

$$\Rightarrow Ex_i^r \leq \sigma_i^2\varsigma^{r-2} \tag{16}$$

Therefore,

$$F_i = \sum_{r=2}^{\infty}\frac{s^{r-2}E(x_i^r)}{r!\sigma_i^2} \leq \sum_{r=2}^{\infty}\frac{s^{r-2}\sigma_i^2\varsigma^{r-2}}{r!\sigma_i^2}$$

Therefore,

$$F_i \leq \frac{1}{s^2\varsigma^2}\sum_{r=2}^{\infty}\frac{s^r\varsigma^r}{r!} = \frac{1}{s^2\varsigma_i^2}(e^{s\varsigma}-1-s\varsigma)$$

Applying this to (16) we get,

$$Ex_i^r \leq e^{s^2\sigma_i^2\frac{(e^{s\varsigma}-1-s\varsigma)}{s^2\varsigma^2}}$$

Now using (10) and the fact that $\sigma^2 = \frac{\sigma_i^2}{n}$ we get,

$$P(S_n \geq a) \leq e^{-sa}e^{s^2n\sigma^2\frac{(e^{s\varsigma}-1-s\varsigma)}{s^2\varsigma^2}} \tag{17}$$

Now to obtain the closest bound we minimize R.H.S w.r.t s. Therefore we get

$$\frac{de^{s^2n\sigma^2(\frac{(e^{s\varsigma}-1-s\varsigma)}{s^2\varsigma^2})-sa}}{ds} = e^{s^2n\sigma^2(\frac{(e^{s\varsigma}-1-s\varsigma)}{s^2\varsigma^2})-sa}(n\sigma^2(\frac{(\varsigma e^{s\varsigma}-\varsigma)}{\varsigma^2})-a) = 0$$

Therefore to get a tighter bound we have

$$(\frac{(e^{s\varsigma}-1)}{\varsigma}) = \frac{a}{n\sigma^2}$$

Therefore we have

$$s = \frac{1}{\varsigma} log(\frac{a\varsigma}{n\sigma^2} + 1)$$

Using this s in (17) we get

$$P(S_n \geq a) \leq e^{\frac{n\sigma^2}{\varsigma^2}(\frac{a\varsigma}{n\sigma^2} - log(\frac{a\varsigma}{n\sigma^2}+1)) - \frac{a}{\varsigma} log(\frac{a\varsigma}{n\sigma^2}+1)}$$

$$\leq e^{\frac{n\sigma^2}{\varsigma^2}(\frac{a\varsigma}{n\sigma^2} - log(\frac{a\varsigma}{n\sigma^2}+1) - \frac{a\varsigma}{n\sigma^2} log(\frac{a\varsigma}{n\sigma^2}+1))}$$

Let $H(x) = (1+x)log(1+x) - x$ Therefore we get

$$P(S_n \geq a) \leq e^{\frac{-n\sigma^2}{\varsigma^2} H(\frac{a\varsigma}{n\sigma^2})} \tag{18}$$

This is called the Bennett's inequality [?]. We can derive the Bernstien's inequality by further bounding the function $H(x)$. Let function $G(x) = \frac{3}{2}\frac{x^2}{x+3}$. We see that $H(0) = G(0) = H'(0) = G'(0) = 0$ and we see that $H''(x) = \frac{1}{x+1}$ and $G''(x) = \frac{27}{(x+3)^3}$. Therefore $H''(0) \geq G''(0)$ and further if $f^n(x)$ of a function f represents the n^{th} derivative of the function then we have $H^n(0) \geq G^n(0)$ for any $\{n \geq 2\}$. Therefore as a consequence of Taylor's theorem we have

$$H(x) \geq G(x) \forall x \geq 0$$

Therefore applying this to (18) we get

$$P(\sum_i^n x_i \geq a) \leq e^{\frac{-n\sigma^2}{\varsigma^2} G(\frac{a\varsigma}{n\sigma^2})}$$

$$\Rightarrow P(\sum_i^n x_i \geq a) \leq e^{\frac{-a^2}{2(a\varsigma + 3n\sigma^2)}}$$

Now let $a = n\epsilon$. Therefore,

$$P(\sum_i^n x_i \geq n\epsilon) \leq e^{\frac{-\epsilon^2 n^2}{2(n\epsilon\varsigma + 3n\sigma^2)}}$$

Therefore we get,

$$P(\frac{1}{n}\sum_{i=1}^n x_i \geq \epsilon) \leq e^{-\frac{n\epsilon^2}{2\sigma^2 + 2\varsigma\epsilon/3}} \tag{19}$$

■

An interesting phenomenon here is that if $\sigma < \epsilon$ then the upper bound grows as $e^{-n\epsilon}$ rather than $e^{-n\epsilon^2}$ as suggested by Hoeffding's inequality (14).

5 Inequalities of Functions of Random Variables

5.1 Efron Stien's Inequality

Till now we only considered sum of R.V.s. Now we shall consider functions of R.V.s. The Efron Stien [9] inequality is one of the tightest bounds known.

Theorem 7 *Let $S : X^n \to R$ be a measurable function which is invariant under permutation and let the random variable Z be given by $Z = S(x_1, x_2, ..., x_n)$. Then we have*

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

where $Z_i' = S(x_1, ..., x_i', ..., x_n)$ where $\{x_1', ...x_n'\}$ is another sample from the same distribution as that of $\{x_1, ...x_n\}$

Proof Let

$$E_i Z = E[Z|x_1, ..., x_{i-1}, x_{i+1}, ..., x_n]$$

and let $V = Z - EZ$. Now if we define V_i as

$$V_i = E[Z|x_1, ..., x_i] - E[Z|x_1, ..., x_{i-1}], \forall i = 1, .., n$$

then $V = \sum_{i=1}^{n} V_i$ and

$$Var(Z) = EV^2 = E[(\sum_{i=1}^{n} V_i)^2] = E[\sum_{i=1}^{n} V_i^2] + 2E[\sum_{i>j} V_i V_j]$$

Now $E[XY] = E[E[XY|Y]] = E[YE[X|Y]]$, therefore $E[V_j V_i] = E[V_i E[V_j|x_1, ..., x_i]]$. But since $i > j$ $E[V_j|x_1, ..., x_i] = 0$. Therefore we have,

$$Var(Z) = E[\sum_{i=1}^{n} V_i^2] = \sum_{i=1}^{n} E[V_i^2]$$

Now let $E_{x_i^j}$ represent expectation w.r.t variables $\{x_i, ..., x_j\}$.

$$Var(Z) = \sum_{i=1}^{n} E[(E[Z|x_1, ..., x_i] - E[Z|x_1, ..., x_{i-1}])^2]$$

$$= \sum_{i=1}^{n} E_{x_1^i}[(E_{x_{i+1}^n}[Z|x_1,...,x_i] - E_{x_i^n}[Z|x_1,...,x_{i-1}])^2]$$

$$= \sum_{i=1}^{n} E_{x_1^i}[(E_{x_{i+1}^n}[Z|x_1,...,x_i] - E_{x_{i+1}^n}[E_{x_i}[Z|x_1,...,x_{i-1}]])^2]$$

However, x^2 is a convex function and hence we can apply Jensens inequality (6) and hence get,

$$Var(Z) \leq \sum_{i=1}^{n} E_{x_1^i, x_{i+1}^n}[(Z - E_{x_i}[Z|x_1,...,x_{i-1},x_{i+1},...,x_n])^2]$$

Therefore,

$$Var(Z) \leq \sum_{i=1}^{n} E[(Z - E_i[Z])^2]$$

Where $E_i[Z] = E[Z|x_1,...,x_{i-1},x_{i+1},...,x_n]$. Now let x and y be 2 independent samples from the same distribution.

$$E(x-y)^2 = E[x^2 + y^2 - 2xy] = 2E[x^2] - 2(E[x])^2$$

Hence, if x and y are i.i.d's, then $Var\{x\} = E[\frac{1}{2}(x-y)^2]$. Thus we have,

$$E_i[(Z - E_i[Z])^2] = \frac{1}{2}E_i[(Z - Z_i')^2]$$

Thus we have the Efron-Stein inequality as

$$Var(Z) \leq \frac{1}{2}\sum_{i=1}^{n} E[(Z - Z_i')^2] \tag{20}$$

■

Notice that if function S is the sum of the random variables the inequality becomes an equality. Hence the bound is tight. It is often refered to as jacknife bound.

5.2 McDiarmid's Inequality

Theorem 8 *Let $S : X^n \to R$ be a measurable function which is invariant under permutation and let the random variable Z be given by $Z = S(x_1, x_2, ..., x_n)$. Then for any $a > 0$ we have*

$$P(|Z - E[Z]| \geq a) \leq 2e^{2\frac{a^2}{\sum_{i=1}^{n} \varsigma_i^2}}$$

whenever the function has bounded difference [10]. That is

$$sup_{x_1,...,x_n,x_i'}|S(x_1,x_2,...,x_n) - S(x_1,...,x_i',...,x_n)| \leq \varsigma_i$$

where $Z_i' = S(x_1,...,x_i',...,x_n)$ *where* $\{x_1',...x_n'\}$ *is a sample from the same distribution as* $\{x_1,...x_n\}$

Proof Using Chernoff's bound (3) we get

$$P(Z - E[Z] \geq a) \leq e^{-sa}e^{E[Z-E[Z]]}$$

Now let,

$$V_i = E[Z|x_1,...,x_i] - E[Z|x_1,...,x_{i-1}], \forall i = 1,..,n$$

then $V = \sum_{i=1}^{n} V_i = Z - E[Z]$. Therefore,

$$P(Z - E[Z] \geq a) \leq e^{-sa}E[e^{\sum_{i=1}^{n} sV_i}] = e^{-sa}\prod_{i=1}^{n} E[e^{sV_i}] \tag{21}$$

Now Let V_i be bounded by the interval $[L_i, U_i]$. We know that $|Z - Z_i')| \leq \varsigma_i$, hence it follows that $|V_i| \leq \varsigma_i$ and hence $|U_i - L_i| \leq \varsigma_i$. Using (13) on $E[e^{sV_i}]$ we get,

$$E[e^{sV_i}] \leq e^{\frac{s^2(U_i-L_i)^2}{8}} \leq e^{\frac{s^2\varsigma_i^2}{8}}$$

Using this in (21) we get,

$$P(Z - E[Z] \geq a) \leq e^{-as}\prod_{i=1}^{n} e^{\frac{s^2\varsigma_i^2}{8}} = e^{s^2\sum_{i=1}^{n}\frac{\varsigma_i^2}{8} - sa}$$

Now to make the bound tight we simply minimize it with respect to s. Therefore to do that,

$$2s\sum_{i=1}^{n}\frac{\varsigma_i^2}{8} - a = 0$$

$$\Rightarrow s = \frac{4a}{\sum_{i=1}^{n}\varsigma_i^2}$$

Therefore the bound is given by,

$$P(Z - E[Z] \geq a) \leq e^{(\frac{4a}{\sum_{i=1}^{n}\varsigma_i^2})^2\sum_{i=1}^{n}\frac{\varsigma_i^2}{8} - (\frac{4a^2}{\sum_{i=1}^{n}\varsigma_i^2})}$$

$$\Rightarrow P(Z - E[Z] \geq a) \leq e^{-(\frac{2a^2}{\sum_{i=1}^{n} \varsigma_i^2})}$$

Hence we get,

$$P(|Z - E[Z]| \geq a) \leq 2e^{-(\frac{2a^2}{\sum_{i=1}^{n} \varsigma_i^2})} \tag{22}$$

■

5.3 Logarithmic Sobolev Inequality

This inequality is very useful in deriving simple proofs for many known bounds using a method famous as Ledoux method [11]. Before jumping into the therem, let us first prove a useful lemma.

Lemma 9 *For any positive random variable y and $\alpha > 0$ we have*

$$E\{ylog\ y\} - Ey\ log\ Ey\ \leq\ E\{ylog\ y\ -\ ylog\ \alpha - (y - \alpha)\} \tag{23}$$

Proof For any $(x > 0\ )$ we have, $log\ x \leq x - 1$. Therefore,

$$log\frac{\alpha}{Ey} \leq \frac{\alpha}{Ey} - 1$$

Therefore,

$$Ey\ log\frac{\alpha}{Ey} \leq \alpha - Ey$$

adding $E\{ylog\ y\}$ on both sides we get,

$$Ey\ log\alpha - Ey\ logEy + E\{ylog\ y\} \leq \alpha - Ey + E\{ylog\ y\}$$

Therefore simplifying we get the required result as

$$E\{ylog\ y\} - Ey\ log\ Ey\ \leq\ E\{ylog\ y\ -\ ylog\ \alpha - (y - \alpha)\}$$

■

Now just like in the previous two sections, let $S : X^n \rightarrow R$ be a measurable function which is invariant under permutation and let the random variable Z be given by $Z = S(x_1, x_2, ..., x_n)$. Here we assume independence of $(x_1, x_2, ..., x_n)$. $Z_i' = S(x_1, ..., x_i', ..., x_n)$ where $\{x_1', ...x_n'\}$ is another sample from the same distribution as that of $\{x_1, ...x_n\}$. Now we are ready to state and prove the logarithmic Sobolev inequality.

Theorem 10 *If function $\psi(x) = e^x - x - 1$ then,*

$$sE\{Ze^{sZ}\} - E\{e^{sZ}\}log\ E\{e^{sZ}\} \leq \sum_{i=1}^{n} E\{e^{sZ}\psi(-s(Z - Z_i'))\} \tag{24}$$

Proof From Lemma 1 (Equation(23)) we have for any positive variable Y and $\alpha = Y_i' > 0$,

$$E_i\{Ylog\ Y\} - E_iY\ log\ E_iY \ \leq \ E_i\{Ylog\ Y \ - \ Ylog\ Y_i' - (Y - Y_i')\}$$

Now let $Y = e^{sZ}$ and $Y_i' = e^{sZ_i'}$ then,

$$E_i\{Ylog\ Y\} - E_iY\ log\ E_iY \ \leq \ E_i\{e^{sZ}(sZ - sZ_i') - e^{sZ}(1 - e^{sZ_i'-sZ})\}$$

Writing it in terms of function $\psi(x)$ we get,

$$E_i\{Ylog\ Y\} - E_iY\ log\ E_iY \ \leq \ E_i\{e^{sZ}\psi(-s(Z - Z_i'))\} \tag{25}$$

Now let measure P denote the distribution of $(x_1, ..., x_n)$ and let distribution Q be given by,

$$dQ(x_1, ..., x_n) = dP(x_1, ..., x_n)Y(x_1, ..., x_n)$$

Then we have,

$$D(Q||P) = E_Q\{log\ Y\} = E_P\{Ylog\ Y\} = E\{Ylog\ Y\}$$

Now since Y is positive, we have

$$D(Q||P) \geq E\{Ylog\ Y\} - EYlog\ EY \tag{26}$$

However by Han's inequality for relative entropy, rearranging Equation (8) we have,

$$D(Q||P) \leq \sum_{i=1}^{n} D(Q||P) - D(Q^{(i)}||P^{(i)}) \tag{27}$$

Now we have already shown that $D(Q||P) = E\{Ylog\ Y\}$ and further, $E\{E_i\{Ylog\ Y\}\} = E\{Ylog\ Y\}$ therefore, $D(Q||P) = E\{E_i\{Ylog\ Y\}\}$. Now by definition,

$$\frac{dQ^{(i)}(x^{(i)})}{dx^{(i)}} = \int_X \frac{dQ(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)}{dx_1^n}dx_i$$

$$= \int_X \frac{Y(x_1, ..., x_n)dP(x_1, ..., x_{i-1}, x_i, x_{i+1}, ..., x_n)}{dx_1^n}dx_i$$

Due to the independence assumption of the sample, we can rewrite the above as,

$$\frac{dQ^{(i)}(x^{(i)})}{dx^{(i)}} = \frac{dP(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)}{dx^{(i)}} \int_X Y(x_1, ..., x_n) dP_i(x_i)$$

$$= \frac{dP(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n)}{dx^{(i)}} E_i Y$$

Therefore $D(Q^{(i)}||P^{(i)})$ is given by,

$$D(Q^{(i)}||P^{(i)}) = \int_{X^{n-1}} E_i Y log\ E_i Y dP(x^i) = E\{E_i Y log\ E_i Y\}$$

Now using this we get,

$$\sum_{i=1}^{n} D(Q||P) - D(Q^{(i)}||P^{(i)}) = \sum_{i=1}^{n} E\{E_i\{Y log\ Y\}\} - E\{E_i Y log\ E_i Y\}$$

Therefore, using this in Equation (27) we get,

$$D(Q||P) \leq \sum_{i=1}^{n} E\{E_i\{Y log\ Y\}\} - E\{E_i Y log\ E_i Y\}$$

and using this inturn with Equation (26) we get,

$$E\{Y log\ Y\} - EY log\ EY \leq \sum_{i=1}^{n} E\{E_i\{Y log\ Y\} - E_i Y log\ E_i Y\}$$

Using the above with Equation (25) and substituting Y with e^{sZ} we get the first required result as,

$$sE\{Ze^{sZ}\} - E\{e^{sZ}\}log\ E\{e^{sZ}\} \leq \sum_{i=1}^{n} E\{e^{sZ}\psi(-s(Z - Z_i'))\}$$

■

6 Symmetrization Lemma

The symmetrisization lemma is probably one of the easier bounds we review in this notes. However, it is extremely powerful since it allows us to bound the difference between of emperical mean of a function and its expected value, using the difference between emperical means of the function for 2 independent samples of the same size as the original sample. Note that in most literature, the symmetrization lemma stated and proved only bounds zero one functions like loss function or the actual classification function. Here we derive a more generalized version where we prove the lemma for bounding the difference between expectation of any measurable function with bounded variance and its emperical mean.

Lemma 11 *Let $f : X \rightarrow R$ be a measurable function such that $Var\{f\} \leq C$. Let $\hat{E_n}\{f\}$ be the emperical mean of the function $f(x)$ estimated using a set $(x_1, x_2, ..., x_n)$ of n independent identical samples from space X. Then for any $a > 0$ if $n > \frac{8C}{a^2}$ we have,*

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) \geq \frac{1}{2}P(|E\{f\} - \hat{E}_n\{f\}| > a) \tag{28}$$

where $\hat{E_n}'\{f\}$ and $\hat{E_n}''\{f\}$ stand for emperical mean of the function $f(x)$ estimated using samples $(x_1', x_2', ..., x_n')$ and $(x_1'', x_2'', ..., x_n'')$ respectively

Proof By the definition of probability,

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) = \int_{X^{2n}} 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP$$

Where function 1_z is 1 for any $z \geq 0$ and 0 otherwise. Since X^{2n} is the product space $X^n \times X^n$, using Fubini's theorem we have,

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) = \int_{X^n} \int_{X^n} 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP'$$

Now since the set $Y = \{(x_1, x_2, ..., x_n) : |E\{f\} - \hat{E_n}\{f\}| > a\}$ is a subset of X^n and term inside the integral is always non-negetive,

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) \geq \int_Y \int_{X^n} 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP'$$

Now let Let $Z = \{(x_1, x_2, ..., x_n) : |\hat{E_n}\{f\} - E\{f(x)\}| \leq \frac{a}{2}\}$. Clearly for any sample $(x_1, x_2, ..., x_{2n})$, if $(x_1, x_2, ..., x_n) \in Y$ and $(x_{n+1}, x_{n+2}, ..., x_{2n}) \in Z$ it implies that $(x_1, x_2, ..., x_{2n})$ is a sample such that $|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{a}{2}$ Therefore, coming back to the integral since $Z \subset X^n$,

$$\int_Y \int_{X^n} 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP' \geq \int_Y \int_Z 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP'$$

Now since the integral is over Y and Z half spaces as we saw earlier,

$$\int_Y \int_Z 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP' = \int_Y \int_Z 1 dP'' dP'$$
$$= \int_Y P(|\hat{E_n}'\{f\} - E\{f\}| \leq \frac{1}{2}a) dP'$$

Therefore,

$$\int_Y \int_{X^n} 1_{[|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| - \frac{1}{2}a]} dP'' dP' \geq \int_Y P(|\hat{E_n}'\{f\} - E\{f\}| \leq \frac{1}{2}a) dP'$$

Now,

$$P(|\hat{E_n}'\{f\} - E\{f\}| \leq \frac{1}{2}a) = 1 - P(|\hat{E_n}''\{f\} - E\{f\}| > \frac{1}{2}a)$$

Using Equation (2) (often called the chebyshev's inequality) we get,

$$P(|\hat{E_n}''\{f\} - E\{f\}| > \frac{1}{2}a) \leq \frac{4Var\{f\}}{na^2} \leq \frac{4C}{na^2}$$

Now if we choose n such that $n > \frac{8C}{a^2}$ as per our assumption then,

$$P(|\hat{E_n}''\{f\} - E\{f\}| > \frac{1}{2}a) \leq \frac{1}{2}$$

Therefore,

$$P(|\hat{E_n}''\{f\} - E\{f\}| \leq \frac{1}{2}a) \geq 1 - \frac{1}{2} = \frac{1}{2}$$

Puting this back in the integral we get,

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) \geq \int_Y \frac{1}{2} dP' = \frac{1}{2}\int_Y dP'$$

Therefore we get the final result as,

$$P(|\hat{E_n}'\{f\} - \hat{E_n}''\{f\}| \geq \frac{1}{2}a) \geq \frac{1}{2} P(|E\{f\} - \hat{E}_n\{f\}| > a)$$

■

Note that if we make the function f to be a zero one function then the maximum possible variance is $\frac{1}{4}$. Hence if we set C to $\frac{1}{4}$ then the condition under which the inequality holds becomes, $n > \frac{2}{a^2}$. Further note that if we choose the zero one function $f(x)$ such that it is 1 when say x is a particular value and 0 if not, then the result basically bounds the absolute difference between probability of the event of x taking a particular value and the frequency estimate of x taking that value in a sample of size n using the difference in the frequencies of occurance of the value in 2 independent samples of size n.

References

[1] V.N. Vapnik, A.YA. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities", Theory of Probability and its Applications, 16(2):264-281, 1971.

[2] Gabor Lugosi, "Concentration-of-Measure Inequalities" , Lecture Notes, http://www.econ.upf.es/ lugosi/anu.pdf

[3] A.N. Kolmogorov, "Foundations of Probability", Chelsea Publications, NY, 1956.

[4] Richard L. Wheeden, Antoni Zygmund, "Measure and Integral:An Introduction to Real Analysis", International Series of Monographs in Pure and Applied Mathematics.

[5] W. Hoeffding, "Probability Inequalities for Sums of Bounded Random Variables", Journal of the American Statistical Association, 58:13-30, 1963.

[6] G. Benette, "Probability Inequalities for Sum of Independent Random Variables", Journal of the American Statistical Association, 57:33-45, 1962.

[7] S.N. Bernstein, "The Theory of Probabilities", Gastehizdat Publishing House, Moscow, 1946.

[8] V. V. Petrov, "Sums of Independent Random Variables", Translated by A.A. Brown, Springer Verlag, 1975.

[9] B. Efron, C. Stein, "The Jacknife Estimate of Variance", Annals of Statistics, 9:586-96, 1981.

[10] C. Mcdiarmid, "On the Method of Bounded Differences" , In Surveys in Combinatorics, LMS lecture. note series 141, 1989.

[11] Michel Ledoux, "The Concentration of Measure Phenomenon", American Mathematical Society, Mathematical Surveys and Monographs, Vol 89, 2001.

[12] Olivier Bousquet, Stephane Boucheron, and Gabor Lugosi,"Introduction to Statistical Learning Theory" ,Lecture Notes ,http://www.econ.upf.es/ lugosi/mlss_slt.pdf