Introduction to Concentration Inequalities

Kumar Abhishek
IIIT Hyderabad, India.
kumar.abhishek@research.iiit.ac.in

Sneha Maheshwari
IIT Roorkee, India.
smaheshwari@ma.iitr.ac.in

Sujit Gujar
IIIT Hyderabad, India.
sujit.gujar@iiit.ac.in

Contents

Abstract

In this report, we aim to exemplify concentration inequalities and provide easy to understand proofs for it. Our focus is on the inequalities which are helpful in the design and analysis of machine learning algorithms.

1 What are Concentration Inequalities?

Concentration inequalities furnish us bounds on how random variables deviate from a value (typically, expected value) or help us to understand how well they are concentrated. A random variable with high concentration is one that is close to its mean (or value) with high probability (more than a certain threshold). For example, the strong law of large numbers or weak law of large numbers say that under mild conditions, if we sum a large number of independant random variables, with high probability, the sum is close to the expected value. These are elementary examples of the concentration we are talking about here.

Concentration inequalities quantify the statements of random fluctuations of functions of random variables, typically by bounding the probability that such a function differs from its expected value (or from its median) by more than a certain amount.

In the last decades, many researchers in a variety of areas were thriving to define concentration inequalities because of their importance in numerous applications.

This report is organized as follows. In Section 2, we provide examples where such concentration inequalities are useful. In Section 3, we state and prove, (i) Markov's Inequality, (ii) Chebyshev's Inequality, (iii) Chernoff Bound, (iv) Hoeffding's Lemma, (v) Hoeffding's Inequality, (vi) Azuma's Inequality. In Section 4, we state advanced topics about concentration inequalities, (i) Bennett's Inequality, (ii) Bernstein's Inequality, (iii) Efron-Stein Inequality, (iv) McDiarmid's Inequality.

2 Motivation

2.1 General example

Let's start with a simple example,

Problem: Estimation of probability for a biased coin
Given a biased coin having an unknown probability 'p' of occurring head, we need to estimate the value of p.

- If we toss the coin once if it comes head then the probability of head will be 1. But we are not at all confident for the probability being 1.
- If we toss the coin 100 times and head appears 65 times then we are a bit more confident for the probability being 0.65.
- Similarly, if we toss a coin lets say million times and the head is outcome $6,00,000$ times then we can say that 'p' is 0.60 with a very high confidence.

Thus, to quantify the level of confidence with respect to the number of trials, we can use *concentration inequalities* to have better estimates of 'p'.

2.2 Statistics

In statistics we umpteen applications of concentration inequalities, let's see one of the example,

Problem: Estimation of the population parameter.
In statistics, from an unknown population distribution, we want to infer information through sampling. (For example, one might want to know the population mean of age with probability of empirical mean to be actual mean, etc.)

Following are the questions we need to address:

- How can we estimate the confidence interval (range of values) which would be a good estimate?
- How can we determine the level of significance (confidence level) of that estimate?

We can get the answers to both questions through *concentration inequalities*.

2.3 Algorithms

Zillions of analyses in algorithms (mainly in theoretical computer science) uses concentration inequalities to give upper or lower bounds about the performance of algorithms with a certain probability.
For example,

- **MAX cut problem**: We can solve this problem approximately and to analyze the probability that this algorithm gives a maximum cut we can use Reverse Markov inequality (converse of Markov inequality).

2.4 Machine Learning

In Machine learning, concentration inequalities are profoundly used in analyzing different aspects of learning algorithms. For example,

- Multi-Armed bandits problem: We use concentration inequalities to analyze algorithms such as UCB algorithms, Thompson Sampling for their *regret*, a measure on performance of a MAB algorithm. Here we need good estimates of rewards of each arm with high confidence.

2.5 Miscellaneous

Among the areas of applications, without trying to be exhaustive,

- Statistics
- Learning Theory which includes supervised learning, unsupervised learning, online learning, and reinforcement learning.
- Discrete mathematics
- Statistical mechanics
- Information theory
- High-Dimensional geometry

and the list goes on. In the next section, we will prove important concentration inequalities and illustrate with examples.

3 Inequalities

We begin with the most elegant, yet powerful Markov inequality. Then, we go on explaining Chebyshev's inequality, Chernoff bound, Hoeffding's Lemma and inequality. At the end of this section, we state and prove Azuma's inequality.

3.1 Markov's Inequality

For a positive random variable $X \geq 0$ and $a > 0$, the probability that X is no less than a is less than or equal to the expectation of X divided by a:

$$Pr[X \geq a] \leq \frac{E(X)}{a}$$

Proof.

$$\begin{aligned} E[X] &= \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx \\ &\geq \int_a^\infty xp(x)dx \geq a\int_a^\infty p(x)dx \\ &\geq aPr(X \geq a) \end{aligned} \tag{1}$$

By rearranging the terms,

$$Pr[X \geq a] \leq \frac{E(X)}{a}$$

■

Example 1. *Let R be the weight distribution of a population with $E[R] = 100$. Calculate the probability that a random person weigh at least 200 pounds.*

Solution: As weight is always positive, we can apply Markov's inequality,

$$Pr[R \geq 200] \leq \tfrac{100}{200} \leq \tfrac{1}{2}$$

Corollary: Reverse Markov inequality

Given maximum value 'U' of a random variable 'X',

$$Pr[X \leq a] \leq \tfrac{U-E[X]}{U-a}$$

Note: In the corollary there is no need for the random variable 'X' to be positive.

Proof.

$$\begin{aligned} Pr[X \leq a] &= Pr[U - X \geq U - a] \\ &\leq \frac{E[U-X]}{U-a} \qquad \text{(Applying Markov's inequality)} \\ &\leq \frac{U-E[X]}{U-a} \end{aligned} \tag{2}$$

■

Example 2. *Let 'X' be the random variable denoting the marks of random student. Maximum marks possible is 100 (U) and expected marks 75. What is the probability that a random student scores 50 or less?*
Solution: *We can directly apply reverse Markov inequality,*

$$Pr[X \leq 50] \leq \tfrac{100-75}{100-50} \leq \tfrac{1}{2}$$

Example 3. *Suppose we use Markov's inequality to bound the probability of obtaining more than 3n/4 heads in a sequence of n fair coin flips. Let*

$$X_i = \begin{cases} 1 & \text{if the } i^{th} \text{ coin flip is head} \\ 0 & \text{otherwise} \end{cases}$$

and let $X = \sum_{i=1}^n X_i$ denote the number of heads in the n coin flips. Since $E[X_i] = Pr(X_i = 1) = 1/2$, it follows that $E[X] = \sum_{i=1}^n E[X_i] = n/2$. Applying Markov's inequality, we obtain

$$P(X \geq 3n/4) \leq \frac{E[X]}{3n/4} = \frac{n/2}{3n/4} = \frac{2}{3}$$

Features:

- Upside - This needs almost no assumptions about the random variable.
- Downside - It gives weaker bounds.

Markov's inequality is generally used where the random variable is too complicated to be analyzed by more powerful [1]inequalities.

[1]Powerful inequalities are those whose confidence level are higher for small confidence interval

3.2 Chebyshev's Inequality

For a random variable X expectation and variance should be finite, then $\forall a > 0$,

$$Pr(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2}$$

Proof.

$$\begin{aligned} Pr(|X - E[X]| \geq a) &= Pr[(X - E[x])^2 \geq a^2] \\ &\leq \frac{E[(X - E[X])^2]}{a^2} \text{ (Applying Markov's inequality)} \\ &= \frac{Var[X]}{a^2} \end{aligned} \tag{3}$$

■

Example 4. *Let X be the IQ of random variable with $X \geq 0$, $E[X] = 100$ and $\sigma(X) = 15$. What is the probability of a random person having an IQ of atleast 250?*

Solution: *Let us first calculate using Markov's inequality,*

$$Pr[X \geq 250] \leq \frac{100}{250} \leq 0.4$$

Using Chebyshev's inequality we get,

$$Pr[X - 100 \geq 150] \leq \frac{15^2}{150^2} \leq 0.01$$

We can clearly see the difference on the bounds we got from the two concentration inequalities.

Example 5. *Let us consider the coin-flipping example, and use Chebyshev's inequality to bound the probability of obtaining more than $3n/4$ heads in a sequence of n fair coin flips. Recall that $X_i = 1$ if the i^{th} coin flip is heads and 0 otherwise, and $X = \sum_{i=1}^{n} X_i$ denotes the number of heads in the n coin flips. To use Chebyshev's inequality we need to compute the variance of X. Observe that, since X_i is a bernoulli random variable,*

$$E[(X_i)^2] = E[X_i] = \frac{1}{2}$$

Thus,

$$Var[X_i] = E[(X_i)^2] - (E[X_i])^2 = \frac{1}{4}$$

Now, since $X = \sum_{i=1}^{n} X_i$ and the X_i are independent

$$Var[X] = Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i] = \frac{n}{4}$$

Applying Chebyshev's inequality yields

$$\begin{aligned} P[X \geq 3n/4] &\leq P[|X - E[X]| \geq n/4] \\ &\leq \frac{Var[X]}{(n/4)^2} \\ &= \frac{(n/4)}{(n/4)^2} \\ &= \frac{4}{n} \end{aligned}$$

In fact, we can do slightly better. Chebyshev's inequality yields that $4/n$ is actually a bound on the probability that X is either smaller than $n/4$ or larger than $3n/4$, so by symmetry the probability that X is greater than $3n/4$ is actually $2/n$. Chebyshev's inequality gives a significantly better bound than Markov's inequality for large n.

Usage: Chebyshev's inequality has great utility because it can be applied to any probability distribution in which the mean and variance are defined.

3.3 Chernoff Bound

The generic Chernoff bound for a random variable X states,

$$Pr(X \geq a) = Pr(e^{tX} \geq e^{ta}) \quad \forall t > 0$$

As $e^{tX} \geq 0$ and is monotonically increasing function, we can use Markov's inequality,

$$Pr(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}$$

When $X = X_1 + X_2.... + X_n$ for any $t > 0$,

$$Pr(X \geq a) \leq e^{-ta} \quad E\big[\prod_i e^{tX_i}\big]$$

For better tighter bounds we can optimize over 't'.

Derivation of Chernoff bound for Bernoulli random variable
Let $X_1, X_2, ..., X_n$ be independent rv(random variable), whose sum is X.
Let 'p' be the probability of $X_i = 1$.

$$\begin{aligned} E[e^{tX_i}] &= pe^t + (1-p) \\ &= 1 + p(e^t - 1) \\ &\leq e^{p(e^t-1)} \qquad (1 + x \leq e^x) \end{aligned} \tag{4}$$

$$\begin{aligned} Pr(X \geq a) &\leq \frac{E[e^{tX}]}{e^{at}} \\ &\leq e^{-at}E[e^{\sum_i tX_i}] \\ &\leq e^{-at}E[e^{tX_1}][e^{tX_2}]....E[e^{tX_n}] \\ &\quad \text{(As given independent rv's)} \\ &\leq e^{-at}e^{\sum_i p(e^t-1)} \quad \text{(From Eq. (4))} \end{aligned} \tag{5}$$

Now, substitute the following for $\delta > 0$ in Eq. (5),

$$\begin{aligned} a &= (1+\delta)np \\ &= (1+\delta)E[X] \\ t &= \ln(1+\delta) \end{aligned} \tag{6}$$

We will get,

$$\begin{aligned} Pr(X \geq (1+\delta)np) &\leq \frac{e^{np(1+\delta-1)}}{(1+\delta)^{(1+\delta)np}} \\ &\leq [\frac{e^{\delta}}{(1+\delta)^{1+\delta}}]^{np} \end{aligned} \tag{7}$$

Similarly, we can derive for different random variables.

Example 6. 1 *million people are playing pick* 4 *(*$0000 - 9999$*), i.e., there is a fixed* 4 *digit number and all people have to guess the number to be the winner. Calculate the probability of atleast* 200 *winner's.*

$$Pr[win] = \frac{1}{10000}$$
$$E[Number\ of\ winners] = 100.$$

$$\begin{aligned} Pr[X \geq 200] &= Pr[X \geq (1+\delta)100] \quad (where\ \delta = 1) \\ &\leq [\frac{e}{2^2}]^{100} \\ &\leq (0.67)^{100} = 4.05 * e^{-18} \end{aligned}$$

We got a very small probability, hence we have a tight bound.

Example 7. *Let X be the number of heads in a sequence of n independent fair coin flips.To compare the power of this bound to Chebyshev's bound. consider the probability of having no more than $n/4$ heads or no fewer than $3n/4$ heads in a sequence of n independent fair coin flips. In the previous theorem, we used Chebyshev's inequality to show that*

$$P\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq \frac{4}{n}$$

Using the Chernoff bound in this case, we find that

$$P\left(\left|X - \frac{n}{2}\right| \geq \frac{n}{4}\right) \leq 2exp\left\{-\frac{1}{3}\frac{n}{2}\frac{1}{4}\right\} = 2e^{-\frac{n}{24}}$$

Observe that Chernoff bound gives a bound that is exponentially smaller than the bound obtained using Chebyshev's inequality.

Applications:

- Chernoff bound is used to bound the tails of the distribution for a sum of independent random variables.
- The Chernoff bound is by far the most useful tool in randomized algorithms.
- Application in Networking : Chernoff bound is also used to obtain tight bounds for permutation routing problems which reduce network congestion while routing packets in sparse networks.

Summarizing the above three inequalities,

- Markov's Inequality : This inequality suffices when constant probability bound is sufficient for the task.
- Chebyshev's Inequality : This inequality is the appropriate one when one have a good handle on the variance of the random variable.
- Chernoff bound : This inequality gives sharp concentration bounds for random variables that are sums of independent and bounded random variables (most commonly, sums of independent indicator random variables).

3.4 Hoeffding's Lemma

Hoeffding's lemma is an inequality that bounds the moment-generating function of any bounded random variable.

Note that Markov's inequality bounded first moment of random variable and Chebyshev's bounded second moment of random variable.

Let X be any real valued random variable with $E[X] = \mu$, such that $a \leq X \leq b$ almost surely (that is with probability $= 1$). Then $\forall \lambda \in R$,

$$E[e^{\lambda X}] \leq e^{\lambda\mu} e^{\frac{(\lambda)^2(b-a)^2}{8}}$$

Proof. As exponential function in convex we will use convexity property, we can write X as convex combination of a and b.

$$\begin{aligned} X &= tb + (1-t)a \quad \text{where } t \in [0,1] \\ t &= \frac{X-a}{b-a} \end{aligned} \tag{8}$$

$$\begin{aligned} e^{\lambda X} &= e^{\lambda(tb+(1-t)a)} \\ &\leq te^{\lambda b} + (1-t)e^{\lambda a} \end{aligned} \tag{9}$$

Taking expectation and substituting Eq. (8) in Eq. (9),

$$\begin{aligned} E[e^{\lambda X}] &\leq e^{\lambda b} E\left[\frac{X-a}{b-a}\right] + e^{\lambda a} E\left[\frac{b-X}{b-a}\right] \\ &\leq e^{\lambda b}\left(\frac{\mu-a}{b-a}\right) + e^{\lambda a}\left(\frac{b-\mu}{b-a}\right) \\ &\quad \left(\text{Now substituting } \gamma = \frac{b-\mu}{b-a}\right) \\ &\leq e^{\lambda b}(1-\gamma) + e^{\lambda a}\gamma \end{aligned} \tag{10}$$

Let $u = (b-a)\lambda$. Consider the following function:

$$\begin{aligned} \phi(u) &= \log(\gamma e^{\lambda a} + (1-\gamma)e^{\lambda b}) \\ &= \lambda a + \log((1-\gamma)e^{u} + \gamma) \\ &= (\gamma-1)u + \lambda\mu + \log((1-\gamma)e^{u} + \gamma) \end{aligned} \tag{11}$$

As, $E[e^{\lambda X}] \leq e^{\phi(u)}$. To find the least upper bound, we need to minimize $\phi(u)$.

$\phi(u)$ is twice differentiable and hence using Taylor's theorem for any u there exists $\xi \in [0, u]$ such that,

$$\phi(u) = \phi(0) + u\phi^{'}(0) + \frac{u^2}{2}\phi^{''}(\xi)$$

Using Eq. (11), we can see that $\phi(0) = \lambda\mu$. Also,

$$\begin{aligned} \phi^{'}(0) &= (1-\gamma) + \frac{(1-\gamma)e^{u}}{\gamma + (1-(\gamma)e^{u}} \\ \phi^{''}(u) &= \frac{(1-\gamma)e^{u}}{\gamma + (1-(\gamma)e^{u}}[1 - \frac{(1-\gamma)e^{u}}{\gamma + (1-(\gamma)e^{u}}] \end{aligned} \tag{12}$$

Thus, $\phi^{'}(0) = 0$, $\phi^{''}(u) = p(1-p)$, where $p = \frac{(1-\gamma)e^{u}}{\gamma+(1-(\gamma)e^{u}}$. Thus, $\phi^{''}(u) \leq \frac{1}{4}$. Hence,

$$\phi(u) \leq \lambda\mu + \tfrac{1}{8}u^2 = \lambda\mu + \tfrac{\alpha^2}{8}(b-a)^2$$

Thus,

$$\begin{aligned} E[e^{\lambda X}] &\leq e^{\phi(u)} \\ &\leq e^{\lambda\mu} e^{\frac{\lambda^2(b-a)^2}{8}} \end{aligned} \tag{13}$$

■

3.5 Hoeffding's Inequality

Hoeffding's inequality provides an upper bound on the probability that the sum of independent random variables deviates from its expected value by more than a certain amount.

> Let $X_1, X_2, ..., X_n$ be n independent random variables, and $S_n = X_1 + X_2 + ... + X_n$, where $\forall i, X_i \in [a_i, b_i]$, then according to Hoeffding's inequality,
>
> $$Pr[S_n - E[S_n] \geq t] \leq e^{\frac{-2t^2n^2}{\sum_i (b_i - a_i)^2}}$$

Proof.

$$\begin{aligned} Pr[S_n - E[S_n] \geq t] &= Pr[e^{s(S_n - E[S_n])} \geq e^{st}] \quad (\text{For } \forall s > 0) \\ &\leq \frac{E[e^{s[S_n - E[S_n]]}]}{e^{st}} \quad (\text{Applying Markov's inequality}) \end{aligned} \tag{14}$$

$$\begin{aligned} E[e^{s[S_n - E[S_n]]}] &= E[e^{s\sum_i^n X_i - E[X_i]}] \\ &= E\left[\prod_i^n e^{s(X_i - E[X_i]}\right] \\ &\quad (\text{Substituting } Y_i = X_i - E[X_i]) \end{aligned}$$

$$
\begin{aligned}
E[e^{s[S_n - E[S_n]]}] &= E\left[\prod_i^n e^{sY_i}\right] \\
&\leq \prod_i^n [e^{sE[Y_i]} e^{\frac{s^2(b_i-a_i)^2}{8}}] \quad \text{(Applying Hoeffding's Lemma)} \\
&\leq \prod_i^n e^{\frac{s^2(b_i-a_i)^2}{8}} \quad (E[Y_i] = 0)
\end{aligned} \tag{15}
$$

By substituting Eq. (15) in Eq. (14), we get

$$
Pr[S_n - E[S_n] \geq t] \leq e^{-st + \frac{s^2 \sum_i^n (b_i-a_i)^2}{8}} \tag{16}
$$

To get the best possible upper bound, we find the minimum of the right hand side of the last inequality as a function of s. Define

$$
g(s) = -st + \frac{s^2 \sum_i^n (b_i - a_i)^2}{8}
$$

Note that g is a quadratic equation and achieves its minimum at

$$
s = \frac{4t}{\sum_i^n (b_i - a_i)^2}
$$

Thus we get

$$
Pr[S_n - E[S_n] \geq t] \leq e^{\frac{-2t^2}{\sum_i^n (b_i-a_i)^2}} \tag{17}
$$

■

Usage

One of the main application of Hoeffding's inequality is to analyse the number of required samples needed to obtain a confidence interval by solving the inequality,

$$
Pr[\bar{X} - E[\bar{X}] \geq t] \leq e^{-2nt^2}
$$

Symmetrically, the inequality is also valid for another side of the difference:

$$
Pr[-\bar{X} + E[\bar{X}] \geq t] \leq e^{-2nt^2}
$$

By adding them both up, we can obtain two-sided variant of this inequality:

$$
Pr[|\bar{X} - E[\bar{X}|] \geq t] \leq 2e^{-2nt^2}
$$

This probability can be interpreted as the level of significance α(probability of making an error) for a confidence interval around $E[\bar{X}]$ of size $2t$:

$$
\alpha = P(\bar{X} \notin [E[\bar{X}] - t, E[\bar{X}] + t]) \leq 2e^{-2nt^2}
$$

Solving for the number of required samples n gives us,

$$
n \geq \frac{log(2/\alpha)}{2t^2}
$$

Therefore, we require at least $\frac{log(2/\alpha)}{2t^2}$ samples to acquire $(1-\alpha)$ confidence interval $E[\bar{X}] \pm t$.

3.6 Azuma's Inequality

The Azuma–Hoeffding inequality gives a concentration result for the values of martingales that have bounded differences. That is here random variables are not independent.

Let $Z_0, ..., Z_n$ be a martingale sequence with respect to the filter $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq .... \subseteq \mathcal{F}_n$ such that for $Y_i = Z_i - Z_{i-1}$, we have that for all $i \in [n]$, $|Y_i| = |Z_i - Z_{i-1}| \leq c_i$. Then

$$
Pr[Z_N - Z_0 \geq t] \leq exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right) \text{ and } Pr[Z_0 - Z_n \geq t] \leq exp\left(\frac{-t^2}{2\sum_{i=1}^n c_i^2}\right)
$$

Proof. We first prove one side of inequality. For any $\lambda > 0$, using Chernoff bound and Markov's inequality

$$Pr[Z_n - Z_0 \geq t] = Pr[e^{\lambda(Z_n - Z_0)} \geq e^{\lambda t}] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda(Z_n - Z_0)}]$$

Now conditioning on $\mathcal{F}_{n-1}$, we get

$$\begin{aligned}\mathbb{E}[e^{\lambda(Z_n - Z_0)}] &= \mathbb{E}[e^{\lambda(Y_n + Z_{n-1} - Z_0)}] \\ &= \mathbb{E}[\mathbb{E}[e^{\lambda(Y_n + Z_{n-1} - Z_0)}|\mathcal{F}_{n-1}]] \\ &= \mathbb{E}[e^{\lambda(Z_{n-1} - Z_0)}\mathbb{E}[e^{\lambda Y_n}|\mathcal{F}_{n-1}]]\end{aligned}$$

Using the fact that Z_{n-1} and Z_0 are both measurable in the $\sigma-$algebra $\mathcal{F}_{n-1}$. We not bound the expectation $\mathbb{E}[e^{\lambda Y_n}|\mathcal{F}_{n-1}]$ using convexity of the function e^x. Let $\alpha \in [-1, 1]$ and $M \in \mathbb{R}$ be any real number. Then,

$$\alpha M = \left(\frac{1+\alpha}{2}\right)M - \left(\frac{1-\alpha}{2}\right)M$$

Now using the convexity of the function e^x,

$$e^{\alpha M} \leq \left(\frac{1+\alpha}{2}\right)e^{M} + \left(\frac{1-\alpha}{2}\right)e^{-M}$$

Now taking $\alpha = Y_n/c_n$ and $M = \lambda c_n$, we get

$$e^{\lambda Y_n} \leq \left(\frac{1+(Y_n/c_n)}{2}\right)e^{\lambda c_n} + \left(\frac{1-(Y_n/c_n)}{2}\right)e^{-\lambda c_n}$$

Using $\mathbb{E}[Y_n|\mathcal{F}_{n-1}] = 0$, we get

$$\begin{aligned}\mathbb{E}[e^{\lambda Y_n}|\mathcal{F}_{n-1}] &\leq \mathbb{E}\left[\left(\frac{1+(Y_n/c_n)}{2}\right)e^{\lambda c_n} + \left(\frac{1-(Y_n/c_n)}{2}\right)e^{-\lambda c_n}\middle|\mathcal{F}_{n-1}\right] \\ &= \frac{e^{\lambda c_n} + e^{-\lambda c_n}}{2} \leq e^{\frac{(\lambda c_n)^2}{2}}\end{aligned}$$

where the last step uses the fact $(e^x + e^{-x})/2 \leq e^{\frac{x^2}{2}}$ which uses taylor expansion to verify.

$$Pr[Z_n - Z_0 \geq t] \leq e^{-\lambda t}e^{\lambda^2 c_n^2/2}\mathbb{E}[e^{\lambda(Z_{n-1} - Z_0)}]$$

Continuing by same process, we can deduce

$$Pr[Z_n - Z_0 \geq t] \leq exp\bigg(-\lambda t + (\lambda^2/2)\sum_{i=1}^{n} c_i^2\bigg)$$

Since above equation holds for any $\lambda > 0$, we can optimize over λ to minimize the above bound. On calculating the above expression is minimized for $\lambda = \frac{t}{\sum_{i=1}^{n} c_i^2}$, which gives

$$Pr[Z_n - Z_0 \geq t] \leq exp\bigg(-\frac{t^2}{2\sum_{i=1}^{n} c_i^2}\bigg)$$

■

Similarly it can be proven for $Pr[Z_0 - Z_n \geq t]$.

Example 8. *Some times, we have to find the the interesting patterns, example examining DNA structure.*

Let $X = (X_1, ..., X_n)$ be independent characters chosen from alphabet A where $a = |A|$. Let $B = (b_1, ..., b_k)$ be fixed string of k characters from A. Let F be the number of occurrence of the fixed string B in the random string X.
Let,

$$Z_0 = E[F]$$

and for $1 \leq i \leq n$ let

$$Z_i = E[F|X_1, ..., X_i]$$

The sequence $Z_0, ..., Z_n$ is a Doob martingale, and

$$Z_n = F$$

Since each character in the string X can participate in no more than k possible matches, for any $0 \leq i \leq n$ we have

$$|Z_{i+1} - Z_i| \leq k$$

In other words. the value of X_{i+1} can affect the value of F by at most k in either direction, since X_{i+1} participates in no more than k possible matches. Hence the difference is

$$E[F|X_1, ..., X_{i+1}] - E[F|X_1, ..., X_i]| = |Z_{i+1} - Z_i|$$

must be at most k, Applying Azuma-Hoeffding Inequality yields

$$P[|F - E[F]| \geq \epsilon] \leq 2e^{\frac{-\epsilon^2}{2nk^2}}$$

4 Advanced Inequalities

In this section we now study advanced inequalities, namely: Bennett's Inequality, Bernstein's Inequality, Efron-Stein Inequality, McDiarmid's Inequality.

4.1 Bennett's Inequality

Let $X_1, ..., X_n$ be independent real-valued random variables with zero mean, and $|X_i| \leq 1$ with probability one. Then for any $t > 0$

$$\mathbb{P}\bigg[\sum_{i=1}^{n} X_i > t\bigg] \leq exp\bigg(-n\sigma^2 h\bigg(\frac{t}{n\sigma^2}\bigg)\bigg)$$

where,

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} Var\{X_i\}$$

$$h(u) = (1+u)\log(1+u) - u \; for \; u \geq 0$$

Proof. Given that mean of rv's are zero , that is

$$E[X_i] = 0 \tag{18}$$

Let

$$F_i = \sum_{r=2}^{\infty} \frac{s^{r-2} E(X_i^r)}{r! \sigma_i^2} \tag{19}$$

where $\sigma_i^2 = E(X_i^2) - E(X_i)^2 = Var\{X_i\}$

now , $e^x = 1 + x + \sum_{r=2}^{\infty} \frac{x^r}{r!}$ therefore,

$$\begin{aligned} E(e^{sX_i}) &= 1 + sE(X_i) + \sum_{r=2}^{\infty} \frac{s^r E(X_i^r)}{r!} \\ E(e^{sX_i}) &= 1 + s^2\sigma_i^2 F_i \qquad \text{(Using Eq. (18) and Eq. (19))} \\ &\leq e^{s^2\sigma_i^2 F_i} \end{aligned} \tag{20}$$

Consider the term $E(X_i^r)$. Since expectation of a function is just the Lebesgue integral of the function with respect to probability measure, we have
$E(X_i) = \int_P X_i^{r-1} X_i$. Using Cauchy Schwarz inequality we get,

$$\begin{aligned} E(X_i^r) &= \int_P X_i^{r-1} X_i \\ &\leq \bigg(\int_P \mid X_i^{r-1} \mid^2\bigg)^{1/2} \bigg(\int_P \mid X_i \mid^2)^{1/2}\bigg) \\ \Rightarrow E(X_i^r) &\leq \sigma_i \bigg(\int_P \mid X_i^{r-1} \mid^2\bigg)^{1/2} \end{aligned}$$

Proceeding to use the Cauchy Schwarz inequality recursively k times we get

$$E(X_i^r) \leq \sigma_i^{1+\frac{1}{2}+\frac{1}{2^2}+...+\frac{1}{2^{k-1}}} \left(\int_P \mid X_i^{(2^k r - 2^{k-1} - 1)} \mid \right)^{1/2^k}$$

$$= \sigma_i^{2(1-\frac{1}{2^k})} \left(\int_P \mid X_i^{(2^k r - 2^{k-1} - 1}) \right) \mid)^{1/2^k}$$

Now we know that $|X_i| \leq 1$. Therefore,

$$\left(\int_P \mid X_i^{(2^k r - 2^{k-1} - 1)} \mid \right)^{1/2^k} \leq 1$$

Hence, we get

$$E(X_i^r) \leq \sigma_i^{2(1-\frac{1}{2^k})}$$

Taking limit $k \to \infty$ we get

$$\begin{aligned} E(X_i^r) &\leq lim_{k\to\infty} \left\{ \sigma_i^{2(1-\frac{1}{2^k})} \right\} \\ &\Rightarrow E(X_i^r) \leq \sigma_i^2 \end{aligned} \tag{21}$$

Therefore, from Eq. (19) and Eq. (20) we get

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

Therefore,

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

Applying this to Eq. (20) we get ,

$$E(e^{sX_i}) \leq e^{s^2 \sigma_i^2 \frac{1}{s^2}(e^s - 1 - s)} \tag{22}$$

Using Chernoff Bound and Markov's inequality , we say that

$$P[X \geq t] \leq e^{-st} E[e^{sX}]$$

where $X = X_1 + X_2 + ... + X_n$ hence,

$$\begin{aligned} P\left[\sum_{i=1}^{n} X_i > t \right] &\leq e^{-st} E\left[\prod_{i=1}^{n} e^{sX_i} \right] \\ &= e^{-st} \prod_{i=1}^{n} E[e^{sX_i}] \qquad \text{(as given independent rv's)} \end{aligned}$$

Using Eq. (22) to this we get,

$$P\left[\sum_{i=1}^{n} X_i > t \right] \leq e^{-st} \prod_{i=1}^{n} e^{s^2 \sigma_i^2 \frac{1}{s^2}(e^s - 1 - s)}$$

As $\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} \sigma_i^2$, hence,

$$\begin{aligned} P\left[\sum_{i=1}^{n} X_i > t \right] &\leq e^{-st} e^{\sum_{i=1}^{n} \sigma_i^2 (e^s - 1 - s)} \\ &= e^{-st} e^{n\sigma^2 (e^s - 1 - s)} \end{aligned} \tag{23}$$

now to obtain the closest bound we minimize R.H.S w.r.t s, therefore we get

$$\frac{d e^{n\sigma^2(e^s - 1 - s) - st}}{ds} = e^{n\sigma^2(e^s - 1 - s) - st}(n\sigma^2(e^s - 1) - t) = 0$$

$$\Rightarrow e^s - 1 = \frac{t}{n\sigma^2}$$

We get,

$$s = \log\left(1 + \frac{t}{n\sigma^2}\right)$$

Using s in Eq. (23), we have

$$\begin{aligned} P\left[\sum_{i=1}^{n} X_i > t\right] &\leq e^{-log(1+\frac{t}{n\sigma^2})t + n\sigma^2(e^{log(1+\frac{t}{n\sigma^2})} - 1 - log(1+\frac{t}{n\sigma^2}))} \\ &= e^{-log(1+\frac{t}{n\sigma^2})t + n\sigma^2(\frac{t}{n\sigma^2} - log(1+\frac{t}{n\sigma^2}))} \\ &= e^{n\sigma^2(\frac{t}{n\sigma^2} - log(1+\frac{t}{n\sigma^2}) - \frac{t}{n\sigma^2}log(1+\frac{t}{n\sigma^2}))} \end{aligned}$$

Let $h(u) = (1+u)log(1+u) - u$ for $u > 0$, therefore we get

$$P\left[\sum_{i=1}^{n} X_i > t\right] \leq e^{-n\sigma^2 h\left(\frac{t}{n\sigma^2}\right)} \tag{24}$$

■

4.2 Bernstein's Inequality

Under the same conditions defined in the Bennett's inequality, for any $\epsilon > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i > \epsilon\right\} \leq exp\left(-\frac{n\epsilon^2}{2(\sigma^2 + \epsilon/3)}\right)$$

Proof. We can derive the Bernstein's inequality by further bounding the function h(x). Let the function be, $G(x) = \frac{3}{2}\frac{x^2}{x+3}$. Now consider a function $\phi(x) = h(x) - G(x)$. $\phi''(x) = \frac{x^3+9x^2}{(x+1)(x+3)^3}$, For all $x \geq 0$, $\phi \geq 0$ implies $\phi'(x)$ is increasing, i.e., for all $x \geq 0$, $\phi'(x) \geq 0$, and therefore $\phi(x)$ is increasing, hence $\phi(x) \geq 0$ for all $x \geq 0$, Hence we have

$$h(x) \geq G(x) \quad \forall x \geq 0$$

Therefore using Eq. (24) we get

$$\begin{aligned} P\left[\sum_{i=1}^{n} X_i > t\right] &\leq e^{-n\sigma^2 G(\frac{t}{n\sigma^2})} \\ \Rightarrow P\left[\sum_{i=1}^{n} X_i > t\right] &\leq e^{(\frac{-3t^2}{2(t+3n\sigma^2)})} \end{aligned}$$

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

$$P\left[\frac{1}{n}\sum_{i=1}^{n} X_i > \epsilon\right] \leq e^{-\frac{n\epsilon^2}{2\sigma^2 + 2\frac{\epsilon}{3}}} \tag{25}$$

■

Example 9. *We have $n = 2$ investments. Expected payoff of Investment 1 is $\mu_1 = \$50$ with standard deviation of $\sigma_1 = \$25$. Investment 2 has expected payoff $\mu_2 = \$70$ with standard deviation $\sigma_2 = \$20$. Investment 1 has a floor on its payoff of $L_1 = \$25$ and the upper bound of this payoff if $M_1 = \$65$. Meanwhile, Investment 2 has it's floor payoff of $L_2 = \$60$ and ceiling payoff be $M_2 = \$80$. For the portfolio to be worthwhile, we are told that the total payoff of both investments must be at least \$130. We apply Bennett's inequality, Bernstein's inequality and Hoeffding's inequality to this portfolio problem. If we calculate the probability bound using generic form of Bennett's inequality*

$$P\left\{\frac{1}{n}(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E[X_i]) \geq t\right\} \leq \exp\left(\frac{-nv}{s^2}h\left(\frac{ts}{v}\right)\right)$$

where
$h(x) = (1+x)\ln(1+x) - x$

$s = \max_i(M_i - \mu_i)$
$v = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$

The probability of complementary event specified in the inequality in turns out to be at least 0.9545 *for the values give*

According to the generic form of Bernstein's inequality,

$$P\left\{(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} E[X_i]) \geq t\right\} \leq \exp\left(\frac{-t^2}{2(n\sigma^2 + (t/3))}\right)$$

where
$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$
Bernstein's gives the probability to be at least 0.9525.

Applying Hoeffding's inequality to the same, we get

$$P\left\{\sum_{i=1}^{n}(X_i - E[X_i]) \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum_{i=1}^{n}(M_i - L_i)^2}\right)$$

where M_i and L_i are as specified in the example.
Hoeffding's gives the probability to be least 0.9048.

Clearly Hoeffding's inequality gives the tightest bound in most of the cases.

4.3 Efron-Stein Inequality

Let χ be some set and let $g : \chi^n \to \mathbb{R}$ be a measurable function of n variables, $Z = g(X_1, ..., X_n)$ and its expected value is $\mathbb{E}(Z)$ where $X_1, ..., X_n$ are arbitrary independent (not necessarily identically distributed!) random variables taking values in χ, Then

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

Where

$$E_i(Z) = E[Z \mid X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n]$$

Proof. Let $V = Z - E(Z)$. Now if we define V_i as

$$V_i = E[Z|X_1, ..., X_i] - E[Z|X_1, ..., X_{i-1}] \quad \forall i = 2, ..., n.$$

and for i=1,

$$V_1 = E[Z|X_1] - E[Z]$$

then

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

and

$$\begin{aligned} Var(Z) &= E(V^2) \\ &= E\left(\left(\sum_{i=1}^{n} V_i\right)^2\right) \\ &= E\left(\sum_{i=1}^{n} V_i^2\right) + 2E\left(\sum_{i>j} V_i V_j\right) \end{aligned} \tag{26}$$

now, $E[XY] = E[E[XY|Y]] = E[YE[X|Y]]$ Therefore

$$E[V_i V_j] = E[V_j E[V_i|X_1, ..., X_j]] \tag{27}$$

Now we calculate

$$\begin{aligned} E[V_i|X_1, ..., X_j] &= E[(E[Z|X_1, ...X_i] - E[Z|X_1, ..., X_{i-1}])|X_1, ..., X_j] \\ &= E[E[(Z|X_1, ...X_i)|X_1, ..., X_j] - E[(Z|X_1, ..., X_{i-1})|X_1, ..., X_j]] \end{aligned}$$

Since $i > j$ and $i - 1 \geq j$ Then by Towering property

$$E[V_i|X_1, ..., X_j] = E[E[Z|X_1, ..., X_j] - E[Z|X_1, ..., X_j]] = 0$$

Using this in Eq. (27) we get,

$$E[V_i V_j] = 0$$

Hence we have,

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

Bounding $E[V_i^2]$,

$$\begin{aligned} V_i^2 &= (E[Z|X_1, ..., X_i] - E[Z|X_1, ..., X_{i-1}])^2 \\ &= (E[E[Z|X_1, ..., X_n] - E[Z|X_1, ..., X_{i-1}, X_{i+1}, ..., X_n]|X_1, ..., X_i])^2 \\ &\leq E[(E[Z|X_1, ..., X_n] - E[Z|X_1, ..., X_{i-1}, X_{i+1}, ..., X_n])^2|X_1, ..., X_i] \\ &= E[(Z - E_i(Z))^2|X_1, ..., X_i] \end{aligned}$$

Summing over all i's and taking expectation on both sides. As we know quadratic function is convex and hence we can apply Jensens inequality.

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

■

Example 10. ***Kernel density estimation***
Let $X_1, ..., X_n$ be i.i.d. real samples drawn according to some density ϕ. The kernel density estimate is

$$\phi_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big)$$

where $h > 0$, and K is a nonnegative "kernel" $\int K = 1$. The L_1 error is

$$Z = f(X_1, ..., X_n) = \int |\phi(x) - \phi_n(x)| dx.$$

It is easy to see that

$$|f(X_1, ..., X_n) - f(X_1, ..., X_i', ..., X_n)| \leq \frac{1}{nh} \int \Big|K\Big(\frac{x - X_i}{h}\Big) - K\Big(\frac{x - X_i'}{h}\Big)\Big| \leq \frac{2}{n}$$

so we get

$$Var(Z) \leq \frac{2}{n}.$$

4.4 McDiarmid's Inequality

Let $X_1, ..., X_m$ be independent random variables all taking values in the set χ. Further, let $f : \chi^m \to \mathbb{R}$ be a function of $X_1, ..., X_m$ that satisfies $\forall i, \forall X_1, ..., X_m, X_i^{'} \in \chi$,

$$|f(X_1, ..., X_i, ..., X_m) - f(X_1, ..., X_i^{'}, ..., X_m)| \leq c_i$$

Then for all $\epsilon > 0$,

$$Pr[f - \mathbb{E}[f] \geq \epsilon] \leq exp\Big(\frac{-2\epsilon^2}{\sum_{i=1}^{m} c_i^2}\Big)$$

Proof. Let $f_i' = f(X_1, ..., X_i', ..., X_n)$
Using Chernoff bound we get,

$$P[f - E[f] \geq \epsilon] \leq e^{-\epsilon s} e^{sE[f - E[f]]}$$

Now let,

$$V_i = E[f|X_1, ..., X_i] - E[f|X_1, ..., X_{i-1}] \quad \forall i = 1, ..., n$$

then

$$V = \sum_{i=1}^{n} V_i = f - E[f]$$

Therefore,

$$\begin{aligned} P[f - E[f] \geq \epsilon] &\leq e^{-\epsilon s} E[e^{\sum_{i=1}^{n} sV_i}] \\ &= e^{-\epsilon s} \prod_{i=1}^{n} E[e^{sV_i}] \end{aligned} \tag{28}$$

Now let V_i be bounded by the interval $[L_i, U_i]$. We know that $|f - f_i'| \leq c_i$, hence it follows that $|V_i| \leq c_i$ and hence $|U_i - L_i| \leq c_i$. Using Hoeffding's lemma on $E[e^{sV_i}]$ we get,

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

Using this in Eq. (28) we get,

$$\begin{aligned} P[f - E[f] \geq \epsilon] &\leq e^{-\epsilon s} \prod_{i=1}^{n} e^{\frac{s^2 c_i^2}{8}} \\ &= e^{-s\epsilon + s^2 \sum_{i=1}^{n} \frac{c_i^2}{8}} \end{aligned}$$

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

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

$$\Rightarrow s = \frac{4\epsilon}{\sum_{i=1}^{n} c_i^2}$$

Hence the bound is given by,

$$\begin{aligned} P[f - E[f] \geq \epsilon] &\leq e^{-\frac{4\epsilon}{\sum_{i=1}^{n} c_i^2}\epsilon + (\frac{4\epsilon}{\sum_{i=1}^{n} c_i^2})^2 \sum_{i=1}^{n} \frac{c_i^2}{8}} \\ \Rightarrow P[f - E[f] \geq \epsilon] &\leq e^{-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i}} \end{aligned}$$

■

Example 11. *Let $X_1, ..., X_n \in A$ be n-tuple i.i.d. random variables whose common distribution is P, i.e. $X_1, ..., X_n \sim P$ and let $P_n(A)$ be the empirical distribution. The empirical distribution assigns the probability $1/n$ to each X_i*

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \in A)$$

Define $\triangle_n \equiv f(X_1, ..., X_n) = sup_A |P_n(A) P(A)|$. Changing one observation changes f by at most $\frac{1}{n}$. Hence,

$$P\bigg(|\triangle_n - E(\triangle_n)| > \epsilon\bigg) \leq 2e^{-2n\epsilon^2}.$$

Example 12. ***Kernel density function***
Similar to the example for Efron-Stein inequality, $X_1, X_2, ..., X_n$ be i.i.d. random variable and $\phi_n(x)$ be the kernel density estimate. If $Z = f(X_1, ...X_n) = \int |\phi(x) - \phi_n(x)| dx$ then, $|f(X_1, ..., X_n) - f(X_1, ..., X_i', ..., X_n)| \leq \frac{2}{n}$. Thus, we can observe that $f(X_n)$ has the bounded differences property with $c_1 = ... = c_n = 2/n$. Applying McDiarmid's inequality on $f(x)$ we get, $P(|f(X_n) - E[f(X_n)]| \geq \epsilon) \leq 2e^{-n\epsilon^2/2}$.

5 References

- Mitzenmacher, Michael; Upfal, Eli (2005). Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press
- Maxim Raginsky, Igal Sason, Concentration of Measure Inequalities in Information Theory, Communications, and Coding.
- Stephane Boucheron, Gabor Lugosi, Pascal Massart, Concentration inequalities A nonasymptotic theory of independence
- Anna Karlin, CSE525: Randomized Algorithms and Probabilistic Analysis, Lecture 18
- Maxim Raginsky, Concentration inequalities