Contents

1 Introduction

1.1 Basic Inequalities

Theorem 1.1.1 (Markov's inequality)**.** *For a random variable X with $\mathbf{P}\{X < 0\} = 0$ and $t > 0$, we have*

$$\mathbf{P}\{X \geq t\} \leq \frac{\mathbf{E}\, X}{t}.$$

It follows that for a non-decreasing function φ which only takes non-negative values,

$$\mathbf{P}\{X \geq t\} = \mathbf{P}\{\varphi(X) \geq \varphi(t)\} \leq \frac{\varphi(X)}{\varphi(t)}.$$

Proof. In the first place, note that

$$\begin{aligned} X &= X \cdot \mathbb{1}_{X \geq t} + X \cdot \mathbb{1}_{X < t} \\ &\geq t \cdot \mathbb{1}_{X \geq t} + 0, \end{aligned}$$

and thus,

$$\mathbf{E}\, X \geq t \cdot \mathbf{E}\, \mathbb{1}_{X \geq t} = t \cdot \mathbf{P}\{X \geq t\}.$$

For the second statement, apply the same argument on the random variable $Y := \varphi(X)$ and the constant $s := \varphi(t)$. □

Theorem 1.1.2 (Chebyshev's inequality)**.** *For $t > 0$ and a random variable X with mean $\mu = \mathbf{E}\, X$ and variance $\sigma^2 = \mathbf{Var}\, X$, then*

$$\mathbf{P}\{|X - \mu| \geq t\} \leq \sigma^2 t^{-2}.$$

Proof. Applying Markov's inequality with $\varphi : x \mapsto x^2$ we obtain,

$$\mathbf{P}\{|X - \mu| \geq t\} = \mathbf{P}\{|X - \mu|^2 \geq t^2\} \leq \frac{\mathbf{E}\left[(X - \mu)^2\right]}{t^2} = \sigma^2 t^{-2}.$$

□

Theorem 1.1.3 (Jensen's inequality)**.** *For any real valued random variable X and convex function φ*

$$\varphi(\mathbf{E}\, X) \leq \mathbf{E}\, \varphi(X)$$

1.2 Why bother?

The concentration inequalities are used to obtain information on how a random variable is distributed at some specific places of its domain. In the most common scenarios, these inequalities will be used to quantify how concentrated a random variable at its tails, for example,

$$\mathbf{P}\{|X-\mu| \geq t\} < f(t) << 1.$$

A concentration inequality is specially useful when this probability cannot be calculated at a low computational cost or estimated with high precision. The following will illustrate a case where using concentration inequalities achieves the best results.

1.2.1 Coin Tossing

A coin tossing game is fair if the chances of winning are equal to the chances of losing. We can verify from a sample of N games that the game is not rigged if the number of heads in the sample is not very distant from the average $N/2$. However, there's a chance that one may classify the coin as rigged, even when the coin is fair. By the *Law of Large Numbers*, we know that the larger the sample, the less likely it is to obtain a false positive. But let's ask ourselves how fast this probability converges to 0.

Let $S_N \sim \mathrm{Bi}(N, 1/2)$ denote the number of heads in a fair coin tossing game. Then,

$$\mu = \mathbf{E}\, S_N = \frac{N}{2}, \qquad \sigma^2 = \mathbf{Var}\, S_N = \frac{N}{4}.$$

For a fixed $\varepsilon > 0$, we may classify a coin tossing game as rigged if, after N trials, the ratio of heads vs tails in the sample is greater than $[1+\varepsilon : 1-\varepsilon]$, or similarly,

$$S_N \geq \mu + \frac{\varepsilon}{2}N = \frac{1+\varepsilon}{2}N.$$

Using the Chebyshev inequality 1.1.2, we assert that

$$\mathbf{P}\left\{S_N \geq \mu + \frac{\varepsilon}{2}N\right\} \leq \mathbf{P}\left\{|S_N - \mu| \geq \frac{\varepsilon}{2}N\right\} \leq \sigma^2 \frac{4}{\varepsilon^2 N^2} = \frac{1}{\varepsilon^2 N}.$$

Therefore, the probability of bad events tends to 0 at least linearly with the number of games.

1.2.2 Central Limit Theorem

The proof of the following theorems can be found in (ref)

Theorem 1.2.1. *Let X_i be a i.i.d. sample, each with expected value μ and variance σ^2. Let $S_N = \sum_{i=1}^N X_i$ and $Z_N = \frac{S_N - N\mu}{\sqrt{N}\sigma^2}$. Then,*

$$Z_N \to Z \sim \mathcal{N}(0,1), \ \textit{in distribution.}$$

Theorem 1.2.2 (Tails of the Normal Distribution). *Let $Z \sim \mathcal{N}(0,1)$, for $t > 0$ we have*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right) \leq \mathbf{P}\{Z \geq t\} \leq \frac{1}{t}\,\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-t^2}{2}\right)$$

With that in mind, we might naively assume that we can obtain better bounds by using the previous theorems since,

2 Exponential Inequalities

2.1 Chernoff-Okamoto Inequalities

Applying Markov's Inequality to $Y = e^{uX}$, we can assert that

$$\mathbf{P}\{X \geq \lambda + t\} \leq e^{-u(\lambda+t)}\mathbf{E}\, e^{uX} = e^{-u(\lambda+t)}(1 - p + pe^u)^n.$$

The right hand equation is minimized when,

$$e^u = \frac{\lambda + t}{(n - \lambda - t)} \cdot \frac{1-p}{p}.$$

Therefore, for $0 \leq t \leq n - \lambda$,

$$\mathbf{P}\{X \geq \lambda + t\} \leq \left(\frac{\lambda}{\lambda + t}\right)^{\lambda+t} \left(\frac{n-\lambda}{n-\lambda-t}\right)^{n-\lambda-t} \tag{2.1}$$

Theorem 2.1.1. *Let X be random variable with the binomial distribution* $\mathrm{Bi}(n, p)$ *with* $\lambda := np = \mathbf{E}\, X$*, then for* $t \geq 0$*,*

$$\mathbf{P}\{X \geq \lambda + t\} \leq \exp\left(-\frac{t^2}{2(\lambda + t/3)}\right) \tag{2.2}$$

$$\mathbf{P}\{X \leq \lambda - t\} \leq \exp\left(-\frac{t^2}{2\lambda}\right) \tag{2.3}$$

Used in: Theorem 3.0.2

Proof. (**TODO** I've already written the proof on paper) □

2.2 Hoeffding-Bernstein inequalities

Theorem 2.2.1. *Let* $\|f\|_\infty < c$*,* $\mathbf{E}\, f(X_1, \ldots, X_m) = 0$ *and* $\sigma^2 = \mathbf{E}\, f^2(X_1, \ldots, X_m)$*. Then for any* $t > 0$*,*

$$\mathbf{P}\{U_m^n(f, P) > t\} \leq \exp\left(\frac{\frac{n}{m}t^2}{2\sigma^2 + \frac{2}{3}ct}\right) \tag{2.1}$$

Used in: Theorem 3.0.4

Proof. Proposition 2.3(a) M.A. Arcones, E. Gine, Limit theorems for U-processes, Ann. Probab. 21 (1993) 14941542

https://sci-hub.se/https://www.jstor.org/stable/2244585 □

3 Application to Estimation of Data Dimension

The article (ref) explains how we can estimate the dimension d of a manifold M embeded on a Euclidean space of dimension m, say $\mathbb{R}^m$. First, we are going to introduce the method they used, and then, we will show how does the exponential inequalities can be used to prove two important results in the paper. The procedure starts with an example on a uniformly distributed sample on a d-sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, but will be later generalized for samples of any distribution on any manifold.

In the first place, let $Z_1, \ldots, Z_k$ be a i.i.d. sample uniformly distributed on $\mathbb{S}^{d-1}$. Then, we have the following formula for the variance of the angles between $Z_i, Z_j, i \neq j$:

$$\beta_d := \mathbf{Var}\left(\arccos \langle Z_i, Z_j \rangle\right) = \begin{cases} \dfrac{\pi^2}{4} - 2\displaystyle\sum_{j=1}^{k} (2j-1)^{-2}, & \text{if } d = 2k+1 \text{ is odd,} \\ \dfrac{\pi^2}{12} - 2\displaystyle\sum_{j=1}^{k} (2j)^{-2}, & \text{if } d = 2k+2 \text{ is even.} \end{cases} \tag{3.1}$$

The previous formula for the angle variance is proven in (ref) and will be skipped (**TODO**: Should it really be skipped?). In order to give more insight on how we will be choosing the estimator $\widehat{d}$ for the dimension of the sphere, consider the following theorem.

Theorem 3.0.1 (Bounds for β_d)**.** *For every $d > 1$,*

$$\frac{1}{d} \leq \beta_d \ \leq \ \frac{1}{d-1}.$$

Proof. The even and the odd cases must be distinguished:

(1): When $d = 2k+2$ is even: In the first place, remember that,

$$\lim_{k \to \infty} \sum_{j=1}^{k} j^{-2} = \frac{\pi^2}{6}.$$

It follows that

$$\begin{aligned} \beta_d &= \frac{\pi^2}{12} - 2\sum_{j=1}^{k} (2j)^{-2} = \frac{\pi^2}{12} - \frac{1}{2}\sum_{j=1}^{k} j^{-2} \\ &= \frac{1}{2}\sum_{j=k+1}^{\infty} j^{-2}. \end{aligned}$$

Since $(j^{-2})_{j\in\mathbb{N}}$ is a monotonically decreasing sequence, it follows that (**TODO**: Improve array syntax)

$$\begin{array}{ccccc} \frac{1}{d} & = & \frac{1}{2k+2} & = & \frac{1}{2}\int_{k+1}^{\infty} x^{-2}dx \\ & \leq & \beta_d & \leq & \frac{1}{2}\int_{k+1/2}^{\infty} x^{-2}dx \\ & = & \frac{1}{2k+1} & = & \frac{1}{d-1}. \end{array}$$

(2): When $d = 2k+3$ is odd: On the other hand, note that

$$\begin{array}{rcl} \lim_{k\to\infty}\sum_{j=1}^{k}(2j-1)^{-2} & = & \lim_{k\to\infty}\sum_{j=1}^{2k-1} j^{-2} - \sum_{j=1}^{k-1}(2j)^{-2} \\ & = & \lim_{k\to\infty}\sum_{j=1}^{2k-1} j^{-2} - \frac{1}{4}\sum_{j=1}^{k-1} j^{-2} \\ & = & \frac{\pi^2}{6} - \frac{\pi^2}{24} = \frac{\pi^2}{8} \end{array}$$

Then,

$$\begin{array}{rcl} \beta_d & = & \frac{\pi^2}{4} - 2\sum_{j=1}^{k}(2j-1)^{-2} \\ & = & 2\sum_{j=k+1}^{\infty}(2j-1)^{-2}. \end{array}$$

Using a similar argument we conclude that (**TODO**: Improve array syntax)

$$\begin{array}{ccccc} \frac{1}{d} & = & \frac{1}{2k+1} & = & 2\int_{k+1}^{\infty}(2x-1)^{-2}dx \\ & \leq & \beta_d & \leq & 2\int_{k+1/2}^{\infty}(2x-1)^{-2}dx \\ & = & \frac{1}{2k+2} & = & \frac{1}{d-1}. \end{array}$$

□

Knowing that for every $d > 1$, β_d is in the interval $[\frac{1}{d}, \frac{1}{d-1}]$, we are going to guess the dimension of the sphere by estimating β_d, and then, taking d from the lower bound of the interval where our estimator is. Since β_d is the variance of the angles in our sphere, our best choice for an estimator is the angle's sample variance,

$$U_k = \binom{k}{2}^{-1} \sum_{i<j\leq k} \left(\arccos\langle Z_i, Z_j\rangle - \frac{\pi^2}{2}\right)^2. \tag{3.2}$$

In Proposition 1. of (ref) the authors prove that it's the Minimum Variance Unbiased Estimator for β_d on the unit sphere. However, the authors also prove that this result can be generalized for any manifold with samples of any distribution.

Let $X_1, \dots, X_n$ be a i.i.d. sample from a random distribution P on a manifold $M \subset \mathbb{R}^m$, and let $p \in M$ a point. For $C > 0 \in \mathbb{R}$, let $k = \lceil C \ln(n) \rceil$ and define $R(n) = L_{k+1}(p)$ as the distance between p and its $(k+1)$-nearest neighbor. W.L.O.G. assume that $p = 0$ and that $X_1, \dots, X_k$ are the k-nearest neighbors of p

Theorem 3.0.2 (Bound k-neighbors). *For any sufficiently large $C > 0$, we have that, there exists n_0 such that, with probability 1, for every $n \geq n_0$,*

$$R(n) \leq f_{p,P,C}(n) = O(\sqrt[d]{\ln(n)/n}). \tag{3.3}$$

The function $f_{p,P,C}$ is a deterministic function which depends on p, P and C.

(**TODO**: I need help connecting the previous theorem with the following idea)

Let $\pi : R^m \to T_p M$ be the orthogonal projection on the Tangent Space of M at p. Also, define $W_i := \pi(X_i)$ and then normalize,

$$Z_i := \frac{X_i}{\|X_i\|}, \qquad \widehat{W}_i := \frac{W_i}{\|W_i\|}.$$

What follows from the previous and the following lemma is that if we know that $W_1, \dots, W_k$ behave similar to a uniformly distributed sample on the sphere $\mathbb{S}^d$, then, $Z_1, \dots, Z_k$ (the normalized k-nearest neighbors of p) also behave like they are uniformly distributed on $\mathbb{S}^d$.

Theorem 3.0.3 (Projection Distance Bounds). *For any $i < j \leq n$,*

(i) $$\|X_i - \pi(X_i)\| = O(\|\pi(X_i)\|^2) \tag{3.4}$$

(ii) $$\|Z_i - \widehat{W}_i\| = O(\|\pi(X_i)\|) \tag{3.5}$$

(iii) The inner products (cosine of angles) can be bounded as it follows:

$$|\langle Z_i, Z_j \rangle - \langle \widehat{W}_i, \widehat{W}_j \rangle| \leq Kr, \tag{3.6}$$

for a constant $K \in \mathbb{R}$, whenever $r \geq \max(\|\pi(X_i)\|, \|\pi(X_j)\|)$.

The last result in the article shows that if we estimate β_d as we did with $U_{k,n} = U_k$ in equation (3.2), and then, extract $\widehat{d}$ from the interval where $U_{k,n}$ is located, it follows that,

Theorem 3.0.4 (Consistency). *When $n \to \infty$,*

$$\mathbf{P}\{\widehat{d} \neq d\} \to 0.$$

Proofs

Theorem 3.0.2. □