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. Let $S_N = \sum_{i=1}^{N} X_i$, with mean $\mu = \mathbf{E}\, S_N$ and variance $\sigma^2 = \mathbf{Var}\, S_N$. If*

$$Z_N = \frac{S_N - N \cdot \mathbf{E}\, X_i}{\sqrt{N \cdot \mathbf{Var}\, X_i}} = \frac{S_N - \mu}{\sqrt{N}\sigma},$$

then,

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

□

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 better bounds can be obtained by using the previous theorem. For a large enough N we can say that for the coin tossing,

$$Z_N = \frac{S_N - N/2}{\sqrt{N/4}}$$

$$\implies \mathbf{P}\left\{S_N \geq \frac{1+\varepsilon}{2}N\right\} = \mathbf{P}\left\{Z_N \geq \varepsilon\sqrt{N}\right\} \sim \mathbf{P}\left\{Z \geq \varepsilon\sqrt{N}\right\}.$$

However, this raises the question of whether we can draw the following conclusion from Theorem 1.2.2:

$$\mathbf{P}\left\{S_N \geq \frac{1+\varepsilon}{2}N\right\} \leq \frac{1}{\varepsilon\sqrt{N}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\varepsilon^2 \cdot N}{2}\right).$$

Unfortunately, the answer is no. The following theorem will show why.

Theorem 1.2.3 (Convergence Rate for Central Limit Theorem)**.** *For Z_N, Z in Theorem 1.2.1, we have:*

$$|\mathbf{P}\{Z_N \geq t\} - \mathbf{P}\{Z \geq t\}| \in O(\tfrac{1}{\sqrt{N}}).$$

□

Since the approximation error is greater than the bound, the previous results cannot be taken into account.

In the context of coin tossing, this may not matter at all because the linear bound obtained using Chebyshev's inequality indicates that the probability of wrongly classifying a fair coin as a rigged coin converges at least linearly to zero. Even the Central Limit Theorem shows in a less precise way this convergence. However, for some specific problems in statistics, these basic tools are not precise enough to solve them. In the following chapters, we will show some examples were better crafted strategies are needed in order to get bounds to the tails of the random variables.

2 Exponential Inequalities

Even if we are satisfied with the linear convergence rate provided by Chebyshev's inequality, there is a simple way to improve this bound with the following result.

Theorem 2.0.1 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_N$ be independent random variables, where $X_i \sim \mathrm{Be}(p_i)$. Then, for every $t > 0$, we have*

$$P\left\{\sum (X_i - \mathbf{E}\, X_i) \geq t\right\} \leq \exp\left(\frac{-2t^2}{N}\right)$$

Proof. TODO □

Another exponential bound we can derive with a similar idea is the following

Theorem 2.0.2 (Chernoff's inequality)**.** *Let $X_1, \ldots, X_N$ be independent random variables, where for every $i = 1, \ldots, N$, $X_i \in [a_i, b_i]$. Define $S_N = \sum X_i$ and let $\mu = \mathbf{E}\, S_N$. Then, for every $t > 0$, we have*

$$P\left\{S_N \geq \mu + t\right\} \leq \exp\left(\frac{-2t^2}{\sum (a_i - b_i)^2}\right)$$

2.1 Chevyshev vs Chernoff vs Hoeffding

2.2 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 2.2.1. *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) □

2.3 Hoeffding-Bernstein inequalities

Theorem 2.3.1. *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 $Z_1, \dots, Z_k$ 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 $d > 1$,*

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

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 $X_1, \ldots, X_n$ 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 $X_1, \ldots, X_k$ 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 $\pi : R^m \to T_pM$ be the orthogonal projection on the Tangent Space of M at p. Also, define $W_i := \pi(X_i)$ and then normalize,

$$Z_i := \frac{X_i}{\|X_i\|}, \quad \widehat{W_i} := \frac{W_i}{\|W_i\|}.$$

What follows from the previous and the following lemma is that if we know that $W_1, \ldots, W_k$ behave similar to a uniformly distributed sample on the sphere $\mathbb{S}^d$, then, $Z_1, \ldots, Z_k$ (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) $\|X_i - \pi(X_i)\| = O(\|\pi(X_i)\|^2)$ (3.4)

(ii) $\|Z_i - \widehat{W_i}\| = O(\|\pi(X_i)\|)$ (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 $n \to \infty$,*

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

3.1 Proofs

Proof Theorem 3.0.1: 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

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

Then,

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

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

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

□

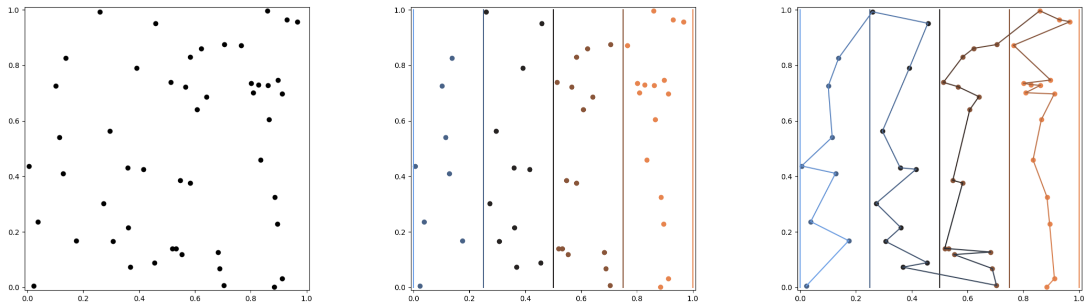
Proof Theorem 3.0.2: □

4 Application to a Heuristic Algorithm for Travelling Salesman

In this section we are going to present an application of the Azuma-Hoeffding inequality to prove the convergence to the mean of a linear approximation algorithm for the *Travelling Salesman Problem.*

The Algorithm

Let $X_1, \dots, X_N$ be a sample of N uniformly distributed points in a compact square $[0, L] \times [0, L]$. The algorithm divides this square in M stripes of width L/M each. Then, it connects each of the points in each of the stripes vertically and connects the top-most of one stripe with the top-most of the next one (or viceversa as the image below shows).



(a) Uniform sample (b) Divide in M stripes (c) Join points vertically

In the reference (ref) the authors assert that by choosing a number of stripes $M^* = \lfloor 0.58 N^{1/2} \rfloor$, one can achieve the best result in comparison to the real TSP solution. If t_N is the TSP solution distance for our sample and d_N is the algorithm's answer with the optimal M^*, then the error is asymptotically:

$$\frac{d_N - t_N}{t_N} \approx 0.23.$$

The result that we are going to prove is that d_N converges with an exponential rate to its mean. To prove our point, we are going to modify the algorithm's trajectory as it follows. Let e_N be trajectory distance that for any empty stripe in the plane we sum the length of its diagonal $\sqrt{L^2 + L^2/M^2}$ and then it skips the empty stripe. When there are no empty stripes $e_N = d_N$ and the probability that any given stripe is empty converges exponentially to 0:

$$\begin{aligned}(1-1/M)^N &= (1-0.58^{-1}N^{-1/2})^N \\ &= \left((1-1/M)^M\right)^{0.58^{-1}N^{1/2}} \\ &\sim \exp(-0.58^{-1}N^{1/2}).\end{aligned}$$

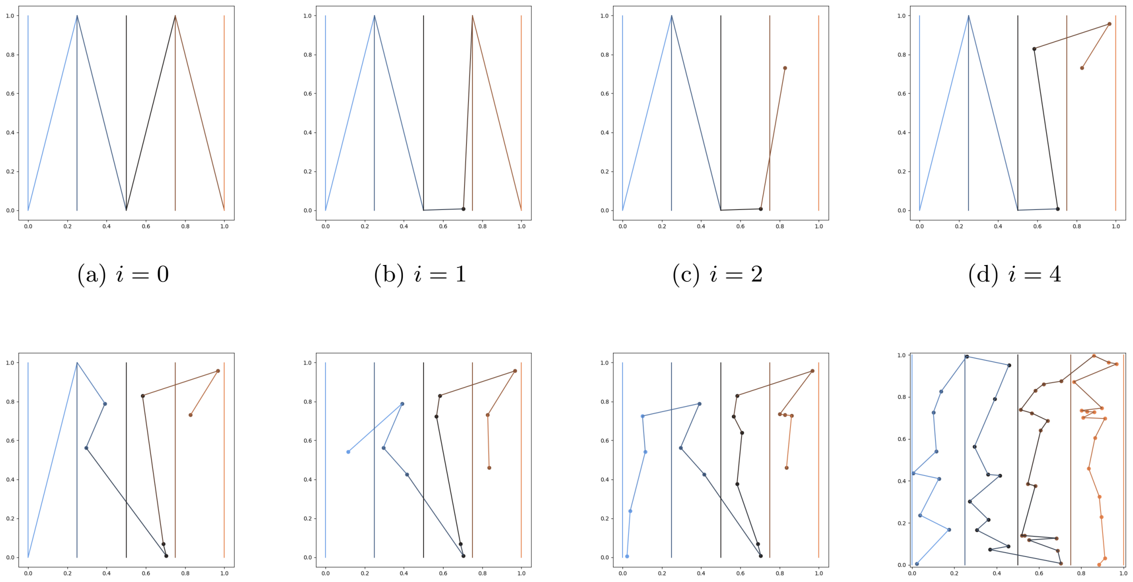
Let $\mathcal{A}_i := \sigma\{X_1, \dots, X_i\}$ be the sigma algebra corresponding to revealing the first i points, $\mathcal{A}_0 = \{\emptyset, [0,L]^2\}$. The expected value of the trajectory e_N given that we only know the positions of the first i points in the sample is $\mathbf{E}\left(e_N|\mathcal{A}_i\right)$. Define

$$Z_i = \mathbf{E}\left(e_N|\mathcal{A}_i\right) - \mathbf{E}\left(e_N|\mathcal{A}_{i-1}\right),$$

As the difference of this expectations when we reveal 1 more point. Note that since

$$\mathbf{E}\left(Z_i|\mathcal{A}_i\right) = \mathbf{E}\left(e_N|\mathcal{A}_i, \mathcal{A}_i\right) - \mathbf{E}\left(e_N|\mathcal{A}_{i-1}, A_i\right) = \mathbf{E}\left(e_N|\mathcal{A}_i\right) - \mathbf{E}\left(e_N|\mathcal{A}_i\right) = 0,$$

The $Z_i's$ form a vertex exposure martingale sequence.



(a) $i = 0$ (b) $i = 1$ (c) $i = 2$ (d) $i = 4$

(e) $i = 7$ (f) $i = 12$ (g) $i = 18$ (h) $i = N = 50$

Figure 4.1: Evolution of the Martingale

Define $e_N^{[i]}$ as the distance of the trajectory when we remove the i-th point from the sample. Intuitively from the figure above and the triangle inequality, we can obtain

$$e_N^{[i]} \le e_N \le e_N + 2L/M,$$

meaning that revealing one point cannot increase more than 2 widths the distance of the trajectory. Thus,

$$\|Z_i\|_\infty = \sup_{X_1,\ldots,X_N} \|\mathbf{E}\left(e_N|\mathcal{A}_i\right) - \mathbf{E}\left(e_N|\mathcal{A}_{i-1}\right)\| \leq 2L/M.$$

On the other hand, by telescopic sums we obtain that

$$e_N - Ee_N = \mathbf{E}\left(e_N|\mathcal{A}_N\right) - \mathbf{E}\left(e_N|\mathcal{A}_0\right) = \sum_{i=1}^{N} Z_i.$$

Therefore, by the Azuma-Hoeffding inequality,

$$\mathbf{P}\{|e_N - Ee_N| > t\} \leq 2\exp\left(\frac{-t^2}{2}\sum_{i=1}^{N}\|Z_i\|_\infty^2\right).$$

Finally,

$$\sum_{i=1}^{N}\|Z_i\|_\infty^2 \leq \frac{4NL^2}{M^2},$$

which implies that

$$\mathbf{P}\{|e_N - Ee_N| > t\} \leq 2\exp\left(\frac{-t^2}{2}\sum_{i=1}^{N}\frac{4NL^2}{M^2}\right) \sim e^{-t^2KN},$$

for some $K \in \mathbb{R}^+$.