Concentration inequalities and martingale inequalities — a survey

Fan Chung [*†] Linyuan Lu [‡†]

May 28, 2006

Abstract

We examine a number of generalized and extended versions of concentration inequalities and martingale inequalities. These inequalities are effective for analyzing processes with quite general conditions as illustrated in an example for an infinite Polya process and webgraphs.

1 Introduction

One of the main tools in probabilistic analysis is the concentration inequalities. Basically, the concentration inequalities are meant to give a sharp prediction of the actual value of a random variable by bounding the error term (from the expected value) with an associated probability. The classical concentration inequalities such as those for the binomial distribution have the best possible error estimates with exponentially small probabilistic bounds. Such concentration inequalities usually require certain independence assumptions (i.e., the random variable can be decomposed as a sum of independent random variables).

When the independence assumptions do not hold, it is still desirable to have similar, albeit slightly weaker, inequalities at our disposal. One approach is the martingale method. If the random variable and the associated probability space can be organized into a chain of events with modified probability spaces and if the incremental changes of the value of the event is "small", then the martingale inequalities provide very good error estimates. The reader is referred to numerous textbooks [5, 17, 20] on this subject.

In the past few years, there has been a great deal of research in analyzing general random graph models for realistic massive graphs which have uneven degree distribution such as the power law [1, 2, 3, 4, 6]. The usual concentration inequalities and martingale inequalities have often been found to be inadequate and in many cases not feasible. The reasons are multi-fold: Due to uneven degree distribution, the error bound of those very large degrees offset the delicate

[*]University of California, San Diego, fan@ucsd.edu

[†]Research supported in part by NSF Grants DMS 0100472 and ITR 0205061

[‡]University of South Carolina, lu@math.sc.edu

analysis in the sparse part of the graph. For the setup of the martingales, a uniform upper bound for the incremental changes are often too poor to be of any use. Furthermore, the graph is dynamically evolving and therefore the probability space is changing at each tick of the time.

In spite of these difficulties, it is highly desirable to extend the classical concentration inequalities and martingale inequalities so that rigorous analysis for random graphs with general degree distributions can be carried out. Indeed, in the course of studying general random graphs, a number of variations and generalizations of concentration inequalities and martingale inequalities have been scattered around. It is the goal of this survey to put together these extensions and generalizations to present a more complete picture. We will examine and compare these inequalities and complete proofs will be given. Needless to say that this survey is far from complete since all the work is quite recent and the selection is heavily influenced by our personal learning experience on this topic. Indeed, many of these inequalities have been included in our previous papers [9, 10, 11, 12].

In addition to numerous variations of the inequalities, we also include an example of an application on a generalization of Polya's urn problem. Due to the fundamental nature of these concentration inequalities and martingale inequalities, they may be useful for many other problems as well.

This paper is organized as follows:

1. Introduction — overview, recent developments and summary.

2. Binomial distribution and its asymptotic behavior — the normalized binomial distribution and Poisson distribution,

3. General Chernoff inequalities — sums of independent random variables in five different concentration inequalities.

4. More concentration inequalities — five more variations of the concentration inequalities.

5. Martingales and Azuma's inequality — basics for martingales and proofs for Azuma's inequality.

6. General martingale inequalities — four general versions of martingale inequalities with proofs.

7. Supermartingales and submartingales — modifying the definitions for martingale and still preserving the effectiveness of the martingale inequalities.

8. The decision tree and relaxed concentration inequalities — instead of the worst case incremental bound (the Lipschitz condition), only certain 'local' conditions are required.

9. A generalized Polya's urn problem — An application for an infinite Polya process by using these general concentration and martingale inequalities.

For webgraphs generated by the preferential attachment scheme, the concentration for the power law degree distribution can be derived in a similar way.

2 The binomial distribution and its asymptotic behavior

Bernoulli trials, named after James Bernoulli, can be thought of as a sequence of coin flips. For some fixed value p, where $0 \leq p \leq 1$, the outcome of the coin tossing process has probability p of getting a "head". Let S_n denote the number of heads after n tosses. We can write S_n as a sum of independent random variables X_i as follows:

$$S_n = X_1 + X_2 + \cdots + X_n$$

where, for each i, the random variable X satisfies

$$\begin{aligned} \Pr(X_i = 1) &= p, \\ \Pr(X_i = 0) &= 1 - p. \end{aligned} \tag{1}$$

A classical question is to determine the distribution of S_n. It is not too difficult to see that S_n has the *binomial distribution* $B(n, p)$:

$$\Pr(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad \text{for } k = 0, 1, 2, \ldots, n.$$

The expectation and variance of $B(n, p)$ are

$$\mathrm{E}(S_n) = np, \qquad \mathrm{Var}(S_n) = np(1-p).$$

To better understand the asymptotic behavior of the binomial distribution, we compare it with the normal distribution $N(\alpha, \sigma)$, whose density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

where α denotes the expectation and σ^2 is the variance.

The case $N(0, 1)$ is called the *standard normal distribution* whose density function is given by

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \qquad -\infty < x < \infty.$$

When p is a constant, the limit of the binomial distribution, after scaling, is the standard normal distribution and can be viewed as a special case of the Central-Limit Theorem, sometimes called the DeMoivre-Laplace limit Theorem [15].

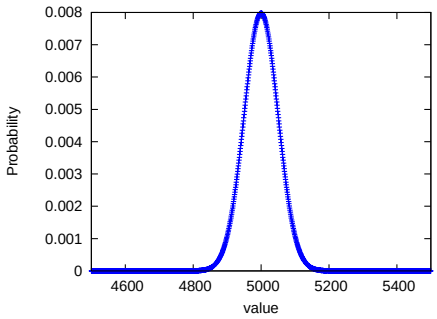

Figure 1: *The Binomial distribution* $B(10000, 0.5)$

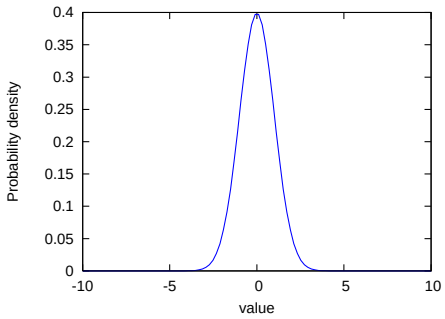

Figure 2: *The Standard normal distribution* $N(0, 1)$

Theorem 1 *The binomial distribution* $B(n, p)$ *for* S_n*, as defined in (1), satisfies, for two constants* a *and* b*,*

$$\lim_{n\to\infty} \Pr(a\sigma < S_n - np < b\sigma) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

where $\sigma = \sqrt{np(1-p)}$ *provided,* $np(1-p) \to \infty$ *as* $n \to \infty$*.*

When np is upper bounded (by a constant), the above theorem is no longer true. For example, for $p = \frac{\lambda}{n}$, the limit distribution of $B(n, p)$ is the so-called *Poisson distribution* $P(\lambda)$:

$$\Pr(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}, \qquad \text{for } k = 0, 1, 2, \cdots.$$

The expectation and variance of the Poisson distribution $P(\lambda)$ is given by

$$\mathrm{E}(X) = \lambda, \qquad \text{and} \qquad \mathrm{Var}(X) = \lambda.$$

Theorem 2 *For* $p = \frac{\lambda}{n}$*, where* λ *is a constant, the limit distribution of binomial distribution* $B(n, p)$ *is the Poisson distribution* $P(\lambda)$*.*

Proof: We consider

$$\begin{aligned}
\lim_{n\to\infty} \Pr(S_n = k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{\lambda^k \prod_{i=0}^{k-1} (1 - \frac{i}{n})}{k!} e^{-p(n-k)} \\
&= \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}$$

□

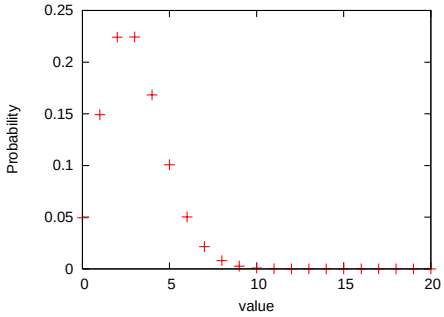

Figure 3: *The Binomial distribution* $B(1000, 0.003)$

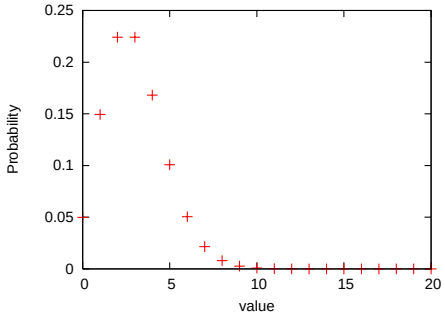

Figure 4: *The Poisson distribution* $P(3)$

As p decreases from $\Theta(1)$ to $\Theta(\frac{1}{n})$, the asymptotic behavior of the binomial distribution $B(n, p)$ changes from the normal distribution to the Poisson distribution. (Some examples are illustrated in Figures 5 and 6). Theorem 1 states that the asymptotic behavior of $B(n, p)$ within the interval $(np - C\sigma, np + C\sigma)$ (for any constant C) is close to the normal distribution. In some applications, we might need asymptotic estimates beyond this interval.

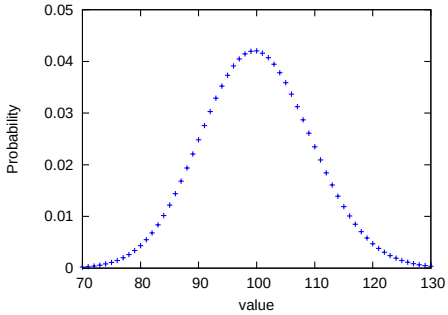

Figure 5: *The Binomial distribution* $B(1000, 0.1)$

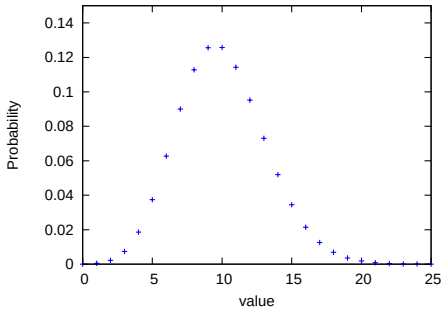

Figure 6: *The Binomial distribution* $B(1000, 0.01)$

3 General Chernoff inequalities

If the random variable under consideration can be expressed as a sum of independent variables, it is possible to derive good estimates. The binomial distribution is one such example where $S_n = \sum_{i=1}^{n} X_i$ and X_i's are independent and identical. In this section, we consider sums of independent variables that are not necessarily identical. To control the probability of how close a sum of random variables is to the expected value, various concentration inequalities are in play.

A typical version of the Chernoff inequalities, attributed to Herman Chernoff, can be stated as follows:

Theorem 3 *[8] Let $X_1, \ldots, X_n$ be independent random variables with $\mathrm{E}(X_i) = 0$ and $|X_i| \leq 1$ for all i. Let $X = \sum_{i=1}^n X_i$ and let σ^2 be the variance of X_i. Then*

$$\Pr(|X| \geq k\sigma) \leq 2e^{-k^2/4n},$$

for any $0 \leq k \leq 2\sigma$.

If the random variables X_i under consideration assume non-negative values, the following version of Chernoff inequalities is often useful.

Theorem 4 *[8] Let $X_1, \ldots, X_n$ be independent random variables with*

$$\Pr(X_i = 1) = p_i, \qquad \Pr(X_i = 0) = 1 - p_i.$$

We consider the sum $X = \sum_{i=1}^n X_i$, with expectation $\mathrm{E}(X) = \sum_{i=1}^n p_i$. Then we have

$$\begin{array}{lll} \textit{(Lower tail)} & \Pr(X \leq \mathrm{E}(X) - \lambda) & \leq \quad e^{-\lambda^2/2\mathrm{E}(X)}, \\ \textit{(Upper tail)} & \Pr(X \geq \mathrm{E}(X) + \lambda) & \leq \quad e^{-\frac{\lambda^2}{2(\mathrm{E}(X)+\lambda/3)}}. \end{array}$$

We remark that the term $\lambda/3$ appearing in the exponent of the bound for the upper tail is significant. This covers the case when the limit distribution is Poisson as well as normal.

There are many variations of the Chernoff inequalities. Due to the fundamental nature of these inequalities, we will state several versions and then prove the strongest version from which all the other inequalities can be deduced. (See Figure 7 for the flowchart of these theorems.) In this section, we will prove Theorem 8 and deduce Theorems 6 and 5. Theorems 10 and 11 will be stated and proved in the next section. Theorems 9, 7, 13, 14 on the lower tail can be deduced by reflecting X to $-X$.

The following inequality is a generalization of the Chernoff inequalities for the binomial distribution:

Theorem 5 *[9] Let $X_1, \ldots, X_n$ be independent random variables with*

$$\Pr(X_i = 1) = p_i, \qquad \Pr(X_i = 0) = 1 - p_i.$$

For $X = \sum_{i=1}^n a_i X_i$ with $a_i > 0$, we have $\mathrm{E}(X) = \sum_{i=1}^n a_i p_i$ and we define $\nu = \sum_{i=1}^n a_i^2 p_i$. Then we have

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \quad \leq \quad e^{-\lambda^2/2\nu} \tag{2}$$

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \quad \leq \quad e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}} \tag{3}$$

where $a = \max\{a_1, a_2, \ldots, a_n\}$.

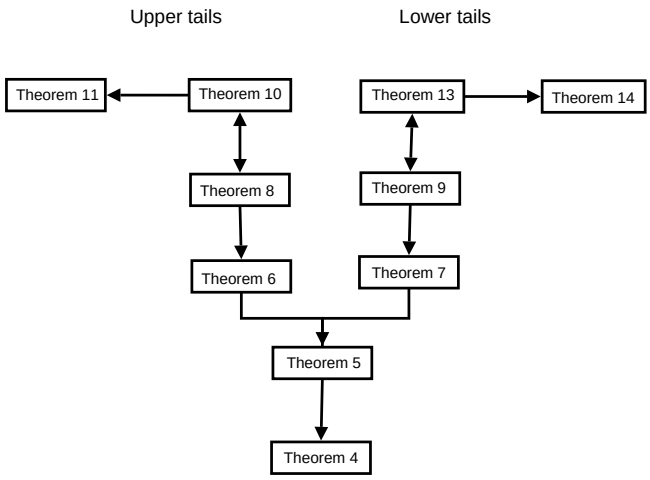

Figure 7: The flowchart for theorems on the sum of independent variables.

To compare inequalities (2) to (3), we consider an example in Figure 8. The cumulative distribution is the function $\Pr(X > x)$. The dotted curve in Figure 8 illustrates the cumulative distribution of the binomial distribution $B(1000, 0.1)$ with the value ranging from 0 to 1 as x goes from $-\infty$ to ∞. The solid curve at the lower-left corner is the bound $e^{-\lambda^2/2\nu}$ for the lower tail. The solid curve at the upper-right corner is the bound $1 - e^{-\frac{\lambda^2}{2(\nu+a\lambda/3)}}$ for the upper tail.

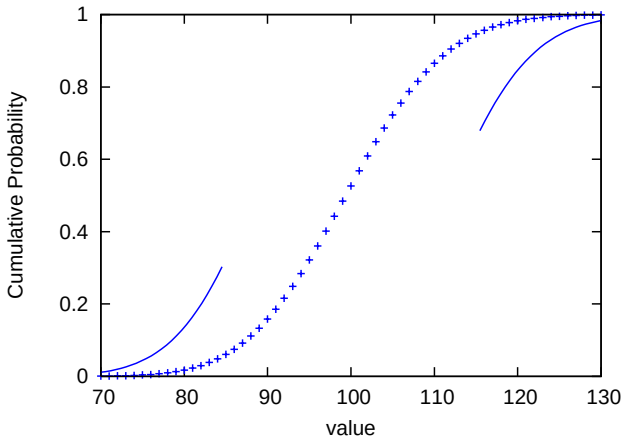

Figure 8: *Chernoff inequalities*

The inequality (3) in the above theorem is a corollary of the following general concentration inequality (also see Theorem 2.7 in the survey paper by McDiarmid [20]).

Theorem 6 *[20] Let X_i ($1 \leq i \leq n$) be independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + M$, for $1 \leq i \leq n$. We consider the sum $X = \sum_{i=1}^{n} X_i$ with expectation $\mathrm{E}(X) = \sum_{i=1}^{n} \mathrm{E}(X_i)$ and variance $\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$. Then we have*

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + M\lambda/3)}}.$$

In the other direction, we have the following inequality.

Theorem 7 *If $X_1, X_2, \ldots, X_n$ are non-negative independent random variables, we have the following bounds for the sum $X = \sum_{i=1}^{n} X_i$:*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^{n} \mathrm{E}(X_i^2)}}.$$

A strengthened version of the above theorem is as follows:

Theorem 8 *Suppose X_i are independent random variables satisfying $X_i \leq M$, for $1 \leq i \leq n$. Let $X = \sum_{i=1}^{n} X_i$ and $\|X\| = \sqrt{\sum_{i=1}^{n} \mathrm{E}(X_i^2)}$. Then we have*

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}.$$

Replacing X by $-X$ in the proof of Theorem 8, we have the following theorem for the lower tail.

Theorem 9 *Let X_i be independent random variables satisfying $X_i \geq -M$, for $1 \leq i \leq n$. Let $X = \sum_{i=1}^{n} X_i$ and $\|X\| = \sqrt{\sum_{i=1}^{n} \mathrm{E}(X_i^2)}$. Then we have*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}.$$

Before we give the proof of Theorems 8, we will first show the implications of Theorems 8 and 9. Namely, we will show that the other concentration inequalities can be derived from Theorems 8 and 9.

Fact: Theorem 8 $\Longrightarrow$ Theorem 6:

Proof: Let $X_i' = X_i - \mathrm{E}(X_i)$ and $X' = \sum_{i=1}^{n} X_i' = X - \mathrm{E}(X)$. We have

$$X_i' \leq M \quad \text{for } 1 \leq i \leq n.$$

We also have

$$\begin{aligned} \|X'\|^2 &= \sum_{i=1}^{n} \mathrm{E}(X_i'^2) \\ &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &= \mathrm{Var}(X). \end{aligned}$$

Applying Theorem 8, we get

$$\begin{aligned}\Pr(X \geq \mathrm{E}(X) + \lambda) &= \Pr(X' \geq \lambda) \\ &\leq e^{-\frac{\lambda^2}{2(\|X'\|^2 + M\lambda/3)}} \\ &\leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + M\lambda/3)}}.\end{aligned}$$

□

Fact: Theorem 9 $\Longrightarrow$ Theorem 7
The proof is straightforward by choosing $M = 0$.

Fact: Theorem 6 and 7 $\Longrightarrow$ Theorem 5
Proof: We define $Y_i = a_i X_i$. Note that

$$\|X\|^2 = \sum_{i=1}^{n} \mathrm{E}(Y_i^2) = \sum_{i=1}^{n} a_i^2 p_i = \nu.$$

Equation (2) follows from Theorem 7 since Y_i's are non-negatives.

For the other direction, we have

$$Y_i \leq a_i \leq a \leq \mathrm{E}(Y_i) + a.$$

Equation (3) follows from Theorem 6. □

Fact: Theorem 8 and Theorem 9 $\Longrightarrow$ Theorem 3
The proof is by choosing $Y = X - \mathrm{E}(X)$, $M = 1$ and applying Theorems 8 and Theorem 9 to Y.

Fact: Theorem 5 $\Longrightarrow$ Theorem 4
The proof follows by choosing $a_1 = a_2 = \cdots = a_n = 1$.

Finally, we give the complete proof of Theorem 8 and thus finish the proofs for all the above theorems on Chernoff inequalities.
Proof of Theorem 8: We consider

$$\mathrm{E}(e^{tX}) = \mathrm{E}(e^{t\sum_i X_i}) = \prod_{i=1}^{n} \mathrm{E}(e^{tX_i})$$

since the X_i's are independent.

We define $g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!} = \frac{2(e^y - 1 - y)}{y^2}$, and use the following facts about g:

- $g(0) = 1$.
- $g(y) \leq 1$, for $y < 0$.
- $g(y)$ is monotone increasing, for $y \geq 0$.

- For $y < 3$, we have

$$g(y) = 2\sum_{k=2}^{\infty} \frac{y^{k-2}}{k!} \leq \sum_{k=2}^{\infty} \frac{y^{k-2}}{3^{k-2}} = \frac{1}{1-y/3}$$

since $k! \geq 2 \cdot 3^{k-2}$.

Then we have, for $k \geq 2$,

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \prod_{i=1}^{n} \mathrm{E}(e^{tX_i}) \\
&= \prod_{i=1}^{n} \mathrm{E}(\sum_{k=0}^{\infty} \frac{t^k X_i^k}{k!}) \\
&= \prod_{i=1}^{n} \mathrm{E}(1 + t\mathrm{E}(X_i) + \frac{1}{2}t^2 X_i^2 g(tX_i)) \\
&\leq \prod_{i=1}^{n} (1 + t\mathrm{E}(X_i) + \frac{1}{2}t^2 \mathrm{E}(X_i^2) g(tM)) \\
&\leq \prod_{i=1}^{n} e^{t\mathrm{E}(X_i) + \frac{1}{2}t^2 \mathrm{E}(X_i^2) g(tM)} \\
&= e^{t\mathrm{E}(X) + \frac{1}{2}t^2 g(tM) \sum_{i=1}^{n} \mathrm{E}(X_i^2)} \\
&= e^{t\mathrm{E}(X) + \frac{1}{2}t^2 g(tM) \|X\|^2}.
\end{aligned}$$

Hence, for t satisfying $tM < 3$, we have

$$\begin{aligned}
\Pr(X \geq \mathrm{E}(X) + \lambda) &= \Pr(e^{tX} \geq e^{t\mathrm{E}(X) + t\lambda}) \\
&\leq e^{-t\mathrm{E}(X) - t\lambda} \mathrm{E}(e^{tX}) \\
&\leq e^{-t\lambda + \frac{1}{2}t^2 g(tM) \|X\|^2} \\
&\leq e^{-t\lambda + \frac{1}{2}t^2 \|X\|^2 \frac{1}{1-tM/3}}.
\end{aligned}$$

To minimize the above expression, we choose $t = \frac{\lambda}{\|X\|^2 + M\lambda/3}$. Therefore, $tM < 3$ and we have

$$\begin{aligned}
\Pr(X \geq \mathrm{E}(X) + \lambda) &\leq e^{-t\lambda + \frac{1}{2}t^2 \|X\|^2 \frac{1}{1-tM/3}} \\
&= e^{-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}}.
\end{aligned}$$

The proof is complete. □

4 More concentration inequalities

Here we state several variations and extensions of the concentration inequalities in Theorem 8. We first consider the upper tail.

Theorem 10 *Let X_i denote independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + a_i + M$, for $1 \leq i \leq n$. For, $X = \sum_{i=1}^n X_i$, we have*

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=1}^n a_i^2 + M\lambda/3)}}.$$

Proof: Let $X_i' = X_i - \mathrm{E}(X_i) - a_i$ and $X' = \sum_{i=1}^n X_i'$. We have

$$X_i' \leq M \quad \text{for } 1 \leq i \leq n.$$

$$\begin{aligned} X' - \mathrm{E}(X') &= \sum_{i=1}^n (X_i' - \mathrm{E}(X_i')) \\ &= \sum_{i=1}^n (X_i' + a_i) \\ &= \sum_{i=1}^n (X_i - \mathrm{E}(X_i)) \\ &= X - \mathrm{E}(X). \end{aligned}$$

Thus,

$$\begin{aligned} \|X'\|^2 &= \sum_{i=1}^n \mathrm{E}({X'}_i^2) \\ &= \sum_{i=1}^n \mathrm{E}((X_i - \mathrm{E}(X_i) - a_i)^2) \\ &= \sum_{i=1}^n \mathrm{E}((X_i - \mathrm{E}(X_i))^2 + a_i^2 \\ &= \mathrm{Var}(X) + \sum_{i=1}^n a_i^2. \end{aligned}$$

By applying Theorem 8, the proof is finished. □

Theorem 11 *Suppose X_i are independent random variables satisfying $X_i \leq \mathrm{E}(X_i) + M_i$, for $0 \leq i \leq n$. We order the X_i's so that the M_i are in increasing order. Let $X = \sum_{i=1}^n X_i$. Then for any $1 \leq k \leq n$, we have*

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=k}^n (M_i - M_k)^2 + M_k\lambda/3)}}.$$

Proof: For fixed k, we choose $M = M_k$ and

$$a_i = \begin{cases} 0 & \text{if } 1 \leq i \leq k, \\ M_i - M_k & \text{if } k \leq i \leq n. \end{cases}$$

We have

$$X_i - \mathrm{E}(X_i) \leq M_i \leq a_i + M_k \quad \text{for } 1 \leq k \leq n,$$

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=k}^{n} (M_i - M_k)^2.$$

Using Theorem 10, we have

$$\Pr(X_i \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X) + \sum_{i=k}^{n} (M_i - M_k)^2 + M_k \lambda/3)}}.$$

□

Example 12 Let $X_1, X_2, \ldots, X_n$ be independent random variables. For $1 \leq i \leq n-1$, suppose X_i follows the same distribution with

$$\Pr(X_i = 0) = 1 - p \quad \text{and} \quad \Pr(X_i = 1) = p,$$

and X_n follows the distribution with

$$\Pr(X_n = 0) = 1 - p \quad \text{and} \quad \Pr(X_n = \sqrt{n}) = p.$$

Consider the sum $X = \sum_{i=1}^{n} X_i$.

We have

$$\begin{aligned} \mathrm{E}(X) &= \sum_{i=1}^{n} \mathrm{E}(X_i) \\ &= (n-1)p + \sqrt{n}p. \end{aligned}$$

$$\begin{aligned} \mathrm{Var}(X) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) \\ &= (n-1)p(1-p) + np(1-p) \\ &= (2n-1)p(1-p). \end{aligned}$$

Apply Theorem 6 with $M = (1-p)\sqrt{n}$. We have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2((2n-1)p(1-p) + (1-p)\sqrt{n}\lambda/3)}}.$$

In particular, for constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\Theta(n^{\epsilon})}.$$

Now we apply Theorem 11 with $M_1 = \ldots = M_{n-1} = (1-p)$ and $M_n = \sqrt{n}(1-p)$. Choosing $k = n-1$, we have

$$\begin{aligned} \mathrm{Var}(X) + (M_n - M_{n-1})^2 &= (2n-1)p(1-p) + (1-p)^2(\sqrt{n}-1)^2 \\ &\leq (2n-1)p(1-p) + (1-p)^2 n \\ &\leq (1-p^2)n. \end{aligned}$$

Thus,

$$\Pr(X_i \geq \mathrm{E}(X) + \lambda) \leq e^{-\frac{\lambda^2}{2((1-p^2)n+(1-p)^2\lambda/3)}}.$$

For constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \geq \mathrm{E}(X) + \lambda) \leq e^{-\Theta(n^{2\epsilon})}.$$

From the above examples, we note that Theorem 11 gives a significantly better bound than that in Theorem 6 if the random variables X_i have very different upper bounds.

For completeness, we also list the corresponding theorems for the lower tails. (These can be derived by replacing X by $-X$.)

Theorem 13 *Let X_i denote independent random variables satisfying $X_i \geq \mathrm{E}(X_i) - a_i - M$, for $0 \leq i \leq n$. For $X = \sum_{i=1}^{n} X_i$, we have*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X)+\sum_{i=1}^{n} a_i^2+M\lambda/3)}}.$$

Theorem 14 *Let X_i denote independent random variables satisfying $X_i \geq \mathrm{E}(X_i) - M_i$, for $0 \leq i \leq n$. We order the X_i's so that the M_i are in increasing order. Let $X = \sum_{i=1}^{n} X_i$. Then for any $1 \leq k \leq n$, we have*

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(\mathrm{Var}(X)+\sum_{i=k}^{n}(M_i-M_k)^2+M_k\lambda/3)}}.$$

Continuing the above example, we choose $M_1 = M_2 = \ldots = M_{n-1} = p$, and $M_n = \sqrt{n}p$. We choose $k = n-1$, so we have

$$\begin{aligned} \mathrm{Var}(X) + (M_n - M_{n-1})^2 &= (2n-1)p(1-p) + p^2(\sqrt{n}-1)^2 \\ &\leq (2n-1)p(1-p) + p^2 n \\ &\leq p(2-p)n. \end{aligned}$$

Using Theorem 14, we have

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2(p(2-p)n+p^2\lambda/3)}}.$$

For a constant $p \in (0,1)$ and $\lambda = \Theta(n^{\frac{1}{2}+\epsilon})$, we have

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\Theta(n^{2\epsilon})}.$$

5 Martingales and Azuma's inequality

A martingale is a sequence of random variables $X_0, X_1, \ldots$ with finite means such that the conditional expectation of X_{n+1} given $X_0, X_1, \ldots, X_n$ is equal to X_n.

The above definition is given in the classical book of Feller [15], p. 210. However, the conditional expectation depends on the random variables under

consideration and can be difficult to deal with in various cases. In this survey we will use the following definition which is concise and basically equivalent for the finite cases.

Suppose that Ω is a probability space with a probability distribution p. Let $\mathcal{F}$ denote a σ-field on Ω. (A σ-field on Ω is a collection of subsets of Ω which contains $\emptyset$ and Ω, and is closed under unions, intersections, and complementation.) In a σ-field $\mathcal{F}$ of Ω, the smallest set in $\mathcal{F}$ containing an element x is the intersection of all sets in $\mathcal{F}$ containing x. A function $f : \Omega \to \mathbb{R}$ is said to be $\mathcal{F}$-measurable if $f(x) = f(y)$ for any y in the smallest set containing x. (For more terminology on martingales, the reader is referred to [17].)

If $f : \Omega \to \mathbb{R}$ is a function, we define the expectation $\mathrm{E}(f) = \mathrm{E}(f(x) \mid x \in \Omega)$ by

$$\mathrm{E}(f) = \mathrm{E}(f(x) \mid x \in \Omega) := \sum_{x \in \Omega} f(x)p(x).$$

If $\mathcal{F}$ is a σ-field on Ω, we define the conditional expectation $\mathrm{E}(f \mid \mathcal{F}) : \Omega \to \mathbb{R}$ by the formula

$$\mathrm{E}(f \mid \mathcal{F})(x) := \frac{1}{\sum_{y \in \mathcal{F}(x)} p(y)} \sum_{y \in \mathcal{F}(x)} f(y)p(y)$$

where $\mathcal{F}(x)$ is the smallest element of $\mathcal{F}$ which contains x.

A *filter* $\mathbf{F}$ is an increasing chain of σ-subfields

$$\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

A martingale (obtained from) X is associated with a filter $\mathbf{F}$ and a sequence of random variables $X_0, X_1, \ldots, X_n$ satisfying $X_i = \mathrm{E}(X \mid \mathcal{F}_i)$ and, in particular, $X_0 = \mathrm{E}(X)$ and $X_n = X$.

Example 15 For given independent random variables $Y_1, Y_2, \ldots, Y_n$, we can define a martingale $X = Y_1 + Y_2 + \cdots + Y_n$ as follows. Let $\mathcal{F}_i$ be the σ-field generated by $Y_1, \ldots, Y_i$. (In other words, $\mathcal{F}_i$ is the minimum σ-field so that $Y_1, \ldots, Y_i$ are $\mathcal{F}_i$-measurable.) We have a natural filter $\mathbf{F}$:

$$\{0, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

Let $X_i = \sum_{j=1}^{i} Y_j + \sum_{j=i+1}^{n} \mathrm{E}(Y_j)$. Then, $X_0, X_1, X_2, \ldots, X_n$ forms a martingale corresponding to the filter $\mathbf{F}$.

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ a vector with positive entries, the martingale X is said to be $\mathbf{c}$-Lipschitz if

$$|X_i - X_{i-1}| \leq c_i \tag{4}$$

for $i = 1, 2, \ldots, n$. A powerful tool for controlling martingales is the following:

Theorem 16 (Azuma's inequality) *If a martingale X is* $\mathbf{c}$*-Lipschitz, then*

$$\Pr(|X - \mathrm{E}(X)| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}, \tag{5}$$

where $\mathbf{c} = (c_1, \ldots, c_n)$.

Theorem 17 *Let $X_1, X_2, \ldots, X_n$ be independent random variables satisfying*

$$|X_i - \mathrm{E}(X_i)| \leq c_i \quad \text{for } 1 \leq i \leq n.$$

Then we have the following bound for the sum $X = \sum_{i=1}^n X_i$.

$$\Pr(|X - \mathrm{E}(X)| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}.$$

Proof of Azuma's inequality: For a fixed t, we consider the convex function $f(x) = e^{tx}$. For any $|x| \leq c$, $f(x)$ is below the line segment from $(-c, f(-c))$ to $(c, f(c))$. In other words, we have

$$e^{tx} \leq \frac{1}{2c}(e^{tc} - e^{-tc})x + \frac{1}{2}(e^{tc} + e^{-tc}).$$

Therefore, we can write

$$\begin{aligned}
\mathrm{E}(e^{t(X_i - X_{i-1})}|\mathcal{F}_{i-1}) &\leq \mathrm{E}(\frac{1}{2c_i}(e^{tc_i} - e^{-tc_i})(X_i - X_{i-1}) + \frac{1}{2}(e^{tc_i} + e^{-tc_i})|\mathcal{F}_{i-1}) \\
&= \frac{1}{2}(e^{tc_i} + e^{-tc_i}) \\
&\leq e^{t^2 c_i^2/2}.
\end{aligned}$$

Here we apply the conditions $\mathrm{E}(X_i - X_{i-1}|\mathcal{F}_{i-1}) = 0$ and $|X_i - X_{i-1}| \leq c_i$.

Hence,

$$\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) \leq e^{t^2 c_i^2/2} e^{tX_{i-1}}.$$

Inductively, we have

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \mathrm{E}(\mathrm{E}(e^{tX_n}|\mathcal{F}_{n-1})) \\
&\leq e^{t^2 c_n^2/2}\mathrm{E}(e^{tX_{n-1}}) \\
&\leq \cdots \\
&\leq \prod_{i=1}^n e^{t^2 c_i^2/2}\mathrm{E}(e^{tX_0}) \\
&= e^{\frac{1}{2}t^2 \sum_{i=1}^n c_i^2} e^{t\mathrm{E}(X)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\Pr(X \geq \mathrm{E}(X) + \lambda) &= \Pr(e^{t(X - \mathrm{E}(X))} \geq e^{t\lambda}) \\
&\leq e^{-t\lambda}\mathrm{E}(e^{t(X - \mathrm{E}(X))}) \\
&\leq e^{-t\lambda} e^{\frac{1}{2}t^2 \sum_{i=1}^n c_i^2} \\
&= e^{-t\lambda + \frac{1}{2}t^2 \sum_{i=1}^n c_i^2}.
\end{aligned}$$

We choose $t = \frac{\lambda}{\sum_{i=1}^n c_i^2}$ (in order to minimize the above expression). We have

$$\begin{aligned} \Pr(X \geq \mathrm{E}(X) + \lambda) &\leq e^{-t\lambda + \frac{1}{2}t^2 \sum_{i=1}^n c_i^2} \\ &= e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}. \end{aligned}$$

To derive a similar lower bound, we consider $-X_i$ instead of X_i in the preceding proof. Then we obtain the following bound for the lower tail.

$$\Pr(X \leq \mathrm{E}(X) - \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}.$$

□

6 General martingale inequalities

Many problems which can be set up as a martingale do not satisfy the Lipschitz condition. It is desirable to be able to use tools similar to the Azuma inequality in such cases. In this section, we will first state and then prove several extensions of the Azuma inequality (see Figure 9).

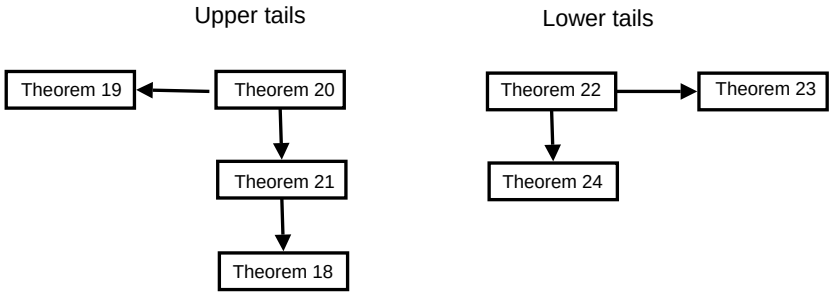

Figure 9: The flowchart for theorems on martingales.

Our starting point is the following well known concentration inequality (see [20]):

Theorem 18 *Let X be the martingale associated with a filter* $\mathbf{F}$ *satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $|X_i - X_{i-1}| \leq M$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - \mathrm{E}(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n \sigma_i^2 + M\lambda/3)}}.$$

Since the sum of independent random variables can be viewed as a martingale (see Example 15), Theorem 18 implies Theorem 6. In a similar way, the following theorem is associated with Theorem 10.

Theorem 19 *Let X be the martingale associated with a filter* $\mathbf{F}$ *satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_i - X_{i-1} \leq M_i$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - \mathrm{E}(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^{n}(\sigma_i^2 + M_i^2)}}.$$

The above theorem can be further generalized:

Theorem 20 *Let X be the martingale associated with a filter* $\mathbf{F}$ *satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_i - X_{i-1} \leq a_i + M$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - \mathrm{E}(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}(\sigma_i^2 + a_i^2) + M\lambda/3)}}.$$

Theorem 20 implies Theorem 18 by choosing $a_1 = a_2 = \cdots = a_n = 0$.

We also have the following theorem corresponding to Theorem 11.

Theorem 21 *Let X be the martingale associated with a filter* $\mathbf{F}$ *satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_i - X_{i-1} \leq M_i$, *for* $1 \leq i \leq n$.

Then, for any M, we have

$$\Pr(X - \mathrm{E}(X) \geq \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}\sigma_i^2 + \sum_{M_i > M}(M_i - M)^2 + M\lambda/3)}}.$$

Theorem 20 implies Theorem 21 by choosing

$$a_i = \begin{cases} 0 & \text{if } M_i \leq M, \\ M_i - M & \text{if } M_i \geq M. \end{cases}$$

It suffices to prove Theorem 20 so that all the above stated theorems hold.

Proof of Theorem 20:

Recall that $g(y) = 2\sum_{k=2}^{\infty}\frac{y^{k-2}}{k!}$ satisfies the following properties:

- $g(y) \leq 1$, for $y < 0$.
- $\lim_{y \to 0} g(y) = 1$.
- $g(y)$ is monotone increasing, for $y \geq 0$.
- When $b < 3$, we have $g(b) \leq \frac{1}{1 - b/3}$.

Since $\mathrm{E}(X_i|\mathcal{F}_{i-1}) = X_{i-1}$ and $X_i - X_{i-1} - a_i \le M$, we have

$$\begin{aligned}
\mathrm{E}(e^{t(X_i - X_{i-1} - a_i)}|\mathcal{F}_{i-1}) &= \mathrm{E}(\sum_{k=0}^{\infty} \frac{t^k}{k!}(X_i - X_{i-1} - a_i)^k|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \mathrm{E}(\sum_{k=2}^{\infty} \frac{t^k}{k!}(X_i - X_{i-1} - a_i)^k|\mathcal{F}_{i-1}) \\
&\le 1 - ta_i + \mathrm{E}(\frac{t^2}{2}(X_i - X_{i-1} - a_i)^2 g(tM)|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \frac{t^2}{2} g(tM)\mathrm{E}((X_i - X_{i-1} - a_i)^2|\mathcal{F}_{i-1}) \\
&= 1 - ta_i + \frac{t^2}{2} g(tM)(\mathrm{E}((X_i - X_{i-1})^2|\mathcal{F}_{i-1}) + a_i^2) \\
&\le 1 - ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2) \\
&\le e^{-ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) &= \mathrm{E}(e^{t(X_i - X_{i-1} - a_i)}|\mathcal{F}_{i-1}) e^{tX_{i-1} + ta_i} \\
&\le e^{-ta_i + \frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2)} e^{tX_{i-1} + ta_i} \\
&= e^{\frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2)} e^{tX_{i-1}}.
\end{aligned}$$

Inductively, we have

$$\begin{aligned}
\mathrm{E}(e^{tX}) &= \mathrm{E}(\mathrm{E}(e^{tX_n}|\mathcal{F}_{n-1})) \\
&\le e^{\frac{t^2}{2} g(tM)(\sigma_n^2 + a_n^2)} \mathrm{E}(e^{tX_{n-1}}) \\
&\le \cdots \\
&\le \prod_{i=1}^{n} e^{\frac{t^2}{2} g(tM)(\sigma_i^2 + a_i^2)} \mathrm{E}(e^{tX_0}) \\
&= e^{\frac{1}{2} t^2 g(tM) \sum_{i=1}^{n} (\sigma_i^2 + a_i^2)} e^{t\mathrm{E}(X)}.
\end{aligned}$$

Then for t satisfying $tM < 3$, we have

$$\begin{aligned}
\Pr(X \ge \mathrm{E}(X) + \lambda) &= \Pr(e^{tX} \ge e^{t\mathrm{E}(X) + t\lambda}) \\
&\le e^{-t\mathrm{E}(X) - t\lambda} \mathrm{E}(e^{tX}) \\
&\le e^{-t\lambda} e^{\frac{1}{2} t^2 g(tM) \sum_{i=1}^{n} (\sigma_i^2 + a_i^2)} \\
&= e^{-t\lambda + \frac{1}{2} t^2 g(tM) \sum_{i=1}^{n} (\sigma_i^2 + a_i^2)} \\
&\le e^{-t\lambda + \frac{1}{2} \frac{t^2}{1 - tM/3} \sum_{i=1}^{n} (\sigma_i^2 + a_i^2)}
\end{aligned}$$

We choose $t = \frac{\lambda}{\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+M\lambda/3}$. Clearly $tM < 3$ and

$$\begin{aligned}\Pr(X \geq \mathrm{E}(X) + \lambda) &\leq e^{-t\lambda+\frac{1}{2}\frac{t^2}{1-tM/3}\sum_{i=1}^{n}(\sigma_i^2+a_i^2)} \\ &= e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+M\lambda/3)}}.\end{aligned}$$

The proof of the theorem is complete. □

For completeness, we state the following theorems for the lower tails. The proofs are almost identical and will be omitted.

Theorem 22 *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_{i-1} - X_i \leq a_i + M$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - \mathrm{E}(X) \leq -\lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+M\lambda/3)}}.$$

Theorem 23 *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_{i-1} - X_i \leq M_i$, *for* $1 \leq i \leq n$.

Then, we have

$$\Pr(X - \mathrm{E}(X) \leq -\lambda) \leq e^{-\frac{\lambda^2}{2\sum_{i=1}^{n}(\sigma_i^2+M_i^2)}}.$$

Theorem 24 *Let X be the martingale associated with a filter $\mathbf{F}$ satisfying*

1. $\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$, *for* $1 \leq i \leq n$;
2. $X_{i-1} - X_i \leq M_i$, *for* $1 \leq i \leq n$.

Then, for any M, we have

$$\Pr(X - \mathrm{E}(X) \leq -\lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}\sigma_i^2+\sum_{M_i>M}(M_i-M)^2+M\lambda/3)}}.$$

7 Supermartingales and Submartingales

In this section, we consider further strengthened versions of the martingale inequalities that were mentioned so far. Instead of a fixed upper bound for the variance, we will assume that the variance $\mathrm{Var}(X_i|\mathcal{F}_{i-1})$ is upper bounded by a linear function of X_{i-1}. Here we assume this linear function is non-negative for all values that X_{i-1} takes. We first need some terminology.

For a filter $\mathbf{F}$:

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

a sequence of random variables $X_0, X_1, \ldots, X_n$ is called a *submartingale* if X_i is $\mathcal{F}_i$-measurable (i.e., $X_i(a) = X_i(b)$ if all elements of $\mathcal{F}_i$ containing a also contain b and vice versa) then $\mathrm{E}(X_i \mid \mathcal{F}_{i-1}) \leq X_{i-1}$, for $1 \leq i \leq n$.

A sequence of random variables $X_0, X_1, \ldots, X_n$ is said to be a *supermartingale* if X_i is $\mathcal{F}_i$-measurable and $\mathrm{E}(X_i \mid \mathcal{F}_{i-1}) \geq X_{i-1}$, for $1 \leq i \leq n$.

To avoid repetition, we will first state a number of useful inequalities for submartingales and supermartingales. Then we will give the proof for the general inequalities in Theorem 27 for submartingales and in Theorem 29 for supermartingales. Furthermore, we will show that all the stated theorems follow from Theorems 27 and 29 (See Figure 10). Note that the inequalities for submartingale and supermartingale are not quite symmetric.

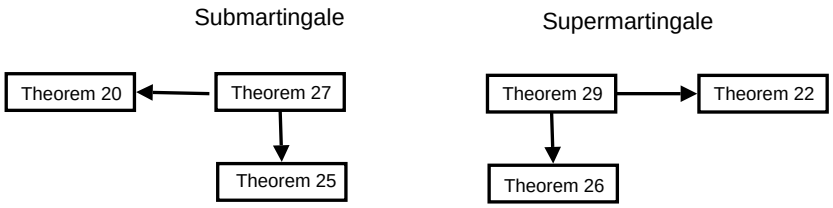

Figure 10: The flowchart for theorems on submartingales and supermartingales

Theorem 25 *Suppose that a submartingale X, associated with a filter $\mathbf{F}$, satisfies*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \phi_i X_{i-1}$$

and

$$X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) \leq M$$

for $1 \leq i \leq n$. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2((X_0+\lambda)(\sum_{i=1}^{n}\phi_i)+M\lambda/3)}}.$$

Theorem 26 *Suppose that a supermartingale X, associated with a filter $\mathbf{F}$, satisfies, for $1 \leq i \leq n$,*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \phi_i X_{i-1}$$

and

$$\mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i \leq M.$$

Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{2(X_0(\sum_{i=1}^{n}\phi_i)+M\lambda/3)}},$$

for any $\lambda \leq X_0$.

Theorem 27 *Suppose that a submartingale X, associated with a filter $\mathbf{F}$, satisfies*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2 + \phi_i X_{i-1}$$

and

$$X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) \le a_i + M$$

for $1 \le i \le n$. Here σ_i, a_i, ϕ_i and M are non-negative constants. Then we have

$$\Pr(X_n \ge X_0 + \lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^n\phi_i)+M\lambda/3)}}.$$

Remark 28 *Theorem 27 implies Theorem 25 by setting all σ_i's and a_i's to zero. Theorem 27 also implies Theorem 20 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

The theorem for a supermartingale is slightly different due to the asymmetry of the condition on the variance.

Theorem 29 *Suppose a supermartingale X, associated with a filter $\mathbf{F}$, satisfies, for $1 \le i \le n$,*

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \le \sigma_i^2 + \phi_i X_{i-1}$$

and

$$\mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i \le a_i + M,$$

where M, a_i's, σ_i's, and ϕ_i's are non-negative constants. Then we have

$$\Pr(X_n \le X_0 - \lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+X_0(\sum_{i=1}^n\phi_i)+M\lambda/3)}},$$

for any $\lambda \le 2X_0 + \frac{\sum_{i=1}^n(\sigma_i^2+a_i^2)}{\sum_{i=1}^n\phi_i}$.

Remark 30 *Theorem 29 implies Theorem 26 by setting all σ_i's and a_i's to zero. Theorem 29 also implies Theorem 22 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

Proof of Theorem 27:

For a positive t (to be chosen later), we consider

$$\begin{aligned}
\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) &= e^{t\mathrm{E}(X_i|\mathcal{F}_{i-1})+ta_i}\mathrm{E}(e^{t(X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)}|\mathcal{F}_{i-1}) \\
&= e^{t\mathrm{E}(X_i|\mathcal{F}_{i-1})+ta_i}\sum_{k=0}^{\infty}\frac{t^k}{k!}\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^k|\mathcal{F}_{i-1}) \\
&\le e^{t\mathrm{E}(X_i|\mathcal{F}_{i-1})+\sum_{k=2}^{\infty}\frac{t^k}{k!}\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^k|\mathcal{F}_{i-1})}
\end{aligned}$$

Recall that $g(y) = 2\sum_{k=2}^{\infty}\frac{y^{k-2}}{k!}$ satisfies

$$g(y) \le g(b) < \frac{1}{1-b/3}$$

for all $y \le b$ and $0 \le b \le 3$.

Since $X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) - a_i \le M$, we have

$$\begin{aligned}\sum_{k=2}^{\infty}\frac{t^k}{k!}\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^k|\mathcal{F}_{i-1}) &\le \frac{g(tM)}{2}t^2\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^2|\mathcal{F}_{i-1})\\ &= \frac{g(tM)}{2}t^2(\mathrm{Var}(X_i|\mathcal{F}_{i-1})+a_i^2)\\ &\le \frac{g(tM)}{2}t^2(\sigma_i^2+\phi_i X_{i-1}+a_i^2).\end{aligned}$$

Since $\mathrm{E}(X_i|\mathcal{F}_{i-1}) \le X_{i-1}$, we have

$$\begin{aligned}\mathrm{E}(e^{tX_i}|\mathcal{F}_{i-1}) &\le e^{t\mathrm{E}(X_i|\mathcal{F}_{i-1})+\sum_{k=2}^{\infty}\frac{t^k}{k!}\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1}-)-a_i)^k|\mathcal{F}_{i-1})}\\ &\le e^{tX_{i-1}+\frac{g(tM)}{2}t^2(\sigma_i^2+\phi_i X_{i-1}+a_i^2)}\\ &= e^{(t+\frac{g(tM)}{2}\phi_i t^2)X_{i-1}}e^{\frac{t^2}{2}g(tM)(\sigma_i^2+a_i^2)}.\end{aligned}$$

We define $t_i \ge 0$ for $0 < i \le n$, satisfying

$$t_{i-1} = t_i + \frac{g(t_0M)}{2}\phi_i t_i^2,$$

while t_0 will be chosen later. Then

$$t_n \le t_{n-1} \le \cdots \le t_0,$$

and

$$\begin{aligned}\mathrm{E}(e^{t_iX_i}|\mathcal{F}_{i-1}) &\le e^{(t_i+\frac{g(t_iM)}{2}\phi_i t_i^2)X_{i-1}}e^{\frac{t_i^2}{2}g(t_iM)(\sigma_i^2+a_i^2)}\\ &\le e^{(t_i+\frac{g(t_0M)}{2}t_i^2\phi_i)X_{i-1}}e^{\frac{t_i^2}{2}g(t_iM)(\sigma_i^2+a_i^2)}\\ &= e^{t_{i-1}X_{i-1}}e^{\frac{t_i^2}{2}g(t_iM)(\sigma_i^2+a_i^2)}\end{aligned}$$

since $g(y)$ is increasing for $y > 0$.

By Markov's inequality, we have

$$\begin{aligned}\Pr(X_n \ge X_0+\lambda) &\le e^{-t_n(X_0+\lambda)}\mathrm{E}(e^{t_nX_n})\\ &= e^{-t_n(X_0+\lambda)}\mathrm{E}(\mathrm{E}(e^{t_nX_n}|\mathcal{F}_{n-1}))\\ &\le e^{-t_n(X_0+\lambda)}\mathrm{E}(e^{t_{n-1}X_{n-1}})e^{\frac{t_i^2}{2}g(t_iM)(\sigma_i^2+a_i^2)}\\ &\le \cdots\\ &\le e^{-t_n(X_0+\lambda)}\mathrm{E}(e^{t_0X_0})e^{\sum_{i=1}^{n}\frac{t_i^2}{2}g(t_iM)(\sigma_i^2+a_i^2)}\\ &\le e^{-t_n(X_0+\lambda)+t_0X_0+\frac{t_0^2}{2}g(t_0M)\sum_{i=1}^{n}(\sigma_i^2+a_i^2)}.\end{aligned}$$

Note that

$$\begin{aligned} t_n &= t_0 - \sum_{i=1}^{n}(t_{i-1} - t_i) \\ &= t_0 - \sum_{i=1}^{n} \frac{g(t_0 M)}{2}\phi_i t_i^2 \\ &\geq t_0 - \frac{g(t_0 M)}{2} t_0^2 \sum_{i=1}^{n} \phi_i. \end{aligned}$$

Hence

$$\begin{aligned} \Pr(X_n \geq X_0 + \lambda) &\leq e^{-t_n(X_0+\lambda)+t_0 X_0 + \frac{t_0^2}{2} g(t_0 M)\sum_{i=1}^{n}(\sigma_i^2+a_i^2)} \\ &\leq e^{-(t_0 - \frac{g(t_0 M)}{2} t_0^2 \sum_{i=1}^{n}\phi_i)(X_0+\lambda)+t_0 X_0 + \frac{t_0^2}{2} g(t_0 M)\sum_{i=1}^{n}(\sigma_i^2+a_i^2)} \\ &= e^{-t_0\lambda + \frac{g(t_0 M)}{2} t_0^2 (\sum_{i=1}^{n}(\sigma_i^2+a_i^2) + (X_0+\lambda)\sum_{i=1}^{n}\phi_i)} \end{aligned}$$

Now we choose $t_0 = \frac{\lambda}{\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^{n}\phi_i)+M\lambda/3}$. Using the fact that $t_0 M < 3$, we have

$$\begin{aligned} \Pr(X_n \geq X_0 + \lambda) &\leq e^{-t_0\lambda + t_0^2(\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+(X_0+\lambda)\sum_{i=1}^{n}\phi_i)\frac{1}{2(1-t_0 M/3)}} \\ &= e^{-\frac{\lambda^2}{2(\sum_{i=1}^{n}(\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^{n}\phi_i)+M\lambda/3)}}. \end{aligned}$$

The proof of the theorem is complete. □

Proof of Theorem 29:

The proof is quite similar to that of Theorem 27. The following inequality still holds.

$$\begin{aligned} \mathrm{E}(e^{-tX_i}|\mathcal{F}_{i-1}) &= e^{-t\mathrm{E}(X_i|\mathcal{F}_{i-1})+ta_i}\mathrm{E}(e^{-t(X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})+a_i)}|\mathcal{F}_{i-1}) \\ &= e^{-t\mathrm{E}(X_i|\mathcal{F}_{i-1})+ta_i}\sum_{k=0}^{\infty}\frac{t^k}{k!}\mathrm{E}((\mathrm{E}(X_i|\mathcal{F}_{i-1})-X_i-a_i)^k|\mathcal{F}_{i-1}) \\ &\leq e^{-t\mathrm{E}(X_i|\mathcal{F}_{i-1})+\sum_{k=2}^{\infty}\frac{t^k}{k!}\mathrm{E}((\mathrm{E}(X_i|\mathcal{F}_{i-1})-X_i-a_i)^k|\mathcal{F}_{i-1})} \\ &\leq e^{-t\mathrm{E}(X_i|\mathcal{F}_{i-1})+\frac{g(tM)}{2}t^2\mathrm{E}((X_i-\mathrm{E}(X_i|\mathcal{F}_{i-1})-a_i)^2)} \\ &\leq e^{-t\mathrm{E}(X_i|\mathcal{F}_{i-1})+\frac{g(tM)}{2}t^2(\mathrm{Var}(X_i|\mathcal{F}_{i-1})+a_i^2)} \\ &\leq e^{-(t-\frac{g(tM)}{2}t^2\phi_i)X_{i-1}}e^{\frac{g(tM)}{2}t^2(\sigma_i^2+a_i^2)}. \end{aligned}$$

We now define $t_i \geq 0$, for $0 \leq i < n$ satisfying

$$t_{i-1} = t_i - \frac{g(t_n M)}{2}\phi_i t_i^2.$$

t_n will be defined later. Then we have

$$t_0 \leq t_1 \leq \cdots \leq t_n,$$

and

$$\begin{aligned}
\mathrm{E}(e^{-t_i X_i}|\mathcal{F}_{i-1}) &\leq e^{-(t_i - \frac{g(t_i M)}{2} t_i^2 \phi_i) X_{i-1}} e^{\frac{g(t_i M)}{2} t_i^2 (\sigma_i^2 + a_i^2)} \\
&\leq e^{-(t_i - \frac{g(t_n M)}{2} t_i^2 \phi_i) X_{i-1}} e^{\frac{g(t_n M)}{2} t_i^2 (\sigma_i^2 + a_i^2)} \\
&= e^{-t_{i-1} X_{i-1}} e^{\frac{g(t_n M)}{2} t_i^2 (\sigma_i^2 + a_i^2)}.
\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}
\Pr(X_n \leq X_0 - \lambda) &= \Pr(-t_n X_n \geq -t_n (X_0 - \lambda)) \\
&\leq e^{t_n (X_0 - \lambda)} \mathrm{E}(e^{-t_n X_n}) \\
&= e^{t_n (X_0 - \lambda)} \mathrm{E}(\mathrm{E}(e^{-t_n X_n}|\mathcal{F}_{n-1})) \\
&\leq e^{t_n (X_0 - \lambda)} \mathrm{E}(e^{-t_{n-1} X_{n-1}}) e^{\frac{g(t_n M)}{2} t_n^2 (\sigma_n^2 + a_n^2)} \\
&\leq \cdots \\
&\leq e^{t_n (X_0 - \lambda)} \mathrm{E}(e^{-t_0 X_0}) e^{\sum_{i=1}^n \frac{g(t_n M)}{2} t_i^2 (\sigma_i^2 + a_i^2)} \\
&\leq e^{t_n (X_0 - \lambda) - t_0 X_0 + \frac{t_n^2}{2} g(t_n M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)}.
\end{aligned}$$

We note

$$\begin{aligned}
t_0 &= t_n + \sum_{i=1}^n (t_{i-1} - t_i) \\
&= t_n - \sum_{i=1}^n \frac{g(t_n M)}{2} \phi_i t_i^2 \\
&\geq t_n - \frac{g(t_n M)}{2} t_n^2 \sum_{i=1}^n \phi_i.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
\Pr(X_n \leq X_0 - \lambda) &\leq e^{t_n (X_0 - \lambda) - t_0 X_0 + \frac{t_n^2}{2} g(t_n M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)} \\
&\leq e^{t_n (X_0 - \lambda) - (t_n - \frac{g(t_n M)}{2} t_n^2) X_0 + \frac{t_n^2}{2} g(t_n M) \sum_{i=1}^n (\sigma_i^2 + a_i^2)} \\
&= e^{-t_n \lambda + \frac{g(t_n M)}{2} t_n^2 (\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (\sum_{i=1}^n \phi_i) X_0)}
\end{aligned}$$

We choose $t_n = \frac{\lambda}{\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (\sum_{i=1}^n \phi_i) X_0 + M\lambda/3}$. We have $t_n M < 3$ and

$$\begin{aligned}
\Pr(X_n \leq X_0 - \lambda) &\leq e^{-t_n \lambda + t_n^2 (\sum_{i=1}^n (\sigma_i^2 + a_i^2) + (\sum_{i=1}^n \phi_i) X_0) \frac{1}{2(1 - t_n M/3)}} \\
&\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2 + a_i^2) + X_0 (\sum_{i=1}^n \phi_i) + M\lambda/3)}}.
\end{aligned}$$

It remains to verify that all t_i's are non-negative. Indeed,

$$\begin{aligned} t_i &\geq t_0 \\ &\geq t_n - \frac{g(t_n M)}{2} t_n^2 \sum_{i=1}^{n} \phi_i \\ &\geq t_n(1 - \frac{1}{2(1 - t_n M/3)} t_n \sum_{i=1}^{n} \phi_i) \\ &= t_n(1 - \frac{\lambda}{2X_0 + \frac{\sum_{i=1}^{n}(\sigma_i^2 + a_i^2)}{\sum_{i=1}^{n} \phi_i}}) \\ &\geq 0. \end{aligned}$$

The proof of the theorem is complete. □

8 The decision tree and relaxed concentration inequalities

In this section, we will extend and generalize previous theorems to a martingale which is not strictly Lipschitz but is *nearly* Lipschitz. Namely, the (Lipschitz-like) assumptions are allowed to fail for relatively small subsets of the probability space and we can still have similar but weaker concentration inequalities. Similar techniques have been introduced by Kim and Vu [19] in their important work on deriving concentration inequalities for multivariate polynomials. The basic setup for decision trees can be found in [5] and has been used in the work of Alon, Kim and Spencer [7]. Wormald [22] considers martingales with a 'stopping time' that has a similar flavor. Here we use a rather general setting and we shall give a complete proof here.

We are only interested in finite probability spaces and we use the following computational model. The random variable X can be evaluated by a sequence of decisions $Y_1, Y_2, \ldots, Y_n$. Each decision has finitely many outputs. The probability that an output is chosen depends on the previous history. We can describe the process by a decision tree T, a complete rooted tree with depth n. Each edge uv of T is associated with a probability p_{uv} depending on the decision made from u to v. Note that for any node u, we have

$$\sum_v p_{u,v} = 1.$$

We allow p_{uv} to be zero and thus include the case of having fewer than r outputs, for some fixed r. Let Ω_i denote the probability space obtained after the first i decisions. Suppose $\Omega = \Omega_n$ and X is the random variable on Ω. Let $\pi_i : \Omega \to \Omega_i$ be the projection mapping each point to the subset of points with the same first i decisions. Let $\mathcal{F}_i$ be the σ-field generated by $Y_1, Y_2, \ldots, Y_i$. (In fact, $\mathcal{F}_i = \pi_i^{-1}(2^{\Omega_i})$ is the full σ-field via the projection π_i.) The $\mathcal{F}_i$ form a natural

filter:

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F}.$$

The leaves of the decision tree are exactly the elements of Ω. Let $X_0, X_1, \ldots, X_n = X$ denote the sequence of decisions to evaluate X. Note that X_i is $\mathcal{F}_i$-measurable, and can be interpreted as a labeling on nodes at depth i.

There is one-to-one correspondence between the following:

- A sequence of random variables $X_0, X_1, \ldots, X_n$ satisfying X_i is $\mathcal{F}_i$-measurable, for $i = 0, 1, \ldots, n$.
- A vertex labeling of the decision tree T, $f \colon V(T) \to \mathbb{R}$.

In order to simplify and unify the proofs for various general types of martingales, here we introduce a definition for a function $f : V(T) \to \mathbb{R}$. We say f satisfies an *admissible* condition P if $P = \{P_v\}$ holds for every vertex v.
Examples of admissible conditions:

1. **Supermartingale:** For $1 \leq i \leq n$, we have

   $$\mathrm{E}(X_i|\mathcal{F}_{i-1}) \geq X_{i-1}.$$

   Thus the admissible condition P_u holds if

   $$f(u) \leq \sum_{v \in C(u)} p_{uv} f(v)$$

   where C_u is the set of all children nodes of u and p_{uv} is the transition probability at the edge uv.

2. **Submartingale:** For $1 \leq i \leq n$, we have

   $$\mathrm{E}(X_i|\mathcal{F}_{i-1}) \leq X_{i-1}.$$

   In this case, the admissible condition of the submartingale is

   $$f(u) \geq \sum_{v \in C(u)} p_{uv} f(v).$$

3. **Martingale:** For $1 \leq i \leq n$, we have

   $$\mathrm{E}(X_i|\mathcal{F}_{i-1}) = X_{i-1}.$$

   The admissible condition of the martingale is then:

   $$f(u) = \sum_{v \in C(u)} p_{uv} f(v).$$

4. **c-Lipschitz:** For $1 \leq i \leq n$, we have

$$|X_i - X_{i-1}| \leq c_i.$$

The admissible condition of the **c**-Lipschitz property can be described as follows:

$$|f(u) - f(v)| \leq c_i, \quad \text{for any child } v \in C(u)$$

where the node u is at level i of the decision tree.

5. **Bounded Variance:** For $1 \leq i \leq n$, we have

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2$$

for some constants σ_i.

The admissible condition of the bounded variance property can be described as:

$$\sum_{v \in C(u)} p_{uv} f^2(v) - (\sum_{v \in C(u)} p_{uv} f(v))^2 \leq \sigma_i^2.$$

6. **General Bounded Variance:** For $1 \leq i \leq n$, we have

$$\mathrm{Var}(X_i|\mathcal{F}_{i-1}) \leq \sigma_i^2 + \phi_i X_{i-1}$$

where σ_i, ϕ_i are non-negative constants, and $X_i \geq 0$. The admissible condition of the general bounded variance property can be described as follows:

$$\sum_{v \in C(u)} p_{uv} f^2(v) - (\sum_{v \in C(u)} p_{uv} f(v))^2 \leq \sigma_i^2 + \phi_i f(u), \quad \text{and } f(u) \geq 0$$

where i is the depth of the node u.

7. **Upper-bound:** For $1 \leq i \leq n$, we have

$$X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) \leq a_i + M$$

where a_i's, and M are non-negative constants. The admissible condition of the upper bounded property can be described as follows:

$$f(v) - \sum_{v \in C(u)} p_{uv} f(v) \leq a_i + M, \quad \text{for any child } v \in C(u)$$

where i is the depth of the node u.

8. **Lower-bound:** For $1 \leq i \leq n$, we have

$$\mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i \leq a_i + M$$

where a_i's, and M are non-negative constants. The admissible condition of the lower bounded property can be described as follows:

$$(\sum_{v \in C(u)} p_{uv} f(v)) - f(v) \leq a_i + M, \quad \text{for any child } v \in C(u)$$

where i is the depth of the node u.

For any labeling f on T and fixed vertex r, we can define a new labeling f_r as follows:

$$f_r(u) = \begin{cases} f(r) & \text{if } u \text{ is a descedant of } r. \\ f(u) & \text{otherwise.} \end{cases}$$

A property P is said to be *invariant* under subtree-unification if for any tree labeling f satisfying P, and a vertex r, f_r satisfies P.

We have the following theorem.

Theorem 31 *The eight properties as stated in the preceding examples — supermartingale, submartingale, martingale,* **c***-Lipschitz, bounded variance, general bounded variance, upper-bounded, and lower-bounded properties are all invariant under subtree-unification.*

Proof: We note that these properties are all admissible conditions. Let P denote any one of these. For any node u, if u is not a descendant of r, then f_r and f have the same value on v and its children nodes. Hence, P_u holds for f_r since P_u does for f.

If u is a descendant of r, then $f_r(u)$ takes the same value as $f(r)$ as well as its children nodes. We verify P_u in each case. Assume that u is at level i of the decision tree T.

1. For supermartingale, submartingale, and martingale properties, we have

$$\begin{aligned} \sum_{v \in C(u)} p_{uv} f_r(v) &= \sum_{v \in C(u)} p_{uv} f(r) \\ &= f(r) \sum_{v \in C(u)} p_{uv} \\ &= f(r) \\ &= f_r(u). \end{aligned}$$

   Hence, P_u holds for f_r.

2. For **c**-Lipschitz property, we have

$$|f_r(u) - f_r(v)| = 0 \leq c_i, \quad \text{for any child } v \in C(u).$$

   Again, P_u holds for f_r.

3. For the bounded variance property, we have

$$\begin{aligned} \sum_{v \in C(u)} p_{uv} f_r^2(v) - (\sum_{v \in C(u)} p_{uv} f_r(v))^2 &= \sum_{v \in C(u)} p_{uv} f^2(r) - (\sum_{v \in C(u)} p_{uv} f(r))^2 \\ &= f^2(r) - f^2(r) \\ &= 0 \\ &\leq \sigma_i^2. \end{aligned}$$

4. For the second bounded variance property, we have

$$f_r(u) = f(r) \geq 0.$$

$$\begin{aligned}\sum_{v\in C(u)} p_{uv} f_r^2(v) - (\sum_{v\in C(u)} p_{uv} f_r(v))^2 &= \sum_{v\in C(u)} p_{uv} f^2(r) - (\sum_{v\in C(u)} p_{uv} f(r))^2 \\ &= f^2(r) - f^2(r) \\ &= 0 \\ &\leq \sigma_i^2 + \phi_i f_r(u).\end{aligned}$$

5. For upper-bounded property, we have

$$\begin{aligned}f_r(v) - \sum_{v\in C(u)} p_{uv} f_r(v) &= f(r) - \sum_{v\in C(u)} p_{uv} f(r) \\ &= f(r) - f(r) \\ &= 0 \\ &\leq a_i + M.\end{aligned}$$

   for any child v of u.

6. For the lower-bounded property, we have

$$\begin{aligned}\sum_{v\in C(u)} p_{uv} f_r(v) - f_r(v) &= \sum_{v\in C(u)} p_{uv} f(r) - f(r) \\ &= f(r) - f(r) \\ &= 0 \\ &\leq a_i + M,\end{aligned}$$

   for any child v of u.

Therefore, P_v holds for f_r and any vertex v. □.

For two admissible conditions P and Q, we define PQ to be the property, which is only true when both P and Q are true. If both admissible conditions P and Q are invariant under subtree-unification, then PQ is also invariant under subtree-unification.

For any vertex u of the tree T, an ancestor of u is a vertex lying on the unique path from the root to u. For an admissible condition P, the associated *bad* set B_i over X_i's is defined to be

$$B_i = \{v| \text{ the depth of } v \text{ is } i, \text{ and } P_u \text{ does not hold for some ancestor } u \text{ of } v\}.$$

Lemma 1 *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose each random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. For any admissible condition P, let B_i be the associated bad set of P over X_i. There are random variables $Y_0, \ldots, Y_n$ satisfying:

1. *Y_i is $\mathcal{F}_i$-measurable.*

2. *$Y_0, \ldots, Y_n$ satisfy condition P.*

3. *$\{x : Y_i(x) \neq X_i(x)\} \subset B_i$, for $0 \leq i \leq n$.*

Proof: We modify f and define f' on T as follows. For any vertex u,

$$f'(u) = \begin{cases} f(u) & \text{if } f \text{ satisfies } P_v \text{ for every ancestor } v \text{ of } u \text{ including } u \text{ itself,} \\ f(v) & v \text{ is the ancestor with smallest depth so that } f \text{ fails } P_v. \end{cases}$$

Let S be the set of vertices u satisfying

- f fails P_u,
- f satisfies P_v for every ancestor v of u.

It is clear that f' can be obtained from f by a sequence of subtree-unifications, where S is the set of the roots of subtrees. Furthermore, the order of subtree-unifications does not matter. Since P is invariant under subtree-unifications, the number of vertices that P fails decreases. Now we will show f' satisfies P.

Suppose to the contrary that f' fails P_u for some vertex u. Since P is invariant under subtree-unifications, f also fails P_u. By the definition, there is an ancestor v (of u) in S. After the subtree-unification on subtree rooted at v, P_u is satisfied. This is a contradiction.

Let $Y_0, Y_1, \ldots, Y_n$ be the random variables corresponding to the labeling f'. Y_i's satisfy the desired properties in (1)-(3). □

The following theorem generalizes Azuma's inequality. A similar but more restricted version can be found in [19].

Theorem 32 *For a filter* **F**

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose the random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let $B = B_n$ denote the bad set associated with the following admissible condition:

$$\begin{aligned} \mathrm{E}(X_i|\mathcal{F}_{i-1}) &= X_{i-1} \\ |X_i - X_{i-1}| &\leq c_i \end{aligned}$$

for $1 \leq i \leq n$ where $c_1, c_2, \ldots, c_n$ are non-negative numbers. Then we have

$$\Pr(|X_n - X_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}} + \Pr(B),$$

Proof: We use Lemma 1 which gives random variables $Y_0, Y_1, \ldots, Y_n$ satisfying properties (1)-(3) in the statement of Lemma 1. Then it satisfies

$$\begin{aligned} \mathrm{E}(Y_i|\mathcal{F}_{i-1}) &= Y_{i-1} \\ |Y_i - Y_{i-1}| &\leq c_i. \end{aligned}$$

In other words, $Y_0, \ldots, Y_n$ form a martingale which is $(c_1, \ldots, c_n)$-Lipschitz. By Azuma's inequality, we have

$$\Pr(|Y_n - Y_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}}.$$

Since $Y_0 = X_0$ and $\{x : Y_n(x) \neq X_n(x)\} \subset B_n = B$, we have

$$\begin{aligned} \Pr(|X_n - X_0| \geq \lambda) &\leq \Pr(|Y_n - Y_0| \geq \lambda) + \Pr(X_n \neq Y_n) \\ &\leq 2e^{-\frac{\lambda^2}{2\sum_{i=1}^n c_i^2}} + \Pr(B). \end{aligned}$$

□

For $\mathbf{c} = (c_1, c_2, \ldots, c_n)$ a vector with positive entries, a martingale is said to be near-$\mathbf{c}$-Lipschitz with an exceptional probability η if

$$\sum_i \Pr(|X_i - X_{i-1}| \geq c_i) \leq \eta. \tag{6}$$

Theorem 32 can be restated as follows:

Theorem 33 *For non-negative values, $c_1, c_2, \ldots, c_n$, suppose a martingale X is near-$\mathbf{c}$-Lipschitz with an exceptional probability η. Then X satisfies*

$$\Pr(|X - \mathrm{E}(X)| < a) \leq 2e^{-\frac{a^2}{2\sum_{i=1}^n c_i^2}} + \eta.$$

Now, we can apply the same technique to relax all the theorems in the previous sections.

Here are the relaxed versions of Theorems 20, 25, and 27.

Theorem 34 *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 \\ X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) &\leq a_i + M \end{aligned}$$

for some non-negative constants σ_i and a_i. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2 + a_i^2) + M\lambda/3)}} + \Pr(B).$$

Theorem 35 *For a filter* $\mathbf{F}$

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a non-negative random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \phi_i X_{i-1} \\ X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) &\leq M \end{aligned}$$

for some non-negative constants ϕ_i and M. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2((X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B).$$

Theorem 36 *For a filter* **F**

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a non-negative random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\leq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 + \phi_i X_{i-1} \\ X_i - \mathrm{E}(X_i|\mathcal{F}_{i-1}) &\leq a_i + M \end{aligned}$$

for some non-negative constants σ_i, ϕ_i, a_i and M. Then we have

$$\Pr(X_n \geq X_0 + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2+a_i^2)+(X_0+\lambda)(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B).$$

For supermartingales, we have the following relaxed versions of Theorem 22, 26, and 29.

Theorem 37 *For a filter* **F**

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\geq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 \\ \mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i &\leq a_i + M \end{aligned}$$

for some non-negative constants σ_i, a_i and M. Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2+a_i^2)+M\lambda/3)}} + \Pr(B).$$

Theorem 38 *For a filter* **F**

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\geq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \phi_i X_{i-1} \\ \mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i &\leq M \end{aligned}$$

for some non-negative constants ϕ_i and M. Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{2(X_0(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B).$$

for all $\lambda \leq X_0$.

Theorem 39 *For a filter* **F**

$$\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n = \mathcal{F},$$

suppose a non-negative random variable X_i is $\mathcal{F}_i$-measurable, for $0 \leq i \leq n$. Let B be the bad set associated with the following admissible conditions:

$$\begin{aligned} \mathrm{E}(X_i \mid \mathcal{F}_{i-1}) &\geq X_{i-1} \\ \mathrm{Var}(X_i|\mathcal{F}_{i-1}) &\leq \sigma_i^2 + \phi_i X_{i-1} \\ \mathrm{E}(X_i|\mathcal{F}_{i-1}) - X_i &\leq a_i + M \end{aligned}$$

for some non-negative constants σ_i,ϕ_i, a_i and M. Then we have

$$\Pr(X_n \leq X_0 - \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^n (\sigma_i^2+a_i^2)+X_0(\sum_{i=1}^n \phi_i)+M\lambda/3)}} + \Pr(B),$$

for $\lambda < X_0$.

9 A generalized Polya's urn problem

To see the powerful effect of the concentration and martingale inequalities in the previous sections, the best way is to check out some interesting applications. In the this section we give the probabilistic analysis of the following process involving balls and bins:

> For a fixed $0 \leq p < 1$ and a positive integer $\kappa > 1$, begin with κ bins, each containing one ball and then introduce balls one at a time. For each new ball, with probability p, create a new bin and place the ball in that bin; otherwise, place the ball in an existing bin, such that the probability the ball is placed in a bin is proportional to the number of balls in that bin.

Polya's urn problem (see [18]) is a special case of the above process with $p = 0$ so new bins are never created. For the case of $p > 0$, this infinite Polya

process has a similar flavor as the *preferential attachment* scheme, one of the main models for generating the webgraph among other information networks (see Barabási et al [4, 6]).

In Subsection 9.1, we will show that the infinite Polya process generates a power law distribution so that the expected fraction of bins having k balls is asymptotic to $ck^{-\beta}$, where $\beta = 1 + 1/(1-p)$ and c is a constant. Then the concentration result on the probabilistic error estimates for the power law distribution will be given in Subsection 9.2.

9.1 The expected number of bins with k balls

To analyze the infinite Polya process, we let n_t denote the number of bins at time t and let e_t denote the number of balls at time t. We have

$$e_t = t + \kappa.$$

The number of bins n_t, however, is a sum of t random indicator variables,

$$n_t = \kappa + \sum_{i=1}^{t} s_t$$

where

$$\begin{aligned} \Pr(s_j = 1) &= p, \\ \Pr(s_j = 0) &= 1 - p. \end{aligned}$$

It follows that

$$\mathrm{E}(n_t) = \kappa + pt.$$

To get a handle on the actual value of n_t, we use the binomial concentration inequality as described in Theorem 4. Namely,

$$\Pr(|n_t - \mathrm{E}(n_t)| > a) \leq e^{-a^2/(2pt+2a/3)}.$$

Thus, n_t is exponentially concentrated around $\mathrm{E}(n_t)$.

The problem of interest is the distribution of sizes of bins in the infinite Polya process.

Let $m_{k,t}$ denote the number of bins with k balls at time t. First we note that

$$m_{1,0} = \kappa, \quad \text{and} \quad m_{0,k} = 0.$$

We wish to derive the recurrence for the expected value $\mathrm{E}(m_{k,t})$. Note that a bin with k balls at time t could have come from two cases, either it was a bin with k balls at time $t-1$ and no ball was added to it, or it was a bin with $k-1$ balls at time $t-1$ and a new ball was put in. Let $\mathcal{F}_t$ be the σ-algebra generated by all the possible outcomes at time t.

$$\begin{aligned} \mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) &= m_{k,t-1}(1 - \frac{(1-p)k}{t+\kappa}) + m_{k-1,t-1}(\frac{(1-p)(k-1)}{t+\kappa-1}) \\ \mathrm{E}(m_{k,t}) &= \mathrm{E}(m_{k,t-1})(1 - \frac{(1-p)k}{t+\kappa-1}) + \mathrm{E}(m_{k-1,t-1})(\frac{(1-p)(k-1)}{t+\kappa-1} \quad (7) \end{aligned}$$

For $t > 0$ and $k = 1$, we have

$$\mathrm{E}(m_{1,t}|\mathcal{F}_{t-1}) = m_{1,t-1}(1 - \frac{(1-p)}{t+\kappa-1}) + p.$$
$$\mathrm{E}(m_{1,t}) = \mathrm{E}(m_{1,t-1})(1 - \frac{(1-p)}{t+\kappa-1}) + p. \tag{8}$$

To solve this recurrence, we use the following fact (see [12]):

For a sequence $\{a_t\}$ satisfying the recursive relation $a_{t+1} = (1 - \frac{b_t}{t})a_t + c_t$, $\lim_{t\to\infty} \frac{a_t}{t}$ exists and

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}$$

provided that $\lim_{t\to\infty} b_t = b > 0$ and $\lim_{t\to\infty} c_t = c$.

We proceed by induction on k to show that $\lim_{t\to\infty} \mathrm{E}(m_{k,t})/t$ has a limit M_k for each k.

The first case is $k = 1$. In this case, we apply the above fact with $b_t = b = 1 - p$ and $c_t = c = p$ to deduce that $\lim_{t\to\infty} \mathrm{E}(m_{1,t})/t$ exists and

$$M_1 = \lim_{t\to\infty} \frac{\mathrm{E}(m_{1,t})}{t} = \frac{p}{2-p}.$$

Now we assume that $\lim_{t\to\infty} \mathrm{E}(m_{k-1,t})/t$ exists and we apply the fact again with $b_t = b = k(1-p)$ and $c_t = \mathrm{E}(m_{k-1,t-1})(1-p)(k-1)/(t+\kappa-1)$, so $c = M_{k-1}(1-p)(k-1)$. Thus the limit $\lim_{t\to\infty} \mathrm{E}(m_{k,t})/t$ exists and is equal to

$$M_k = M_{k-1}\frac{(1-p)(k-1)}{1+k(1-p)} = M_{k-1}\frac{k-1}{k+\frac{1}{1-p}}. \tag{9}$$

Thus we can write

$$M_k = \frac{p}{2-p}\prod_{j=2}^{k} \frac{j-1}{j+\frac{1}{1-p}} = \frac{p}{2-p}\frac{\Gamma(k)\Gamma(2+\frac{1}{1-p})}{\Gamma(k+1+\frac{1}{1-p})}$$

where $\Gamma(k)$ is the Gamma function.

We wish to show that the distribution of the bin sizes follows a power law with $M_k \propto k^{-\beta}$ (where $\propto$ means "is proportional to") for large k. If $M_k \propto k^{-\beta}$, then

$$\frac{M_k}{M_{k-1}} = \frac{k^{-\beta}}{(k-1)^{-\beta}} = (1-\frac{1}{k})^{\beta} = 1 - \frac{\beta}{k} + O(\frac{1}{k^2}).$$

From (9) we have

$$\frac{M_k}{M_{k-1}} = \frac{k-1}{k+\frac{1}{1-p}} = 1 - \frac{1+\frac{1}{1-p}}{k+\frac{1}{1-p}} = 1 - \frac{1+\frac{1}{1-p}}{k} + O(\frac{1}{k^2}).$$

Thus we have an approximate power-law with

$$\beta = 1 + \frac{1}{1-p} = 2 + \frac{p}{1-p}.$$

9.2 Concentration on the number of bins with k balls

Since the expected value can be quite different from the actual number of bins with k balls at time t, we give a (probabilistic) estimate of the difference.

We will prove the following theorem.

Theorem 40 *For the infinite Polya process, asymptotically almost surely the number of bins with k balls at time t is*

$$M_k(t+\kappa) + O(2\sqrt{k^3(t+\kappa)\ln(t+\kappa)}).$$

Recall $M_1 = \frac{p}{2-p}$ and $M_k = \frac{p}{2-p}\frac{\Gamma(k)\Gamma(2+\frac{1}{1-p})}{\Gamma(k+1+\frac{1}{1-p})} = O(k^{-(2+\frac{p}{1-p})})$, for $k \geq 2$. In other words, almost surely the distribution of the bin sizes for the infinite Polya process follows a power law with the exponent $\beta = 1 + \frac{1}{1-p}$.

Proof: We have shown that

$$\lim_{t\to\infty} \frac{\mathrm{E}(m_{k,t})}{t} = M_k,$$

where M_k is defined recursively in (9). It is sufficient to show $m_{k,t}$ concentrates on the expected value.

We shall prove the following claim.

Claim: For any fixed $k \geq 1$, for any $c > 0$, with probability at least $1 - 2(t+\kappa+1)^{k-1}e^{-c^2}$, we have

$$|m_{k,t} - M_k(t+\kappa)| \leq 2kc\sqrt{t+\kappa}.$$

To see that the claim implies Theorem 40, we choose $c = \sqrt{k\ln(t+\kappa)}$. Note that

$$2(t+\kappa)^{k-1}e^{-c^2} = 2(t+\kappa+1)^{k-1}(t+\kappa)^{-k} = o(1).$$

From the claim, with probability $1 - o(1)$, we have

$$|m_{k,t} - M_k(t+\kappa)| \leq 2\sqrt{k^3(t+\kappa)\ln(t+\kappa)},$$

as desired.

It remains to prove the claim.

Proof of the Claim: We shall prove it by induction on k.

The base case of $k = 1$:

For $k = 1$, from equation (8), we have

$$\begin{aligned}
& \mathrm{E}(m_{1,t} - M_1(t+\kappa)|\mathcal{F}_{t-1}) \\
= \; & \mathrm{E}(m_{1,t}|\mathcal{F}_{t-1}) - M_1(t+\kappa) \\
= \; & m_{1,t-1}(1 - \frac{1-p}{t+\kappa-1}) + p - M_1(t+\kappa-1) - M_1 \\
= \; & (m_{1,t-1} - M_1(t+\kappa-1))(1 - \frac{1-p}{t+\kappa-1}) + p - M_1(1-p) - M_1 \\
= \; & (m_{1,t-1} - M_1(t+\kappa-1))(1 - \frac{1-p}{t+\kappa-1})
\end{aligned}$$

since $p - M_1(1-p) - M_1 = 0$.

Let $X_{1,t} = \frac{m_{1,t} - M_1(t+\kappa)}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}$. We consider the martingale formed by $1 = X_{1,0}, X_{1,1}, \ldots, X_{1,t}$. We have

$$\begin{aligned}
& X_{1,t} - X_{1,t-1} \\
= & \frac{m_{1,t} - M_1(t+\kappa)}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})} - \frac{m_{1,t-1} - M_1(t+\kappa-1)}{\prod_{j=1}^{t-1}(1-\frac{1-p}{j+\kappa-1})} \\
= & \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}[(m_{1,t} - M_1(t+\kappa)) - (m_{1,t-1} - M_1(t+\kappa-1))(1 - \frac{1-p}{t+\kappa-1})] \\
= & \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}[(m_{1,t} - m_{1,t-1}) + \frac{1-p}{t+\kappa-1}(m_{1,t-1} - M_1(t+\kappa-1)) - M_1].
\end{aligned}$$

We note that $|m_{1,t} - m_{1,t-1}| \leq 1$, $m_{1,t-1} \leq t$, and $M_1 = \frac{p}{2-p} < 1$. We have

$$|X_{1,t} - X_{1,t-1}| \leq \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}. \tag{10}$$

Since $|m_{1,t} - m_{1,t-1}| \leq 1$, we have

$$\begin{aligned}
\mathrm{Var}(m_{1,t}|\mathcal{F}_{t-1}) &\leq \mathrm{E}((m_{1,t} - m_{1,t-1})^2|\mathcal{F}_{t-1}) \\
&\leq 1.
\end{aligned}$$

Therefore, we have the following upper bound for $\mathrm{Var}(X_{1,t}|\mathcal{F}_{t-1})$.

$$\begin{aligned}
\mathrm{Var}(X_{1,t}|\mathcal{F}_{t-1}) &= \mathrm{Var}\big((m_{1,t} - M_1(t+\kappa))\frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}\big|\mathcal{F}_{t-1}\big) \\
&= \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})^2}\mathrm{Var}(m_{1,t} - M_1(t+\kappa)|\mathcal{F}_{t-1}) \\
&= \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})^2}\mathrm{Var}(m_{1,t}|\mathcal{F}_{t-1}) \\
&\leq \frac{1}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})^2}.
\end{aligned} \tag{11}$$

We apply Theorem 19 on the martingale $\{X_{1,t}\}$ with $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1-\frac{1-p}{j+\kappa-1})^2}$, $M = \frac{4}{\prod_{j=1}^{t}(1-\frac{1-p}{j+\kappa-1})}$ and $a_i = 0$. We have

$$\Pr(X_{1,t} \geq \mathrm{E}(X_{1,t}) + \lambda) \leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2 + M\lambda/3)}}.$$

Here $\mathrm{E}(X_{1,t}) = X_{1,0} = 1$. We will use the following approximation.

$$\begin{aligned}\prod_{j=1}^{i}(1-\frac{1-p}{j+\kappa-1}) &= \prod_{j=1}^{i}\frac{j+\kappa-2+p}{j+\kappa-1}\\ &= \frac{\Gamma(\kappa)\Gamma(i+\kappa-1+p)}{\Gamma(\kappa-1+p)\Gamma(i+\kappa)}\\ &\approx C(i+\kappa)^{-1+p}\end{aligned}$$

where $C = \frac{\Gamma(\kappa)}{\Gamma(\kappa-1+p)}$ is a constant depending only on p and κ.

For any $c > 0$, we choose $\lambda = \frac{4c\sqrt{t+\kappa}}{\prod_{j=1}^{t}(1-\frac{1-p}{j})} \approx 4C^{-1}ct^{3/2-p}$. We have

$$\begin{aligned}\sum_{i=1}^{t}\sigma_i^2 &= \sum_{i=1}^{t}\frac{4}{\prod_{j=1}^{i}(1-\frac{1-p}{j})^2}\\ &\approx \sum_{i=1}^{t}4C^{-2}(i+\kappa)^{2-2p}\\ &\approx \frac{4C^{-2}}{3-2p}(t+\kappa)^{3-2p}\\ &< 4C^{-2}(t+\kappa)^{3-2p}.\end{aligned}$$

We note that

$$M\lambda/3 \approx \frac{8}{3}C^{-2}ct^{5/2-2p} < 2C^{-2}t^{3-2p}$$

provided $4c/3 < \sqrt{t+\kappa}$. We have

$$\begin{aligned}\Pr(X_{1,t} \geq 1+\lambda) &\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2+M\lambda/3)}}\\ &< e^{-\frac{16C^{-2}c^2t^{3-2p}}{8C^{-2}t^{3-2p}+2C^{-2}(t+\kappa)^{3-2p}}}\\ &< e^{-c^2}.\end{aligned}$$

Since 1 is much smaller than λ, we can replace $1+\lambda$ by 1 without loss of generality. Thus, with probability at least $1-e^{-c^2}$, we have

$$X_{1,t} \leq \lambda.$$

Similarly, with probability at least $1-e^{-c^2}$, we have

$$m_{1,t} - M_1(t+\kappa) \leq 2c\sqrt{t+\kappa}. \tag{12}$$

We remark that inequality (12) holds for any $c > 0$. In fact, it is trivial when $4c/3 > \sqrt{t+\kappa}$ since $|m_{1,t} - M_1(t+\kappa)| \leq 2t$ always holds.

Similarly, by applying Theorem 23 on the martingale, the following lower bound

$$m_{1,t} - M_1(t+\kappa) \geq -2c\sqrt{t+\kappa} \tag{13}$$

holds with probability at least $1 - e^{-c^2}$.

We have proved the claim for $k = 1$.

The inductive step:

Suppose the claim holds for $k-1$. For k, we define

$$X_{k,t} = \frac{m_{k,t} - M_k(t+\kappa) - 2(k-1)c\sqrt{t+\kappa}}{\prod_{j=1}^{t}(1 - \frac{(1-p)k}{j+\kappa-1})}.$$

We have

$$\begin{aligned}
&\mathrm{E}(m_{k,t} - M_k(t+\kappa) - 2(k-1)c\sqrt{t+\kappa}|\mathcal{F}_{t-1}) \\
&\quad = \mathrm{E}(m_{k,t}|\mathcal{F}_{t-1}) - M_k(t+\kappa) - 2(k-1)c\sqrt{t+\kappa} \\
&\quad = m_{k,t-1}(1 - \frac{(1-p)k}{t+\kappa-1}) + m_{k-1,t-1}(\frac{(1-p)(k-1)}{t+\kappa-1}) \\
&\qquad -M_k(t+\kappa) - 2(k-1)c\sqrt{t+\kappa}.
\end{aligned}$$

By the induction hypothesis, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$|m_{k-1,t-1} - M_{k-1}(t+\kappa)| \le 2(k-1)c\sqrt{t+\kappa}.$$

By using this estimate, with probability at least $1 - 2t^{k-2}e^{-c^2}$, we have

$$\begin{aligned}
&\mathrm{E}(m_{k,t} - M_k(t+\kappa) - 2(k-1)c\sqrt{t+\kappa}|\mathcal{F}_{t-1}) \\
\le\ &(1 - \frac{(1-p)k}{t})(m_{k,t-1} - M_k(t+\kappa-1) - 2(k-1)c\sqrt{t+\kappa-1})
\end{aligned}$$

from the fact that $M_k \le M_{k-1}$ as seen in (9).

Therefore, $0 = X_{k,0}, X_{k,1}, \cdots, X_{k,t}$ forms a submartingale with failure probability at most $2t^{k-2}e^{-c^2}$.

Similar to inequalities (10) and (11), it can be easily shown that

$$|X_{k,t} - X_{k,t-1}| \le \frac{4}{\prod_{j=1}^{t}(1 - \frac{(1-p)k}{j+\kappa-1})} \tag{14}$$

and

$$\mathrm{Var}(X_{k,t}|\mathcal{F}_{t-1}) \le \frac{4}{\prod_{j=1}^{t}(1 - \frac{(1-p)k}{j+\kappa-1})^2}.$$

We apply Theorem 35 on the submartingale with $\sigma_i^2 = \frac{4}{\prod_{j=1}^{i}(1-\frac{(1-p)k}{j+\kappa-1})^2}$, $M = \frac{4}{\prod_{j=1}^{t}(1-\frac{(1-p)\kappa}{j+\kappa-1})}$ and $a_i = 0$. We have

$$\Pr(X_{k,t} \ge \mathrm{E}(X_{k,t}) + \lambda) \le e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2 + M\lambda/3)}} + \Pr(B),$$

where $\Pr(B) \le t^{k-1}e^{-c^2}$ by induction hypothesis.

Here $\mathrm{E}(X_{k,t}) = X_{k,0} = 0$. We will use the following approximation.

$$\begin{aligned}\prod_{j=1}^{i}(1-\frac{(1-p)k}{j+\kappa-1}) &= \prod_{j=1}^{i}\frac{j-(1-p)k}{j+\kappa-1}\\ &= \frac{\Gamma(\kappa)}{\Gamma(1-(1-p)k)}\frac{\Gamma(i+1-(1-p)k)}{\Gamma(i+\kappa)}\\ &\approx C_k(i+\kappa)^{-(1-p)k}\end{aligned}$$

where $C_k = \frac{\Gamma(\kappa)}{\Gamma(1-(1-p)k)}$ is a constant depending only on k, p and κ.

For any $c > 0$, we choose $\lambda = \frac{4c\sqrt{t+\kappa}}{\prod_{j=1}^{t}(1-\frac{(1-p)k}{j})} \approx 4C_k^{-1}ct^{3/2-p}$. We have

$$\begin{aligned}\sum_{i=1}^{t}\sigma_i^2 &= \sum_{i=1}^{t}\frac{4}{\prod_{j=1}^{i}(1-\frac{(1-p)k}{j})^2}\\ &\approx \sum_{i=1}^{t}4C_k^{-2}(i+\kappa)^{2k(1-p)}\\ &\approx \frac{4C_k^{-2}}{1+2k(1-p)}(t+\kappa)^{1+2k(1-p)}\\ &< 4C_k^{-2}(t+\kappa)^{1+2k(1-p)}.\end{aligned}$$

We note that

$$M\lambda/3 \approx \frac{8}{3}C_k^{-2}c(t+\kappa)^{1/2+2(1-p)} < 2C_k^{-2}(t+\kappa)^{1+2(1-p)}$$

provided $4c/3 < \sqrt{t+\kappa}$. We have

$$\begin{aligned}\Pr(X_{k,t}\geq\lambda) &\leq e^{-\frac{\lambda^2}{2(\sum_{i=1}^{t}\sigma_i^2+M\lambda/3)}}+\Pr(B)\\ &< e^{-\frac{16C_k^{-2}c^2(t+\kappa)^{1+2k(1-p)}}{8C_k^{-2}(t+\kappa)^{1+2(1-p)}+2C_k^{-2}(t+\kappa)^{1+2(1-p)}}}+\Pr(B)\\ &< e^{-c^2}+(t+\kappa)^{k-1}e^{-c^2}\\ &\leq (t+\kappa+1)^{k-1}e^{-c^2}.\end{aligned}$$

With probability at least $1-(t+\kappa+1)^{k-1}e^{-c^2}$, we have

$$X_{k,t} \leq \lambda.$$

Equivalently, with probability at least $1-(t+\kappa+1)^{k-1}e^{-c^2}$, we have

$$m_{k,t} - M_k(t+\kappa) \leq 2ck\sqrt{t+\kappa}. \tag{15}$$

We remark that inequality (15) holds for any $c > 0$. In fact, it is trivial when $4c/3 > \sqrt{t+\kappa}$ since $|m_{k,t} - M_k(t+\kappa)| \leq 2(t+\kappa)$ always holds.

To obtain the lower bound, we consider

$$X'_{k,t} = \frac{m_{k,t} - M_k(t+\kappa) + 2(k-1)c\sqrt{t+\kappa}}{\prod_{j=1}^{t}(1 - \frac{(1-p)k}{j+\kappa-1})}.$$

It can be easily shown that $X'_{k,t}$ is nearly a supermartingale. Similarly, if applying Theorem 38 to $X'_{k,t}$, the following lower bound

$$m_{k,t} - M_k(t+\kappa) \geq -2kc\sqrt{t+\kappa} \tag{16}$$

holds with probability at least $1 - (t+\kappa+1)^{k-1}e^{-c^2}$.

Together these prove the statement for k. The proof of Theorem 40 is complete. □

The above methods for proving concentration of the power law distribution for the infinite Polya process can be easily carried out for many other problems. One of the most popular models for generating random graphs (which simulate webgraphs and various information networks) is the so-called *preferential attachment scheme*. The problem on the degree distribution of the preferential attachment scheme can be viewed as a variation of the Polya process as we will see. Before we proceed, we first give a short description for the preferential attachment scheme [3, 21]:

- With probability p, for some fixed p, add a new vertex v, and add an edge $\{u, v\}$ from v by randomly and independently choosing u in proportion to the degree of u in the current graph. The initial graph, say, is one single vertex with a loop.
- Otherwise, add a new edge $\{r, s\}$ by independently choosing vertices r and s with probability proportional to their degrees. Here r and s could be the same vertex.

The above preferential attachment scheme can be rewritten as the following variation of the Polya process:

- Start with one bin containing one ball.
- At each step, with probability p, add two balls, one to a new bin and one to an existing bin with probability proportional to the bin size. with probability $1 - p$, add two balls, each of which is independently placed to an existing bin with probability proportional to the bin size.

As we can see, the bins are the vertices; at each time step the bins that the two balls are placed are associated with an edge; the bin size is exactly the degree of the vertex.

It is not difficult to show the expected degrees of the preferential attachment model satisfy a power law distribution with exponent $1 + 2/(2-p)$ (see [3, 21]). The concentration results for the power law degree distribution of the preferential attachment scheme can be proved in a very similar way as what we have done in this section for the Polya process. The details of the proof can be found in a forthcoming book [13].

References

[1] J. Abello, A. Buchsbaum, and J. Westbrook, A functional approach to external graph algorithms, *Proc. 6th European Symposium on Algorithms,* pp. 332–343, 1998.

[2] W. Aiello, F. Chung and L. Lu, A random graph model for massive graphs, *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, (2000) 171-180.

[3] W. Aiello, F. Chung and L. Lu, Random evolution in massive graphs, Extended abstract appeared in *The 42th Annual Symposium on Foundation of Computer Sciences*, October, 2001. Paper version appeared in *Handbook of Massive Data Sets*, (Eds. J. Abello, et. al.), Kluwer Academic Publishers (2002), 97-122.

[4] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Review of Modern Physics* **74** (2002), 47-97.

[5] N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley and Sons, New York, 1992.

[6] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999) 509-512.

[7] N. Alon, J.-H. Kim and J. H. Spencer, Nearly perfect matchings in regular simple hypergraphs, *Israel J. Math.* **100** (1997), 171-187.

[8] H. Chernoff, A note on an inequality involving the normal distribution, *Ann. Probab.*, **9**, (1981), 533-535.

[9] F. Chung and L. Lu, Connected components in random graphs with given expected degree sequences, *Annals of Combinatorics* **6**, (2002), 125–145.

[10] F. Chung and L. Lu, The average distances in random graphs with given expected degrees, *Proceeding of National Academy of Science* **99**, (2002), 15879–15882.

[11] F. Chung, L. Lu and V. Vu, The spectra of random graphs with given expected degrees, *Proceedings of National Academy of Sciences*, **100**, no. 11, (2003), 6313-6318.

[12] F. Chung and L. Lu, Coupling online and offline analyses for random power law graphs, *Internet Mathematics*, **1** (2004), 409-461.

[13] F. Chung and L. Lu, *Complex Graphs and Networks*, CBMS Lecture Notes, in preparation.

[14] F. Chung, S. Handjani and D. Jungreis, Generalizations of Polya's urn problem, *Annals of Combinatorics*, **7**, (2003), 141-153.

[15] W. Feller, Martingales, *An Introduction to Probability Theory and its Applications*, Vol. 2, New York, Wiley, 1971.

[16] R. L. Graham, D. E. Knuth and O. Patashnik, *Concrete Mathematics*, Second edition, Addison-Wesley Publishing Company, Reading, MA, 1994.

[17] S. Janson, T. Łuczak, and A. Rucínski, *Random Graphs*, Wiley-Interscience, New York, 2000.

[18] N. Johnson and S. Kotz, *Urn Models and their Applications: An approach to Modern Discrete Probability Theory*, Wiley, New York, 1977.

[19] J. H. Kim and V. Vu, Concentration of multivariate polynomials and its applications, *Combinatorica*, **20** (3) (2000), 417-434.

[20] C. McDiarmid, Concentration, *Probabilistic methods for algorithmic discrete mathematics*, 195–248, *Algorithms Combin.*, 16, Springer, Berlin, 1998.

[21] M. Mitzenmacher, A brief history of generative models for power law and lognormal distribution, *Internet Mathematics*, **1**, (2004), 226-251.

[22] N. C. Wormald, The differential equation method for random processes and greedy algorithms, in *Lectures on Approximation and Randomized Algorithms*, (M. Karonski and H. J. Proemel, eds), pp. 73-155, PWN, Warsaw, 1999.