CHAPTER 2

Old and New Concentration Inequalities

In the study of random graphs or any randomly chosen objects, the "tools of the trade" mainly concern various concentration inequalities and martingale inequalities.

Suppose we wish to predict the outcome of a problem of interest. One reasonable guess is the expected value of the object. However, how can we tell how good the expected value is to the actual outcome of the event? It can be very useful if such a prediction can be accompanied by a guarantee of its accuracy (within a certain error estimate, for example). This is exactly the role that the concentration inequalities play. In fact, analysis can easily go astray without the rigorous control coming from the concentration inequalities.

In our study of random power law graphs, the usual concentration inequalities are simply not enough. The reasons are multi-fold: Due to uneven degree distribution, the error bound of those very large degrees offset the delicate analysis in the sparse part of the graph. Furthermore, our graph is dynamically evolving and therefore the probability space is changing at each tick of the clock. The problems arising in the analysis of random power law graphs provide impetus for improving our technical tools.

Indeed, in the course of our study of general random graphs, we need to use several strengthened versions of concentration inequalities and martingale inequalities. They are interesting in their own right and are useful for many other problems as well.

In the next several sections, we state and prove a number of variations and generalizations of concentration inequalities and martingale inequalities. Many of these will be used in later chapters.

2.1. 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_i satisfies

$$\begin{aligned} \Pr(X_i = 1) &= p, \\ \Pr(X_i = 0) &= 1 - p. \end{aligned} \tag{2.1}$$

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 $N(a, \sigma)$, whose density function is given by

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

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

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

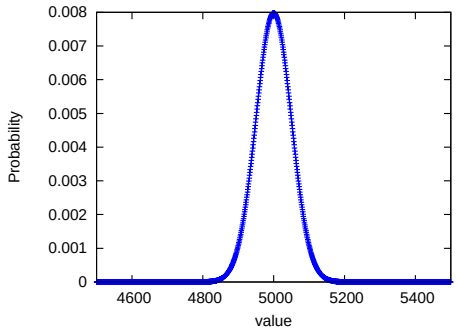

FIGURE 1. The Binomial distribution $B(10000, 0.5)$.

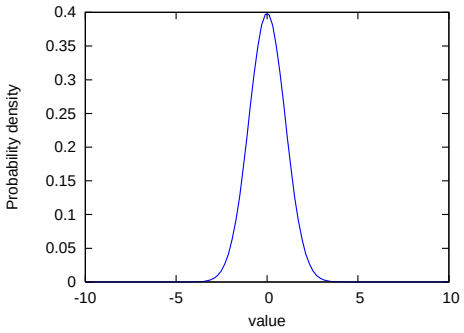

FIGURE 2. The Standard normal distribution $N(0, 1)$.

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 [**53**].

THEOREM 2.1. *The binomial distribution $B(n,p)$ for S_n, as defined in (2.1), 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 $n \to \infty$.

PROOF. We use *Stirling's formula* for $n!$ (see [**70**]).

$$\begin{aligned} n! &= (1+o(1))\sqrt{2\pi n}(\frac{n}{e})^n \\ \text{or, equivalently, } n! &\approx \sqrt{2\pi n}(\frac{n}{e})^n. \end{aligned}$$

For any constant a and b, we have

$$\begin{aligned} & \Pr(a\sigma < S_n - np < b\sigma) \\ = & \sum_{a\sigma<k-np<b\sigma} \binom{n}{k} p^k (1-p)^{n-k} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \frac{n^n}{k^k (n-k)^{n-k}} p^k (1-p)^{n-k} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi np(1-p)}} \big(\frac{np}{k}\big)^{k+1/2} \big(\frac{n(1-p)}{n-k}\big)^{n-k+1/2} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma} \big(1+\frac{k-np}{np}\big)^{-k-1/2} \big(1-\frac{k-np}{n(1-p)}\big)^{-n+k-1/2}. \end{aligned}$$

To approximate the above sum, we consider the following slightly simpler expression. Here, to estimate the lower order term, we use the fact that $k = np + O(\sigma)$ and $1 + x = e^{\ln(1+x)} = e^{x-x^2+O(x^3)}$, for $x = o(1)$. To proceed, we have

$$\begin{aligned} & \Pr(a\sigma < S_n - np < b\sigma) \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma} \big(1+\frac{k-np}{np}\big)^{-k} \big(1-\frac{k-np}{n(1-p)}\big)^{-n+k} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{k(k-np)}{np} + \frac{(n-k)(k-np)}{n(1-p)} + \frac{k(k-np)^2}{n^2p^2} + \frac{(n-k)(k-np)^2}{n^2(1-p)^2} + O(\frac{1}{\sigma})} \\ = & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{k-np}{\sigma})^2 + O(\frac{1}{\sigma})} \\ \approx & \sum_{a\sigma<k-np<b\sigma} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{k-np}{\sigma})^2}. \end{aligned}$$

Now, we set $x = x_k = \frac{k-np}{\sigma}$, and the increment '$dx$'$= x_k - x_{k-1} = 1/\sigma$. Note that $a < x_1 < x_2 < \cdots < b$ forms a $1/\sigma$-net for the interval (a,b). As n approaches infinity, the limit exists. We have

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

Thus, the limit distribution of the normalized binomial distribution is the normal distribution. □

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* $P(\lambda)$:

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

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

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

THEOREM 2.2. *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

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

□

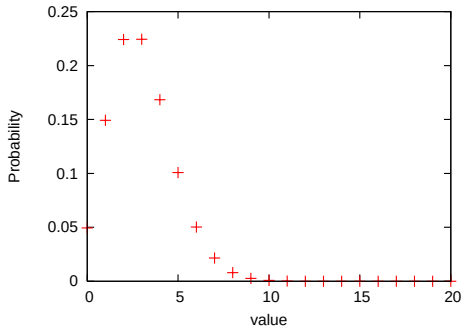

FIGURE 3. The Binomial distribution $B(1000, 0.003)$.

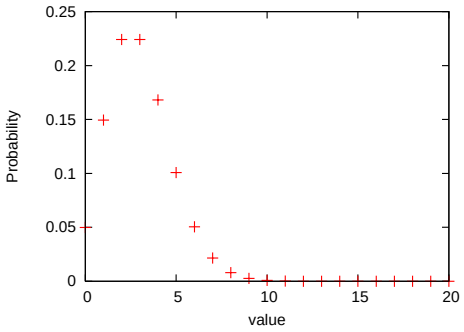

FIGURE 4. The Poisson distribution $P(3)$.

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 2.1 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.

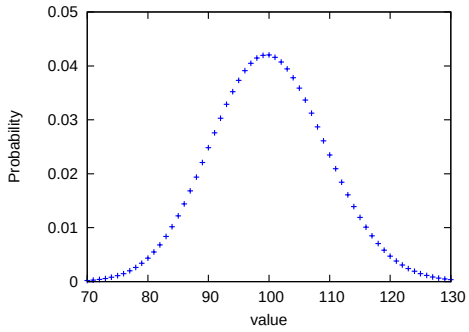

FIGURE 5. The Binomial distribution $B(1000, 0.1)$.

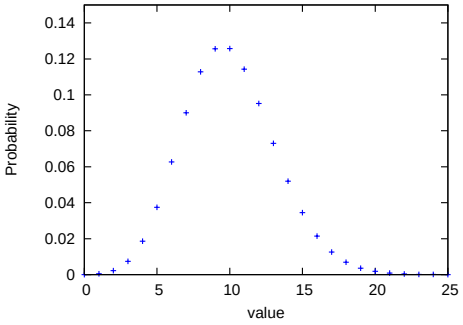

FIGURE 6. The Binomial distribution $B(1000, 0.01)$.

2.2. 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 the 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 2.3. [**28**] *Let $X_1, \ldots, X_n$ be independent random variables such that $\mathrm{E}(X_i) = 0$ and $|X_i| \leq 1$ for all i. Let $X = \sum_{i=1}^n X_i$ and σ^2 be the variance of X_i. Then*

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

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 2.4. [**28**] *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

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

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 2.8 and deduce Theorems 2.6 and 2.5. Theorems 2.10 and 2.11 will be stated and proved in the next section. Theorems 2.9, 2.7, 2.13, and 2.14 on the lower tail can be deduced by reflecting X to $-X$.

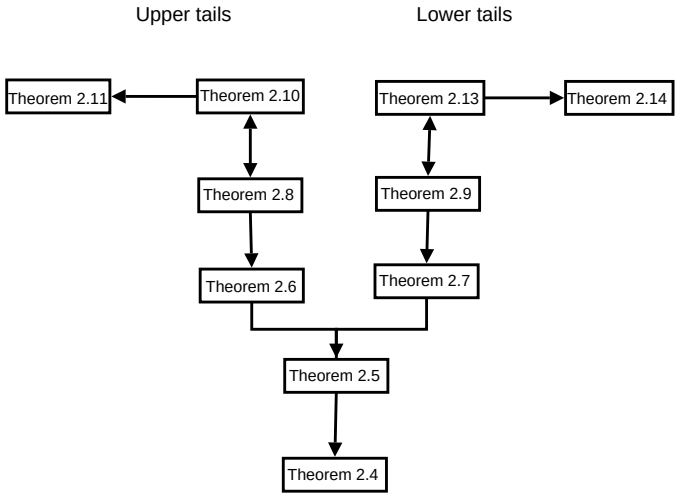

FIGURE 7. The flowchart for theorems on the sum of independent variables.

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

THEOREM 2.5. [**34**] *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

$$\Pr(X \le \mathrm{E}(X) - \lambda) \le e^{-\lambda^2/2\nu} \tag{2.2}$$

$$\Pr(X \ge \mathrm{E}(X) + \lambda) \le e^{-\frac{\lambda^2}{2(\nu + a\lambda/3)}} \tag{2.3}$$

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

To compare inequalities (2.2) to (2.3), 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.

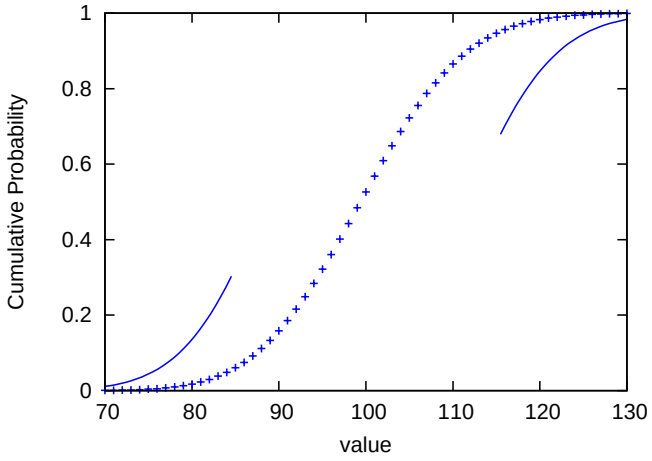

FIGURE 8. Chernoff inequalities.

The inequality (2.3) in the above theorem is a corollary of the following general concentration inequality (also see Theorem 2.7 in the survey paper by McDiarmid [**99**]).

THEOREM 2.6. [**99**] *Let X_i 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 2.7. *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 2.8. *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 2.8, we have the following theorem for the lower tail.

THEOREM 2.9. *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 Theorem 2.8, we will first show the implications of Theorems 2.8 and 2.9. Namely, we will show that the other concentration inequalities can be derived from Theorems 2.8 and 2.9.

Fact: Theorem 2.8 $\Longrightarrow$ Theorem 2.6:

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

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

Applying Theorem 2.8, 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 2.9 $\Longrightarrow$ Theorem 2.7
The proof is straightforward by choosing $M = 0$.

Fact: Theorem 2.6 and 2.7 $\Longrightarrow$ Theorem 2.5

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 (2.2) now follows from Theorem 2.7 since the Y_i's are non-negative.

For the other direction, we have

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

Equation (2.3) now follows from Theorem 2.6. □

Fact: Theorem 2.8 and Theorem 2.9 $\Longrightarrow$ Theorem 2.3

The proof is by choosing $Y = X - \mathrm{E}(X)$, $M = 1$ and applying Theorems 2.8 and 2.9 to Y.

Fact: Theorem 2.5 $\Longrightarrow$ Theorem 2.4

The proof follows by choosing $a_1 = a_2 = \cdots = a_n = 1$.

Finally, we give the complete proof of Theorem 2.8 and thus finish the proofs for all the above theorems on Chernoff inequalities.

Proof of Theorem 2.8: 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:

- $g(0) = 1$.
- $g(y) \leq 1$, 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. □

2.3. More concentration inequalities

Here we state several variations and extensions of the concentration inequalities in Theorem 2.8. We first consider the upper tail.

THEOREM 2.10. *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,

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

By applying Theorem 2.8, the proof is finished. □

THEOREM 2.11. *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 2.10, 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 2.12. 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 2.6 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 2.11 with $M_1 = \cdots = M_{n-1} = (1-p)$ and $M_n = \sqrt{n}(1-p)$. Choosing $k = n-1$, we have

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

Thus,

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

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

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

From the above examples, we note that Theorem 2.11 gives a significantly better bound than that in Theorem 2.6 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 2.13. *Let X_i denote independent random variables satisfying $X_i \ge \mathrm{E}(X_i) - a_i - M$, for $0 \le i \le n$. For $X = \sum_{i=1}^n X_i$, we have*

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

THEOREM 2.14. *Let X_i denote independent random variables satisfying $X_i \ge \mathrm{E}(X_i) - M_i$, for $0 \le i \le n$. 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 $1 \le k \le n$, we have*

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

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

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

Using Theorem 2.14, we have

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

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

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

2.4. A concentration inequality with a large error estimate

In the previous section, we saw that the Chernoff inequality gives very good probabilistic estimates when a random variable is close to its expected value. Suppose we allow the error bound to the expected value to be a positive fraction of the expected value. Then we can obtain even better bounds for the probability of the tails. The following two concentration inequalities can be found in [**105**].

THEOREM 2.15. *Let X be a sum of independent random indicator variables. For any $\epsilon > 0$,*

$$\Pr(X \geq (1+\epsilon)\mathrm{E}(X)) \leq \left[\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right]^{\mathrm{E}(X)}. \tag{2.4}$$

THEOREM 2.16. *Let X be a sum of independent random indicator variables. For any $0 < \epsilon < 1$,*

$$\Pr(X \leq \epsilon\mathrm{E}(X)) \leq e^{-(1-\epsilon)^2\mathrm{E}(X)/2}. \tag{2.5}$$

The above inequalities, however, are still not enough for our applications in Chapter 7. We need the following somewhat stronger concentration inequality for the lower tail.

THEOREM 2.17. *Let X be the sum of independent random indicator variables. For any $0 \leq \epsilon \leq e^{-1}$, we have*

$$\Pr(X \leq \epsilon\mathrm{E}(X)) \leq e^{-(1-2\epsilon(1-\ln\epsilon))\mathrm{E}(X)}. \tag{2.6}$$

PROOF. Suppose that $X = \sum_{i=1}^{n} X_i$, where X_i's are independent random variables with

$$\Pr(X_i = 0) = 1 - p_i \text{ and } \Pr(X_i = 1) = p_i.$$

We have

$$\begin{aligned}
\Pr(X \le \epsilon \mathrm{E}(X)) &= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \Pr(X = k) \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i \prod_{i \notin S} (1 - p_i) \\
&\le \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i \prod_{i \notin S} e^{-p_i} \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\sum_{i \notin S} p_i} \\
&= \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\sum_{i=1}^{n} p_i + \sum_{i \in S} p_i} \\
&\le \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \sum_{|S|=k} \prod_{i \in S} p_i e^{-\mathrm{E}(X) + k} \\
&\le \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} e^{-\mathrm{E}(X)+k} \frac{(\sum_{i=1}^{n} p_i)^k}{k!} \\
&= e^{-\mathrm{E}(X)} \sum_{k=0}^{\lfloor \epsilon \mathrm{E}(X) \rfloor} \frac{(e\mathrm{E}(X))^k}{k!}.
\end{aligned}$$

When $\epsilon \mathrm{E}(X) < 1$, the statement is true since

$$\Pr(X \le \epsilon \mathrm{E}(X)) \le e^{-\mathrm{E}(X)} \le e^{-(1 - 2\epsilon(1 - \ln \epsilon))\mathrm{E}(X)}.$$

Now we consider the case $\epsilon \mathrm{E}(X) \ge 1$.

Note that $g(k) = \frac{(e\mathrm{E}(X))^k}{k!}$ increases when $k < e\mathrm{E}(X)$. Let $k_0 = \lfloor \epsilon \mathrm{E}(X) \rfloor \le \epsilon \mathrm{E}(X)$.

We have

$$\begin{aligned}
\Pr(X \le \epsilon \mathrm{E}(X)) &\le e^{-\mathrm{E}(X)} \sum_{k=0}^{k_0} \frac{(e\mathrm{E}(X))^k}{k!} \\
&\le e^{-\mathrm{E}(X)} (k_0 + 1) \frac{(e\mathrm{E}(X))^{k_0}}{k_0!}.
\end{aligned}$$

By using Stirling's formula

$$n! \approx \sqrt{2\pi n} (\frac{n}{e})^n \ge (\frac{n}{e})^n,$$

we have

$$\begin{aligned}\Pr(X \le \epsilon \mathrm{E}(X)) &\le e^{-\mathrm{E}(X)}(k_0+1)\frac{(e\mathrm{E}(X))^{k_0}}{k_0!} \\ &\le e^{-\mathrm{E}(X)}(k_0+1)(\frac{e^2\mathrm{E}(X)}{k_0})^{k_0} \\ &\le e^{-\mathrm{E}(X)}(\epsilon\mathrm{E}(X)+1)(\frac{e^2}{\epsilon})^{\epsilon\mathrm{E}(X)} \\ &= (\epsilon\mathrm{E}(X)+1)e^{-(1-2\epsilon+\epsilon\ln\epsilon)\mathrm{E}(X)}.\end{aligned}$$

Here we replaced k_0 by $\epsilon\mathrm{E}(X)$ since the function $(x+1)(\frac{e^2\mathrm{E}(X)}{x})^x$ is increasing for $x < e\mathrm{E}(X)$.

To simplify the above expression, we have

$$\mathrm{E}(X) \ge \frac{1}{\epsilon} \ge \frac{1}{1-\epsilon}$$

since $\epsilon\mathrm{E}(X) \ge 1$ and $\epsilon \le e^{-1} \le 1-\epsilon$. Thus, $\epsilon\mathrm{E}(X)+1 \le \mathrm{E}(X)$.

Also, we have $\mathrm{E}(X) \ge \frac{1}{\epsilon} \ge e$. The function $\frac{\ln x}{x}$ is decreasing for $x \ge e$. Thus,

$$\frac{\ln \mathrm{E}(X)}{\mathrm{E}(X)} \le \frac{\ln \frac{1}{\epsilon}}{\frac{1}{\epsilon}} = -\epsilon\ln\epsilon.$$

We have

$$\begin{aligned}\Pr(X \le \epsilon\mathrm{E}(X)) &\le (\epsilon\mathrm{E}(X)+1)e^{-(1-2\epsilon+\epsilon\ln\epsilon)\mathrm{E}(X)} \\ &\le \mathrm{E}(X)e^{-(1-2\epsilon)\mathrm{E}(X)}e^{-\epsilon\ln\epsilon\mathrm{E}(X)} \\ &\le e^{-(1-2\epsilon)\mathrm{E}(X)}e^{-2\epsilon\ln\epsilon\mathrm{E}(X)} \\ &= e^{-(1-2\epsilon(1-\ln\epsilon))\mathrm{E}(X)}.\end{aligned}$$

The proof of Theorem 2.17 is complete. □

2.5. 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 (see [**53**], 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 book 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 [**80**].)

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 2.18. 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, $X_0, X_1, X_2, \ldots, X_n$ form 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

$$|X_i - X_{i-1}| \le c_i \tag{2.7}$$

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

Theorem 2.19 (Azuma's inequality). *If a martingale X is $\mathbf{c}$-Lipschitz, then*

$$\Pr(|X - \mathrm{E}(X)| \ge \lambda) \le 2e^{-\frac{\lambda^2}{2\sum_{i=1}^{n} c_i^2}}, \tag{2.8}$$

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

Theorem 2.20. *Let $X_1, X_2, \ldots, X_n$ be independent random variables satisfying*

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

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

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

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.

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

□

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 Azuma's inequality in such cases. In this section, we will first state and then prove several extensions of Azuma's inequality (see Figure 9).

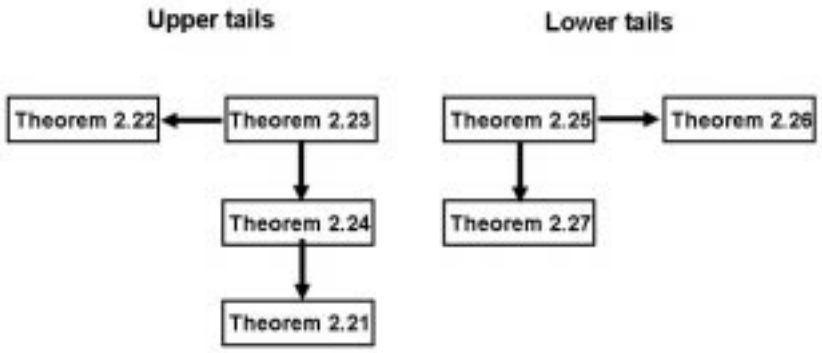

FIGURE 9. The flowchart for theorems on martingales.

Our starting point is the following well known concentration inequality (see [**99**]):

THEOREM 2.21. *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) $X_i - X_{i-1} \leq M$, *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 2.18), Theorem 2.21 implies Theorem 2.6. In a similar way, the following theorem is associated with Theorem 2.10.

THEOREM 2.22. *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) $X_i - X_{i-1} \leq M_i$, *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 2.23. *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) $X_i - X_{i-1} \leq a_i + M$, *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 2.23 implies Theorem 2.21 by choosing $a_1 = a_2 = \cdots = a_n = 0$.

We also have the following theorem corresponding to Theorem 2.11.

THEOREM 2.24. *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) $X_i - X_{i-1} \leq M_i$, *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 2.23 implies Theorem 2.24 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 2.23 so that all the above stated theorems hold.

Proof of Theorem 2.23:

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

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

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 2.25. *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 2.26. *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 2.27. *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

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

2.7. 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 $\mathrm{E}(X_i \mid \mathcal{F}_{i-1}) \geq X_{i-1}$, 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 $\mathrm{E}(X_i \mid \mathcal{F}_{i-1}) \leq X_{i-1}$, 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 2.32 for submartingales and in Theorem 2.30 for supermartingales. Furthermore, we will show that all the stated theorems follow from

Theorems 2.32 and 2.30 (See Figure 10). Note that the inequalities for submartingales and supermartingales are not quite symmetric.

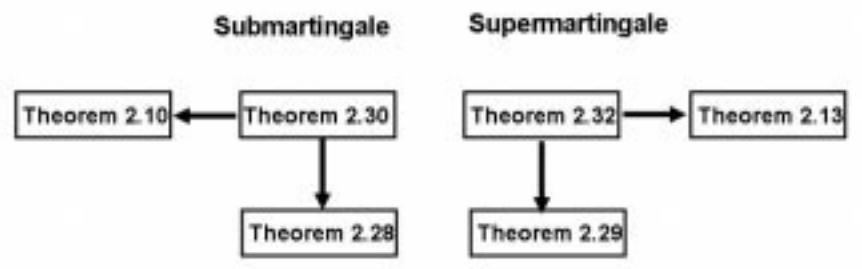

FIGURE 10. The flowchart for theorems on submartingales and supermartingales.

THEOREM 2.28. *Suppose that a supermartingale X, associated with a filter $\mathbf{F}$, satisfies*

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

and

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

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

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

THEOREM 2.29. *Suppose that a submartingale X, associated with a filter $\mathbf{F}$, satisfies, for $1 \le i \le n$,*

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

and

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

Then we have

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

for any $\lambda \le X_0$.

THEOREM 2.30. *Suppose that a supermartingale 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 2.31. *Theorem 2.30 implies Theorem 2.28 by setting all σ_i's and a_i's to zero. Theorem 2.30 also implies Theorem 2.23 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

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

THEOREM 2.32. *Suppose a submartingale 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 $\lambda \le 2X_0 + \frac{\sum_{i=1}^n(\sigma_i^2+a_i^2)}{\sum_{i=1}^n\phi}$.

REMARK 2.33. *Theorem 2.32 implies Theorem 2.29 by setting all σ_i's and a_i's to zero. Theorem 2.32 also implies Theorem 2.25 by choosing $\phi_1 = \cdots = \phi_n = 0$.*

Proof of Theorem 2.30:

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

$$t_{i-1} = t_i + \frac{g(t_0 M)}{2}\phi_i t_i^2,$$

while t_0 will be chosen later. Then

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

and

$$\begin{aligned}
\mathrm{E}(e^{t_i X_i}|\mathcal{F}_{i-1}) &\leq e^{(t_i + \frac{g(t_i M)}{2}\phi_i t_i^2)X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\
&\leq e^{(t_i + \frac{g(t_0 M)}{2} t_i^2 \phi_i)X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)} \\
&= e^{t_{i-1} X_{i-1}} e^{\frac{t_i^2}{2} g(t_i M)(\sigma_i^2 + a_i^2)}
\end{aligned}$$

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

By Markov's inequality, we have

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

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 2.32:

The proof is quite similar to that of Theorem 2.30. 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

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

while t_n will be defined later. Then we have

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

and

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

By Markov's inequality, we have

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

We note

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

Thus, we have

$$\begin{aligned}\Pr(X_n \le X_0 - \lambda) &\le e^{t_n(X_0-\lambda)-t_0X_0+\frac{t_n^2}{2}g(t_nM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\ &\le e^{t_n(X_0-\lambda)-(t_n-\frac{g(t_nM)}{2}t_n^2)X_0+\frac{t_n^2}{2}g(t_nM)\sum_{i=1}^n(\sigma_i^2+a_i^2)} \\ &= e^{-t_n\lambda+\frac{g(t_nM)}{2}t_n^2(\sum_{i=1}^n(\sigma_i^2+a_i^2)+(\sum_{i=1}^n\phi_i)X_0)}.\end{aligned}$$

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 $t_nM < 3$ and

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

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

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

The proof of the theorem is complete. □

2.8. 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 [**81**] 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 [**4**]. Wormald [**124**] 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

$$\sum_v p_{uv} = 1.$$

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 $\pi_i \colon \Omega \to \Omega_i$ 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) **Submartingale:** For $1 \le i \le n$, we have

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

Thus the admissible condition P_u holds if

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

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

(2) **Supermartingale:** For $1 \le i \le n$, we have

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

In this case, the admissible condition of the submartingale is

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

(3) **Martingale:** For $1 \le i \le n$, 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 $1 \le i \le n$, we have

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

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-bounded:** 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-bounded:** 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:

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

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 2.34. *The eight properties as stated in the preceding examples — submartingale, supermartingale, martingale,* **c***-Lipschitz, bounded variance, general bounded variance, upper-bounded, and lower-bounded — 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

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

Again, P_u holds for f_r.

(3) For the bounded variance property, we have

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

(4) For the general bounded variance property, we have

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

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

(5) For the 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 2.35. *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_j *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 the 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'. Then the Y_i's satisfy the desired properties. □

The following theorem generalizes Azuma's inequality. A similar but more restricted version can be found in [**81**].

THEOREM 2.36. *For a filter* $\mathbf{F}$

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

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

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

where $c_1, c_2, \ldots, c_n$ are non-negative numbers. Let $B = \cup_i^n B_i$ denote the union of all bad sets. 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 2.35 which gives random variables $Y_0, Y_1, \ldots, Y_n$ satisfying properties (1)-(3) in the statement of Lemma 2.35. 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 $\{x : Y_n(x) \neq X_n(x)\} \subset \cup_{i=1^n} B_i = B$, 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

$$\sum_i \Pr(|X_i - X_{i-1}| \geq c_i) \leq \eta. \tag{2.9}$$

Theorem 2.36 can be restated as follows:

THEOREM 2.37. *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*

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

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

Here are the relaxed versions of Theorems 2.23, 2.28, and 2.30.

THEOREM 2.38. *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_i be the bad set associated with the following admissible conditions:

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

where σ_i, a_i and M are non-negative constants. Let $B = \cup_{i=1}^n B_i$ be the union of all bad sets. 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 2.39. *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_i be the bad set associated with the following admissible conditions:

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

where ϕ_i and M are non-negative constants. Let $B = \cup_{i=1}^n B_i$ be the union of all bad sets. 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 2.40. *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_i be the bad set associated with the following admissible conditions:

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

where σ_i, ϕ_i, a_i and M are non-negative constants. Let $B = \cup_{i=1}^n B_i$ be the union of all bad sets. 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 submartingales, we have the following relaxed versions of Theorems 2.25, 2.29, and 2.32.

THEOREM 2.41. *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_i *be the bad set associated with the following admissible conditions:*

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

where σ_i, a_i *and* M *are non-negative constants. Let* $B = \cup_{i=1}^n B_i$ *be the union of all bad sets. 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 2.42. *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_i *be the bad set associated with the following admissible conditions:*

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

where ϕ_i *and* M *are non-negative constants. Let* $B = \cup_{i=1}^n B_i$ *be the union of all bad sets. 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 2.43. *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_i *be the bad set associated with the following admissible conditions:*

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

where σ_i, ϕ_i, a_i *and* M *are non-negative constants. Let* $B = \cup_{i=1}^n B_i$ *be the union of all bad sets. 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 $\lambda < X_0$*.*

The best way to see the powerful effect of the concentration and martingale inequalities, as stated in this chapter, is to check out many interesting applications. Indeed, the inequalities here are especially useful for estimating the error bounds in the random graphs that we shall discuss in subsequent chapters. The applications for random graphs of the off-line models are easier than those for the on-line models. The concentration results in Chapter 3 (for the preferential attachment scheme) and Chapter 4 (for the duplication model) are all quite complicated. For a beginner, a good place to start is Chapter 5 on classical random graphs of the Erdős-Rényi model and the generalization of random graph models with given expected degrees. An earlier version of this chapter has appeared as a survey paper [**36**] and includes some further applications.