7
Martingales and Strong Concentration Inequalities

7.1 Concentration Beyond Sums

We are familiar with concentration bounds for sums of independent random variables. Given r.v.s $X_1, X_2, \ldots, X_n$, Chernoff-Hoeffding bounds provide concentration for their sums. For example, if X_i are bounded and independent, then

$$\Pr\left(\left|\sum X_i - E\left[\sum X_i\right]\right| \geq \lambda\right) \leq 2\exp\left(-\frac{\lambda^2}{\ldots}\right).$$

But what if we have some function $f(\vec{X}) = f(X_1, \ldots, X_n)$ and we want to understand if $f(\vec{X})$ is close to its expectation $E[f(\vec{X})]$? Classic Chernoff is not the tool here if $f \neq \sum X_i$.

Example 7.1. Consider the random graph $G(n, p)$. Every edge is chosen independently at random with probability p. Let $f(\vec{X})$ be the size of the maximum matching in $G \sim G(n, p)$. (Here $\vec{X}$ represents the independent r.v.s corresponding to the edges).

We want to say: $\Pr(|f - E[f]| \geq \lambda) \leq$ small.

How can we achieve this? Today, we explore several techniques:

1. Martingales and Azuma-Hoeffding Inequality $\Rightarrow$ Concentration for Lipschitz functions (McDiarmid's Inequality).
2. Talagrand's Inequality $\Rightarrow$ Concentration of certifiable functions.

7.2 Concentration of Lipschitz Functions

Let $f : \mathbb{R}^n \to \mathbb{R}$. We are interested in functions that do not change too much if the input changes slightly.

Definition 7.2 (Lipschitz Condition)**.** Suppose $f(\vec{x})$ is such that

$$f(\vec{x}) \leq f(\vec{y}) + \sum_{i=1}^{n} c_i \cdot \mathbb{I}(x_i \neq y_i).$$

Then we say that f is $\vec{c}$-Lipschitz (with respect to the Hamming metric).

(In general, we want that $|f(\vec{x}) - f(\vec{y})| \leq \alpha \cdot \text{dist}(\vec{x}, \vec{y})$ for some metric. The Lipschitz condition means that a function's value changes little if we change the input by a little bit.)

Theorem 7.3 (Method of Bounded Differences (McDiarmid's Inequality))**.** *Suppose f is $\vec{c}$-Lipschitz for $\vec{c} = (c_1, c_2, \ldots, c_n)$. If $X_1, \ldots, X_n$ are chosen randomly and independently, then*

$$\Pr(|f(\vec{X}) - E[f(\vec{X})]| \geq \lambda) \leq 2 \cdot \exp\left(-\frac{2\lambda^2}{\sum c_i^2}\right).$$

Example 7.4. Let $X_i \in \{-1, +1\}$ (Rademacher variables) and $f(\vec{X}) = \sum X_i$. $E[f] = 0$. If we change one X_i, the sum changes by 2, so $c_i = 2$. $\sum c_i^2 = 4n$. We get concentration for sums of Rademachers:

$$\Pr(|f(\vec{X})| > \lambda) \leq 2 \cdot \exp\left(-\frac{2\lambda^2}{4n}\right) = 2 \cdot \exp\left(-\frac{\lambda^2}{2n}\right).$$

(This recovers Hoeffding's bound.)

7.2.1 *Further Examples*

Example 7.5 (Balls and Bins)**.** n balls, n bins. X_i = bin in which ball i falls (uniform in $\{1, \ldots, n\}$, independent). Let $f(\vec{X})$ = # empty bins. $E[f] = n(1 - 1/n)^n \approx n/e$. The function f is 1-Lipschitz (changing where one ball falls changes the number of empty bins by at most 1). So $c_i = 1$ and $\sum c_i^2 = n$.

$$\Pr(|f - E[f]| > \lambda) \leq 2\exp\left(-\frac{2\lambda^2}{n}\right).$$

Setting $\lambda = O(\sqrt{n \log n})$, we get that # empty bins $= n/e \pm O(\sqrt{n \log n})$ with high probability (w.h.p.).

Example 7.6 (Random Graphs - Edge Exposure)**.** $G \sim G(n, p)$ random graph. We consider the input variables to be the $m = \binom{n}{2}$ potential edges. f = # isolated vertices, or size of max matching, or coloring number $\chi(G)$.

If we change a single edge (present or absent):

- # isolated vertices changes by ≤ 2.
- Size of max matching changes by ≤ 1.
- Coloring # changes by ≤ 1.

In these cases, we can use McDiarmid's Inequality. If $c_i = O(1)$, then $\sum c_i^2 = O(m)$.

$$\Pr(|f - E[f]| \geq \lambda) \leq 2\exp\left(-\frac{2\lambda^2}{O(m)}\right).$$

$\Rightarrow f \in E[f] \pm O(\sqrt{m \log n})$ w.h.p.

Remark 7.7. Note that the above argument applies to any base graph G and we sample a random subset of the edges of G where each edge is picked with probability p. The Erdős-Rényi model is just the special case where G is the complete graph on n vertices. We focus on that in the class for simplicity.

7.2.2 *Limitations and Vertex Exposure*

Hang on! We have $E[f] \leq n$ (size of matching, # isolated vertices, etc.). If $m \approx n^2$, the deviation is $O(n\sqrt{\log n})$. This means we are not very tightly concentrated!

Can we do better by using a different view of the randomness? Yes! (For some cases.)

Suppose we view the edges in "groups". This is often called the **vertex exposure** method, as opposed to the **edge exposure** method used above.

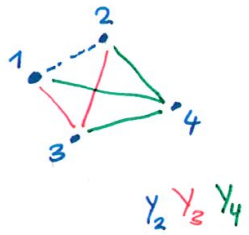


Figure 7.1: Example of the vertex exposure method. Each group corresponds to the edges incident to a vertex.

Group Y_i = edges going from vertex i to vertices $\{1, 2, \ldots, i-1\}$. We have n groups (actually $n-1$) of independent r.v.'s. $f(Y_1, Y_2, \ldots, Y_n)$ is the same function.

Now we check the Lipschitz condition with respect to these groups.

For Matching: If we change one group Y_i (i.e., change all edges incident to vertex i), the max matching size changes by at most 1.

For Coloring χ: If we change all edges incident to vertex i, we may need to choose a new color for i, but the total number of colors changes by at most 1.

Using vertex exposure, we have $\sum c_i^2 = n$. This gives much tighter concentration:

$$f - E[f] \leq O(\sqrt{n \log n}) \quad \text{w.h.p.}$$

Is this still useful? Is $E[f]$ large compared to the deviation $O(\sqrt{n})$? It depends on p. Say $p = 1/2$.

- Max matching size $\approx n/2$. ($G(n,p)$ has a perfect matching w.h.p. if p is large enough).
- Coloring number $\approx \frac{n}{2\log_2 n}$. (Each color class must be an independent set. The largest independent set in $G(n,1/2)$ has size $\approx 2\log_2 n$).

These are much larger than the deviation bounds (at least for $p = 1/2$). For small p, more sophisticated techniques might be needed.

7.3 Martingales and Concentration

To understand how to prove McDiarmid's Inequality, we need the concepts of Martingales and Conditional Expectations.

7.3.1 Review of Conditional Expectation

Let's recap some basic facts about conditional probability. Suppose X, Y are r.v.s on a common sample space Ω. $E[X|Y = y]$ is the average of X over the events where $Y = y$.

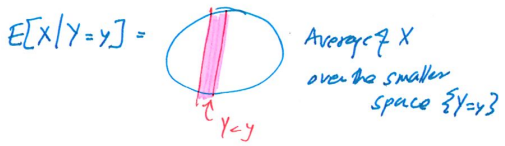


Figure 7.2: Illustration of $E[X|Y = y]$ as the average of X over the part of the sample space where $Y = y$.

$E[X|Y]$ is a function of Y, which takes on some constant value in each part of the partition defined by Y.

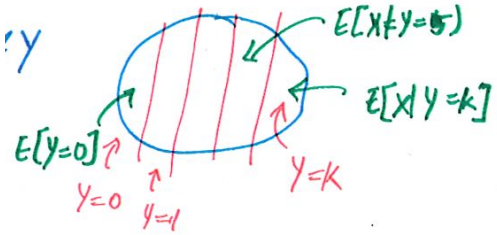


Figure 7.3: Illustration of $E[X|Y]$ as a function of Y.

Fact 7.8 (Tower Property / Law of Total Expectation). $E[X] = \sum_y E[X|Y = y] \cdot \Pr(Y = y) = E_Y[E[X|Y]]$.

More complicated example of the above:

Fact 7.9. $E[X|Y] = E_Z[E[X|Y,Z]|Y]$.

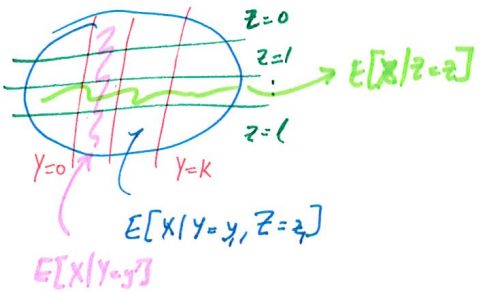


Figure 7.4: Illustration of $E[X|Y] = E_Z[E[X|Y,Z]|Y]$

$E[X]$ is the average over all points in Ω (a constant). $E[X|Y]$ is the average of X over each "partition" for Y (a function of Y). $E[X|Y,Z]$ is the average of X over the finer "grid" defined by Y and Z (a function of Y,Z).

7.3.2 *Martingales*

Martingales model settings where we have correlations between random variables, but we can still prove concentration. They allow us to move beyond merely sums of independent r.v.s.

Definition 7.10 (Martingale)**.** A sequence $(X_k) = X_0, X_1, \ldots$ of r.v.s on a common sample space is a **martingale** if:

1. $E[|X_k|] < \infty$.
2. $E[X_k|X_{k-1}, \ldots, X_0] = X_{k-1}$.

More generally, a sequence (X_k) is a martingale with respect to (w.r.t.) (Z_k) if:

1. $E[|X_k|] < \infty$.
2. X_k is a function of $Z_0, \ldots, Z_k$.
3. $E[X_k|Z_{k-1}, \ldots, Z_0] = X_{k-1}$.

7.3.3 *Doob Martingales*

The most common type of martingale we encounter is the **Doob Martingale**. It formalizes the idea of "exposing" the underlying randomness gradually.

Suppose we have n random variables $Z_1, Z_2, \ldots, Z_n$ over Ω, and a function $f(Z_1, \ldots, Z_n)$. We define the sequence X_t for $t = 0 \ldots n$ as:

$$X_t = E[f(Z_1, \ldots, Z_n)|Z_1, \ldots, Z_t].$$

- $X_0 = E[f]$.

- $X_1 = E[f|Z_1]$.
- ...
- $X_n = E[f|Z_1, \ldots, Z_n] = f(Z_1, \ldots, Z_n)$.

This sequence forms a martingale, representing the process of "exposing the variables one at a time".

Example 7.11. Throw n balls into n bins. Z_i = bin into which i-th ball goes. $f(\vec{Z})$ = number of empty bins.

Example 7.12. $Z_1, \ldots, Z_n$ are n random points in $[0,1]^2$. $f(\vec{Z})$ = length of the shortest tour (TSP) over these points.

7.3.4 Concentration for Martingales (Azuma-Hoeffding)

If the differences between consecutive steps in a martingale are bounded, then the end point is concentrated around the start point.

Theorem 7.13 (Azuma-Hoeffding Inequality)**.** *Suppose (X_k) is a martingale where $|X_k - X_{k-1}| \leq c_k$ for all k. Then*

$$\Pr(|X_n - X_0| > \lambda) \leq 2 \cdot \exp\left(-\frac{\lambda^2}{2\sum c_k^2}\right).$$

Recall that for Doob martingales, $X_0 = E[f]$ and $X_n = f(\vec{Z})$. So Azuma-Hoeffding bounds the probability that the function deviates from its mean:

$$\Pr(|f - E[f]| > \lambda) \leq 2 \cdot \exp\left(-\frac{\lambda^2}{2\sum c_k^2}\right).$$

(Note the difference in the constant in the exponent compared to McDiarmid's inequality).

Exercise 7.14. Prove McDiarmid's Inequality (up to constants) from Azuma-Hoeffding. (Hint: Show that the Lipschitz condition on f implies bounded differences for the corresponding Doob martingale when Z_i are independent).

7.4 Extensions and Other Inequalities

Sometimes we want concentration bounds that are much smaller than $O(\sqrt{n \log n})$, perhaps depending on the mean or variance (more "Chernoff-like").

7.4.1 *Freedman's Inequality*

Freedman's Inequality provides concentration based on the conditional variance. Let (X_k) be a martingale with differences $Y_k = X_k - X_{k-1}$. Suppose $|Y_k| \leq c$. Define the predictable quadratic variation (sum of conditional variances):

$$W_t = \sum_{i=1}^{t} E[Y_i^2 | Z_1, \ldots, Z_{i-1}].$$

Theorem 7.15 (Freedman's Inequality (Simplified))**.** *Then for any $\lambda, \tau > 0$:*

$$\Pr(|X_n - X_0| \geq \lambda \text{ and } W_n \leq \tau^2) \leq 2 \exp\left(-\frac{\lambda^2}{\tau^2 + c\lambda}\right).$$

If we typically have $\tau^2 \ll \sum c_k^2$, this bound is much better.

7.4.2 *Talagrand's Concentration Inequality*

We now consider a special case of another famous inequality: **Talagrand's Concentration Inequality**.

It is useful for functions that are Lipschitz AND have "small certificates" for their value. It can give good concentration even when the expectation is small (e.g., maximum matching in $G(n,p)$ when p is very small).

Let $X_1, \ldots, X_m$ be independent variables. Let f be c-Lipschitz.

Definition 7.16 (h-certifiable)**.** f is ***h*-certifiable** if: whenever $f(x) \geq s$, there exists a subset of coordinates I (a certificate) such that $|I| \leq h(s)$, and $f(y) \geq s$ whenever y agrees with x on the coordinates in I.

Example 7.17 (Max Matching)**.** $f(X_1, \ldots, X_m)$ = max matching in a graph. $X_i \in \{0,1\}$ (edge indicators). If the max matching size is $\geq s$, there is a matching of size s. If we reveal these s edges (variables) and they are present, the max matching size is guaranteed to be $\geq s$. So $h(s) = s$.

Theorem 7.18 (Application of Talagrand's Concentration Inequality)**.** *If f is c-Lipschitz and h-certifiable:*

$$\Pr\left(\left|f - M_f\right| \geq \lambda\right) \leq 4 \exp\left(-\frac{\lambda^2}{4c^2 h(M_f + \lambda)}\right).$$

*Here M_f is the **median** of f. ($\Pr(f \geq M_f) \geq 1/2$ and $\Pr(f \leq M_f) \geq 1/2$).*

This provides "Concentration around the median" rather than around the mean. (If f is bounded and the probability of deviation is small, the mean and median are close).

Application to Matching. For the matching example: $h(s) = s$ and $c = 1$.

$$\Pr(|f - M_f| \geq \lambda) \leq 4 \exp\left(-\frac{\lambda^2}{4(M_f + \lambda)}\right).$$

Let's set $\lambda = 10\sqrt{M_f \log n}$. Assuming M_f is large enough so that $\lambda \ll M_f$.

$$\leq 4 \exp\left(-\frac{100 M_f \log n}{4(M_f + \lambda)}\right) \approx 4 \exp(-O(\log n)) \leq 1/\text{poly}(n).$$

We have tight concentration around the median matching value M_f. The deviation is $\pm O(\sqrt{M_f \log n})$.

7.4.3 Side Note: Median vs Mean

If a function is concentrated around the median, it must also be concentrated around the mean, provided the function is bounded.

Suppose $\Pr(|f - M| \geq \lambda) \leq 1/\text{poly}(n)$ (concentration around median M), and $|f| \leq n$. Then $|M - \mu| \leq \lambda + O(1)$, where $\mu = E[f]$.

Proof Sketch. Suppose $\mu > M + \lambda + c$.

$$\mu = E[f] = E[f|f \geq M + \lambda] \Pr[f \geq M + \lambda] + E[f|f < M + \lambda] \Pr[f < M + \lambda]$$
$$\leq n \cdot \frac{1}{\text{poly}(n)} + (M + \lambda) \cdot 1.$$

This implies $\mu \leq M + \lambda + n/\text{poly}(n)$. This is not possible unless $c \leq n/\text{poly}(n)$.

A similar argument holds for $\mu < M - \lambda - c$.

So $|M - \mu| \leq \lambda + O(1)$. Therefore, $\Pr(|f - \mu| \geq 2\lambda + O(1)) \leq 1/\text{poly}(n)$. □

7.5 Wrap up

We saw:

- **McDiarmid's Inequality** (for Lipschitz functions of independent r.v.s). This is a special case of Talagrand's Concentration Inequality.
- A quick glimpse of **Freedman's Inequality** (which is more "Chernoff-like", depends on the variance/mean).
- **Martingales and Doob Martingales**: important tools when dealing with dependent choices. (Applications mentioned include Fair Matching and next lecture on Discrepancy Algorithms).

We did not get to:

- Stopping times and the Optional Stopping Theorem.
- Wald's Inequality.

(These are useful tools for analyzing expected running times of randomized algorithms).

7.6 Application: Online Edge Coloring (Online Fair Matchings)

In the online edge coloring problem, we are given a max degree Δ and a sequence of edges $e_1, e_2, \ldots, e_m$ on a vertex set V with maximum degree Δ. The edges arrive one by one and we must color each edge as it arrives with a color such that no two adjacent edges share the same color. The goal is to minimize the number of colors used. There has been recent progress on this problem by Blikstad, Svensson, Vintan, and Wajc that settles a conjecture of Bar-Noy, Motwani, and Naor from 1992. They basically show the following:

Theorem 7.19. *There is a randomized online algorithm that uses at most $(1 + o(1))\Delta$ colors in expectation whenever $\Delta = \omega(\sqrt{\log n})$. Moreover, there is a deterministic online algorithm that uses at most $(1 + o(1))\Delta$ colors whenever $\Delta = \omega(\log n)$.*

Both the randomized (and perhaps surpringly the deterministic) algorithm are based on analyzing Martingales, which lead to rather clean and tight algorithms. We explain this connection in a simplified (but very related) setting of online fair matchings. Specifically, given the max degree Δ, we wish to devise a randomized algorithm that maintains a matching M in an online fashion (whenever an edge is presented we need to irrevocably decide whether to include it in our matching or not) such that the following holds:

$$\Pr[e \in M] \geq \frac{1}{\Delta + q} \quad \forall e \in E\,,$$

where q is a small "error" term that is $o(\Delta)$. We can notice that if we have an online edge coloring algorithm that uses at most $\Delta + q$ colors, then it is easy to achieve a fair matching algorithm. Simply output one of the color classes uniformly at random. This shows that the online edge coloring problem is at least as hard as the online fair matching problem. Surprisingly, one can also (when randomization is allowed) go the other direction (up to losing some small factors $o(\log n)$ etc.). In any case, we will focus on the online fair matching problem and illustrate the main ideas there.

7.6.1 The Algorithm

After thinking about this for a while, there is almost only one reasonable algorithm that comes to mind. At each time step t, we will for

each pair of vertices (i.e., potential edge e) maintain a *bias* $Q_t(e)$ that we will use to sample the edge e if it arrives at time t. We will initialize $Q_1(e) = 1/(\Delta + q)$ for all edges e. Indeed, why would we ever want to sample an edge with probability larger than $1/(\Delta + q)$? That would only be detrimental, as it would mean we are over-sampling some edges and decreasing the probability that we can take later arriving edges. These biases will change over time, as we see now.

1. When an edge $e_t = (u_t, v_t)$ arrives at time t, we sample it with probability equal to its current bias $Q_t(e_t)$.
2. If we sample e_t, we add it to our matching M and then zero out the bias for all edges e incident to either u_t or v_t, i.e., set $Q_{t+1}(e) = 0$ for all these edges. Indeed, we cannot select any of these neighboring edges in a matching.
3. On the other hand, if we do not sample e_t, we will keep the matching unchanged and update the neighbors' biases as follows:

$$Q_{t+1}(e) := \frac{Q_t(e)}{1 - Q_t(e_t)} \quad \forall e \text{ incident to } u_t \text{ or } v_t\,.$$

The reason for this update is simple: that we want to "boost" the biases of the edges incident to u_t or v_t since we did not get to sample e_t. Indeed, for an edge e incident to u_t or v_t we have that the probability that we sample that edge if it were to arrive at time $t+1$ to be

$$\mathbb{E}[Q_{t+1}(e) \mid Q_t(e)] = Q_t(e_t) \cdot 0 + (1 - Q_t(e_t))\frac{Q_t(e)}{1 - Q_t(e_t)} = Q_t(e),$$

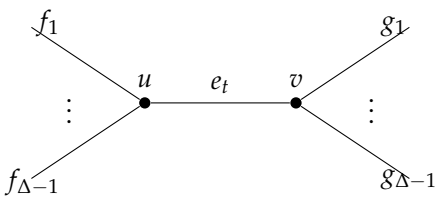
so these biases give us a Martingale!

In other words, that each edge e has the same probability of being included in the matching if it were to arrive at time t or $t+1$, or more generally at any time step. Specifically, each edge e is sampled in the matching with probability exactly $1/(\Delta + q)$.

7.6.2 *The Issue*

However, an issue with the above "algorithm": it is only well-defined if we can ensure that $Q_t(e_t) \leq 1$ for all t and e_t.

To get a feeling for this condition let us consider the following example:



Here we think that the edges arrive in the order

$$f_1, f_2, \ldots, f_{\Delta-1}, g_1, g_2, \ldots, g_{\Delta-1}, e_t.$$

The only chance that we have to pick e_t in the matching is that none of the other edges are selected in the matching. If this is the whole graph, we have that

$$Q_t(e_t) = Q_1(e_t) \cdot \frac{1}{\prod_{i=1}^{\Delta-1}(1 - Q_i(f_i))} \cdot \frac{1}{\prod_{j=1}^{\Delta-1}(1 - Q_{\Delta-1+j}(g_j))},$$

which simplifies to

$$Q_t(e_t) = \frac{1}{\Delta + q} \cdot \frac{1}{S_t(u)} \cdot \frac{1}{S_t(v)},$$

Indeed, observe that $Q_2(f_2) = \frac{Q_1(f_2)}{1-Q_1(f_1)}$ since f_1 was not chosen, and hence $1 - Q_2(f_2) = \frac{1-Q_1(f_1)-Q_1(f_2)}{1-Q_1(f_1)}$, etc.

where

$$S_t(u) = 1 - \sum_{i=1}^{\Delta-1} Q_1(f_i) \quad \text{and} \quad S_t(v) = 1 - \sum_{j=1}^{\Delta-1} Q_1(g_j).$$

Calculations aside, this is rather natural as this is the probability that u is free if we match every f_i with probabiliity $Q_1(f_i)$ and similarly for v. Crucially, note that since each $S_t(u)$ is a linear function of the edge biases, it itself is a Martingale.

We thus have that both $S_t(u)$ and $S_t(v)$ are at least $1 - \frac{\Delta-1}{\Delta+q} \geq q/(2\Delta)$, where we used that $q \leq \Delta$. This gives us that

$$Q_t(e_t) = \frac{1}{(\Delta + q)} \frac{1}{S_t(u)} \frac{1}{S_t(v)} \leq 4\frac{\Delta}{q^2} \leq 1/\sqrt{\Delta}$$

if we set $q = 4\Delta^{3/4}$. So this case is great in that $P(e_t)$ is not only smaller than 1 but actually very small (less than $1/\sqrt{\Delta}$).

However, what if there are edges that arrive before the edges in the figure, incident to the edges f_i and g_i? This may, if not taken, increase the scaling factors $1/S_t(u)$ and $1/S_t(v)$. Moreover, it is very hard to get a control on these $P(\cdot)$ values as they may be arbitrarily correlated.

7.6.3 *Concentration Save The Day*

Martingale concentration allows us to handle this issue and only consider the local 2-hop neighborhood of the edge e_t. Specifically, the scaling factors $S_t(u)$ form a Martingale for each u. So even if $Q_1(f_i)$ was not the original bias of the neighboring edge f_i, but the bias due to all the edges that came before, in expectation it equals $1/(\Delta + q)$.

We can now use the known Martingale concentration inequalities to show that the scaling factors $S_t(u)$ do not deviate too much from its initial value (which is at least $q/(2\Delta)$) over all arriving edges

incident to f_i's. A union bound then shows that $P(e_t) \leq O(\frac{1}{\sqrt{\Delta}})$ with high probability.

In the formal proof, we need to be a bit more careful to bound the step size of the Martingale: this is done by never selecting edges with probability much larger than $1/\sqrt{\Delta}$ (if an edge arrives with probability larger than this, we simply reject it). This then allows us to use Freedman's inequality to conclude that the Martingale does not deviate too much from its initial value.