3 A few more good inequalities, martingale variety

3.1 From independence to martingales ............................ 1
3.2 Hoeffding’s inequality for martingales ........................ 4
3.3 Bennett’s inequality for martingales ............................ 8
3.4 *Concentration of random polynomials ...................... 10
3.5 *Proof of the Kim-Vu inequality ............................... 14
3.6 Problems ......................................................... 18
3.7 Notes ............................................................. 18

Printed: 8 October 2015
Chapter 3

A few more good inequalities, martingale variety

Section 3.1 introduces the method for bounding tail probabilities using moment generating functions.

Section 3.2 discusses the Hoeffding inequality, both for sums of independent bounded random variables and for martingales with bounded increments.

Section 3.3 discusses the Bennett inequality, both for sums of independent random variables that are bounded above by a constant and their martingale analogs.

Section 3.4 presents an extended application of a martingale version of the Bennett inequality to derive a version of the Kim-Vu inequality for polynomials in independent, bounded random variables.

3.1 From independence to martingales

Throughout this Chapter \( \{(S_i, \mathcal{F}_i) : i = 1, \ldots, n\} \) is a martingale on some probability space \((\Omega, \mathcal{F}, \mathbb{P})\). That is, we have a sub-sigma fields \( \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F} \) and integrable, \( \mathcal{F}_i \)-measurable random variables \( S_i \) for which \( \mathbb{P}_{\mathcal{F}_{i-1}} S_i = S_{i-1} \) almost surely. Equivalently, the martingale differences \( \xi_i := S_i - S_{i-1} \) are integrable, \( \mathcal{F}_i \)-measurable, and \( \mathbb{P}_{\mathcal{F}_{i-1}} \xi_i = 0 \) almost surely, for \( i = 1, \ldots, n \). In that case \( S_i = S_0 + \xi_1 + \cdots + \xi_i = S_{i-1} + \xi_i \).

For simplicity I assume \( \mathcal{F}_0 \) is trivial, so that \( S_0 = \mu = \mathbb{P}S_n \). Also, I usually omit the “almost sure” qualifiers, unless there is some good reason to
remind you that the conditional expectations are unique only up to almost sure equivalence. And to avoid a plethora of subscripts, I abbreviate the conditional expectation operator $\mathbb{P}_{\mathcal{F}_{i-1}}$ to $\mathbb{P}_{i-1}$.

The inequalities from Sections 2.5 and 2.6 have martingale analogs, which have proved particularly useful for the development of concentration inequalities. The analog of the Hoeffding inequality is often attributed to Azuma (1967), even though Hoeffding (1963, pages 17–18) had already noted that

"The inequalities of this section can be strengthened in the following way. Let $S_m = X_1 + \cdots + X_m$ for $m = 1, 2, \ldots, n$.

Furthermore, the inequalities of Theorems 1 and 2 remain true if the assumption that $X_1, X_2, \ldots, X_n$ are independent is replaced by the weaker assumption that the sequence $S_m' = S_m - ES_m$, $m = 1, 2, \ldots, n$, is a martingale, that is, \ldots. Indeed, Doob’s inequality (2.17) is true under this assumption. On the other hand, (2.18) implies that the conditional mean of $X_m$ for $S_m' - 1$ fixed is equal to its unconditional mean. A slight modification of the proofs of Theorems 1 and 2 yields the stated result."

In principle, the martingale results can be proved by conditional analogs of the moment generating function technique described in Section 2.2, with just a few precautions to avoid problems with negligible sets. In that Section I started with a random variable $X$ that lives on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. By means of arguments involving differentiation inside expectations I derived a Taylor expansion for the cumulant generating function,

$$L(\theta) = \log \mathbb{E}e^{\theta X} = \theta \mathbb{E}X + {\frac{1}{2}}\theta^2 \text{var}_{\theta^*}(X),$$

where $\theta^*$ lies between 0 and $\theta$ and the variance is calculated under the probability measure $\mathbb{P}_{\theta^*}$ that has density $\exp (\theta^*X - L(\theta^*))$ with respect to $\mathbb{P}$.

The obvious analog for expectations conditional on some sub-sigma-field $\mathcal{G}$ of $\mathcal{F}$ is

$$\log \mathbb{P}_{\mathcal{G}} e^{\theta X} = \theta \mathbb{P}_{\mathcal{G}} X + {\frac{1}{2}}\theta^2 \text{var}_{\mathcal{G}, \theta^*}(X) \quad \text{almost surely.}$$

Here the conditional expectation $\mathbb{P}_{\mathcal{G}} e^{\theta X}$ is defined in the usual Kolmogorov way as the almost surely unique $\mathcal{G}$-measurable random variable $\mathcal{M}_{\omega}(\theta)$ for
3.1 From independence to martingales

which

\[ \mathbb{P} \left( g(\omega) e^{\theta X} \right) = \mathbb{P} \left( g(\omega) M_\omega(\theta) \right) \]

at least for all nonnegative, \( \mathcal{G} \)-measurable random variables \( g \). The random variable \( M_\omega(\theta) \) could be changed on some \( \mathcal{G} \)-measurable negligible set \( N_\theta \) without affecting the defining equality. Unfortunately, such negligible ambiguities can become a major obstacle if we want the map \( \theta \mapsto M_\omega(\theta) \) to be differentiable, or even continuous.

Regular conditional distributions eliminate the obstacle, by consolidating all the bookkeeping with negligible sets into one step.

Recall that the distribution of a real-valued random variable \( X \) is a probability measure \( \mathbb{P} \) on \( \mathcal{B}(\mathbb{R}) \) for which \( \mathbb{P} f(X) = Pf \), at least for all bounded, \( \mathcal{B}(\mathbb{R}) \)-measurable real functions \( f \). A regular conditional distribution for \( X \) given a sub-sigma-field \( \mathcal{G} \) is a Markov kernel, a set \( \{ \mathbb{P}_\omega : \omega \in \Omega \} \) of probability measures on \( \mathcal{B}(\mathbb{R}) \) for which \( \mathbb{P}_\omega f \) is a version of the conditional expectation \( \mathbb{P}_G f(X) \) whenever it is well defined. That is, \( \omega \mapsto \mathbb{P}_\omega f \) is \( \mathcal{G} \)-measurable and \( \mathbb{P}_\omega g(\omega) f(X(\omega)) = \mathbb{P} (g(\omega) \mathbb{P}_\omega f) \) for a suitably large collection of \( \mathcal{B}(\mathbb{R}) \)-measurable functions \( f \) and \( \mathcal{G} \)-measurable functions \( g \). Such regular conditional distributions always exist (Breiman, 1968, Section 4.3).

**Remark.** As noted by Breiman (1968, page 80), regular conditional distributions also justify the familiar idea that for the purpose of calculating \( \mathbb{P}_G \) conditional expectations we may treat \( \mathcal{G} \)-measurable functions as constants. For example, if \( f \) is a bounded \( \mathcal{B}(\mathbb{R}^2) \setminus \mathcal{B}(\mathbb{R}) \)-measurable real function on \( \mathbb{R}^2 \) and \( G \) is a \( \mathcal{G} \)-measurable random variable then \( \mathbb{P}_G f(X,G) = \mathbb{P}_G f(x,G(\omega)) \) almost surely. You might find it instructive to prove this assertion using a generating class argument, starting from the fact that it is valid when \( f \) factorizes: \( f(x,y) = f_1(x) f_2(y) \).

If we choose \( P_\omega e^{\theta x} \) as the version of the conditional moment generating function \( M_\omega(\theta) \) then the arguments from Section 2.2 work as before: on the interior of the interval \( J_\omega = \{ \theta \in \mathbb{R} : M_\omega(\theta) < \infty \} \) we have

\[
L_\omega(\theta) := \log M_\omega(\theta)
\]

\[
L_\omega'(\theta) = P_\omega \left( x e^{\theta x} / M_\omega(\theta) \right) = \mathbb{P}_{\omega,\theta} x
\]

\[
L_\omega''(\theta) = \text{var}_{\omega,\theta}(x) = \mathbb{P}_{\omega,\theta} (x - \mu_{\omega,\theta})^2
\]

where \( \mu_{\omega,\theta} := \mathbb{P}_{\omega,\theta} x \).

As before, the probability measure \( P_{\omega,\theta} \) is defined on \( \mathcal{B}(\mathbb{R}) \) by its density, \( dP_{\omega,\theta}/dP_\omega = e^{x\theta}/M_\omega(\theta) \). If 0 and \( \theta \) are interior points of \( J_\omega \) then
equality \(<2>\) takes the cleaner form
\[
L_\omega(\theta) = \theta P_\omega x + \frac{1}{2} \theta^2 \text{var}_{\omega, \theta^*}(x) \quad \text{for all } \omega,
\]
with \(\theta^*\), which might depend on \(\omega\), lying between 0 and \(\theta\).

Now suppose there exist functions \(a_\omega\) and \(b_\omega\) for which \(P_\omega[a_\omega, b_\omega] = 1\). As in Section 2.5 we have
\[
P_\omega e^{\theta(x-P_\omega x)} \leq \exp\left(\frac{1}{8} \theta^2 (b_\omega - a_\omega)^2\right)
\]
so that
\[
P_\omega e^{\theta(x-P_\omega x)} \leq \exp\left(\frac{1}{8} \theta^2 R_\omega^2\right) \quad \text{if } R_\omega \geq b_\omega - a_\omega.
\]

**Remark.** Strangely enough, the first inequality does not require any sort of measurability assumption about \(a_\omega\) and \(b_\omega\), because \(\omega\) is being held fixed. However, we usually need the range \(R_\omega\) to be \(\mathcal{G}\)-measurable when integrating out \(<3>\).

Similarly, if \(P_\omega(-\infty, b_\omega] = 1\) and \(v_\omega \geq P_\omega x^2\) then
\[
P_\omega e^{\theta(x-P_\omega x)} \leq \exp\left(\frac{1}{2} \theta^2 v_\omega \Delta(\theta b_\omega)\right) \quad \text{for } \theta \geq 0,
\]
where, as before, \(\Delta\) denotes the convex increasing function on \(\mathbb{R}\) for which
\[
\frac{1}{2} x^2 \Delta(x) = \Phi_1(x) = e^x - 1 - x.
\]

### 3.2 Hoeffding’s inequality for martingales

Consider the martingale \(\{(S_i, \mathcal{F}_i) : i = 1, \ldots, n\}\) described at the start of Section 3.1.

Suppose the conditional distribution of the increment \(\xi_i\) given \(\mathcal{F}_{i-1}\) concentrates on an interval whose length is bounded above by an \(\mathcal{F}_{i-1}\)-measurable function \(R_i(\omega)\). Then inequality \(<3>\) gives
\[
P_{i-1} e^{\theta \xi_i} \leq \exp\left(\frac{1}{8} \theta^2 R_i^2\right).
\]

If the \(R_i\)'s are actually constants then
\[
P e^{\theta S_i} = P \left( e^{\theta S_{i-1} P_{i-1} e^{\theta \xi_i}} \right) \leq \exp\left(\frac{1}{8} \theta^2 R_i^2\right) P e^{\theta S_{i-1}},
\]
which, by repeated substitution, gives
\[
\Pr[e^{\theta S_n} \leq \prod_{i=1}^{n} \exp \left( \frac{1}{8} \theta^2 R_i^2 \right) = e^{\theta \Pr S_n + \theta^2 W/8} \text{ where } W = \sum_{1 \leq i \leq n} R_i^2.
\]

For each nonnegative \( x \),
\[
\Pr[S_n \geq x + \Pr S_n] \leq \inf_{\theta \geq 0} \exp(-x\theta + W\theta^2/8) = \exp(-2x^2/W).
\]

We have the same bound as for a sum of independent increments, as Hoeffding promised.

If the \( R_i \)'s are truly random, a slightly more subtle argument leads to the inequality
\[
\Pr[S_n \geq x + \Pr S_n] \leq \Pr[\sum_{i \leq n} R_i^2 > W] + \Pr\left[\sum_{i \leq n} R_i^2 > W\right] \text{ for each constant } W.
\]

Compare with McDiarmid, 1998, Theorem 3.7.

To prove \( \leq 7 \) define \( V_i := \sum_{k \leq i} R_k^2 \), which is \( \mathcal{F}_{i-1} \) measurable. For each \( i \),
\[
\Pr[\exp (\theta S_i - \theta^2 V_i/8) = \Pr \left[ \exp (\theta S_{i-1} - \theta^2 V_{i-1}/8) \Pr_{i-1} e^{\theta \xi} \right] \leq \Pr \exp (\theta S_{i-1} - \theta^2 V_{i-1}/8) \text{ by } \leq 5>.
\]

Repeated substitution now gives
\[
\Pr[\exp (\theta S_n - \theta^2 V_n/8)] \leq \Pr[\exp (\theta S_0)] = e^{\theta \Pr S_n}.
\]

The strange \( V_n \) contribution is not a problem on the set where \( V_n \leq W \):
\[
\Pr[e^{\theta S_n} \{V_n \leq W\} \leq \Pr[e^{\theta S_n + \theta^2(W-V_n)/8} \leq e^{\theta \Pr S_n + \theta^2 W/8}
\]

so that
\[
\Pr[S_n \geq x + \Pr S_n] \leq \Pr[V_n > W] + \Pr[\sum_{i \leq n} R_i^2 > W] \text{ for each constant } W.
\]

And so on, until inequality \( \leq 7 > \) emerges.

**Example.** Suppose \( X = (X_1, \ldots, X_n) \) is a vector of independent random variables and \( f \) is a \( \text{B}(\mathbb{R}^n) \)-measurable function with the **bounded difference** property: for constants \( c_i \),
\[
|f(x) - f(z)| \leq c_i \text{ if } x \text{ and } z \text{ differ only in the } i\text{th coordinate.}
\]
More succinctly,

\[ |f(x_1, \ldots, x_n) - f(z_1, \ldots, z_n)| \leq \sum_{i \leq n} c_i \{x_i \neq z_i\} \quad \text{for all } x, z \in \mathbb{R}^n. \]

Then an appeal to <6> will give

\[ \mathbb{P}\{f(X) \geq t + \mathbb{P}f(X)\} \leq \exp\left( -2t^2 / \sum_i c_i^2 \right). \]

**Remark.** If you crave greater generality, let \( X_i \) take values in some measurable space \( (X_i, A_i) \) and let \( f \) be product-measurable. In that setting you will be grateful that <3> does not require measurability of \( a_\omega \) and \( b_\omega \).

As before, let me simplify our lives by assuming \( \mathbb{P} = \otimes_i P_i \), a product measure, and the \( X_i \)'s are the coordinate maps. Define \( Q_i = \otimes_{i < k \leq n} P_k \), the joint distribution of \( (x_{i+1}, \ldots, x_n) \). Write \( \mathcal{F}_i \) for the sigma-field on \( \mathbb{R}^n \) generated by \( x_1, \ldots, x_i \) and \( \mathcal{F}_0 \) for the trivial sigma-field. Then

\[ S_i := P_i f(x) = f_i(x_1, \ldots, x_i) := Q_i f(1, \ldots, x_n) \quad \text{for } i = 0, 1, \ldots, n \]

is a martingale with \( S_0 = \mathbb{P}f(X) \) and increments

\[ \xi_i = f_i(x_1, \ldots, x_i) - f_{i-1}(x_1, \ldots, x_{i-1}). \]

Note that \( f_i \) inherits a bounded difference property from \( f \):

\[
|f_i(x_1, \ldots, x_i) - f_i(z_1, \ldots, z_i)| \\
\leq Q_i |f(x_1, \ldots, x_i, x_{i+1}, \ldots, x_n) - f(z_1, \ldots, z_i, x_{i+1}, \ldots, x_n)| \\
\leq Q_i \sum_{k \leq i} c_k \{x_k \neq z_k\} = \sum_{k \leq i} c_k \{x_k \neq z_k\}.
\]

The last integration with respect to \( Q_i \) has no effect because the \( x_{i+1}, \ldots, x_n \) coordinates have already disappeared from the bound.

The conditional distribution of \( \xi_i \) given \( \mathcal{F}_{i-1} \) concentrates on the interval \([a(x_1, \ldots, x_{i-1}), b(x_1, \ldots, x_{i-1})]\), where

\[
a(x_1, \ldots, x_{i-1}) = \inf_{y_i} f_i(x_1, \ldots, x_{i-1}, y_i) - f_i(x_1, \ldots, x_i) \\
b(x_1, \ldots, x_{i-1}) = \sup_{z_i} f_i(x_1, \ldots, x_{i-1}, z_i) - f_i(x_1, \ldots, x_i)
\]

**Remark.** You might worry about the measurability of \( a \) and \( b \) if the \( y_i \) and \( z_i \) range over uncountable sets. It’s a good thing that the argument only depends on an upper bound for \( b - a \).
The interval has length at most $c_i$ because
\[ |f_i(x_1, \ldots, x_{i-1}, z_i) - f_i(x_1, \ldots, x_{i-1}, y_i)| \leq c_i \]
And so on.

\[ \square \]

**Remark.** McDiarmid (1989, page 159, after Lemma 4.1) noted that the cruder inequality $-c_i \leq \xi_i \leq c_i$ contributes an unwanted factor of $1/4$ in the exponent of $<10>$. He stressed the importance of working with sharper (conditional) bounds on the squared range, rather than (unconditional) bounds on the increments.

**Example.** (The Erdős-Rényi random graph) The edges of a graph with $m$ vertices are a subset of the set $E$ of all $n = \binom{m}{2}$ pairs of distinct vertices. Two edges are said to be adjacent if they share an endpoint. Three edges form a triangle if together they contain only three vertices: that is, the edges are $\{i, j\}, \{j, k\}$, and $\{k, i\}$ for distinct vertices $i, j, k$. The subset $\mathcal{T}$ of $E^3$ of all triangles has size $N = \binom{\binom{m}{2}}{3}$.

The Erdős-Rényi random graph $G_m(p)$ chooses its edges by a set of independent random variables $\{X_e : e \in E\}$, with each $X_e$ distributed Ber$(p)$. That is, $G_m(p)$ includes $e$ when $X_e = 1$, an event with probability $p$.

The number of triangles in $G_m(p)$ equals
\[ f(X) = \sum_{\{e_1, e_2, e_3\} \in \mathcal{T}} X_{e_1}X_{e_2}X_{e_3} \]

The expected number of triangles is $\mathbb{E}[f(X)] = Np^3$.

A change in a single $x_e$ can have a big effect on $f(x)$. For example, if $x_e = 1$ for all $e \in E$ and just one $x_e$ is changed to zero then $m - 2$ triangles disappear. The constants corresponding to the $c_i$’s in inequality $<9>$ must all be equal to $m - 2$. The two-sided version of inequality $<10>$ gives
\[ \mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq \epsilon\} \leq 2 \exp \left( -\frac{2\epsilon^2}{n(m-2)^2} \right) \]

For what values of $p$ (depending on $m$) might that inequality assert something nontrivial? For example, suppose we wanted to show that
\[ \mathbb{P}\{|f(X) - \mathbb{E}[f(X)]| \geq \frac{1}{2}\mathbb{E}[f(X)]\} \to 0. \]

We know that $n$ grows like $m^2$ and $N$ grows like $m^3$, so we have $\epsilon$ of order $m^3p^3$, which makes the exponent in the inequality have order $m^2p^6$. We would need $m^{1/3}p \to \infty$ to make the bound go to zero.
In fact, arguments based on the Chen-Stein method (as in Chapter 5 of Barbour et al., 1992) show that $f(X_i)$ is approximately Poisson($Np^3$) distributed. To achieve $<12>$ it should suffice to have the mean of the distribution going to infinity, that is, $mp \to \infty$.

Why has inequality $<10>$ fallen so far short of the best answer? The blame lies with the pessimistic choice $c_i = m - 2$ forced by the unlikely configuration where $G_m(p)$ contains all $n$ possible edges. It is much more likely that a possible edge, say $e = \{1, 2\}$, is involved in only about $mp^2$ potential triangles: the expected number of vertices $k \geq 3$ for which $X_{\{1,k\}} = 1 = X_{\{2,k\}}$ equals $(m - 2)p^2$. A change in $X_e$ probably changes $f(X)$ by $O(mp)$. If by some modification of the argument we could reduce the effective $c_i$ to a term of order $mp^2$ then the exponent in the bound from inequality $<10>$ would be changed to something of order $m^2p^2$, which would then match the conclusion suggested by the Poisson approximation. The Kim-Vu inequality in Section 3.4 captures (modulo a few logs) this idea that the degree of concentration should be controlled by average rather than extreme behavior.

To be continued in Example $<21>$.

### 3.3 Bennett’s inequality for martingales

If $X = X_1 + \cdots + X_n$, a sum of independent random variables with each $X_i$ bounded above by some constant $b$ and $W \geq \sum_{i \leq n} P X_i^2$, then Bennett’s inequality from Section 2.6 asserts that

$$P\{X \geq t + P X\} \leq \exp \left(-\frac{t^2}{2W} \psi_{Benn}(bt/w)\right) \quad \text{for } t \geq 0$$

where $\psi_{Benn}$ is the decreasing, convex function on $[-1, \infty)$ for which

$$\frac{1}{2} t^2 \psi_{Benn}(t) = h(t) := (1 + t) \log(1 + t) - t.$$

Inequality $<4>$ extends the result to sums of martingale differences. The following theorem illustrates the method. The resulting bound interests me particularly because of its application to the Kim-Vu inequality in Sections 3.4 and 3.5.

**Theorem.** Suppose $S_i = \mu + \xi_1 + \cdots + \xi_i$ for $i \leq n$ is the martingale, with $\mu = P S_n$, described at the start of Section 3.1. Let $V_i := \sum_{1 \leq k \leq i} v_i$, where $v_i$ is an $\mathcal{F}_{i-1}$-measurable random variable with $v_i \geq P_{S_{i-1}} \xi_i^2$, and...
§3.3 Bennett’s inequality for martingales

Let $M_i$ be an $\mathcal{F}_{i-1}$-measurable random variable for which $\xi_i \leq M_i$. Then, for each $x \geq 0$ and nonnegative constants $b$ and $W$,

$$
P\{S_n \geq x + \mu\} \leq \exp\left(-\frac{x^2}{2W} \psi_{\text{Benn}} \left(\frac{bx}{W}\right)\right) + P\{\max_i M_i > b \text{ or } V_n > W\}.
$$

**Corollary.** If, in addition to the assumptions of the Theorem we actually have two-sided control over the increments, $|\xi_i| \leq M_i$, then $P\{|S_n - \mu| \geq x\}$ is less than twice the upper bound for $P\{S_n \geq x + \mu\}$.

**Remark.** The sequences $\{v_i\}$ and $\{M_i\}$ are said to be **predictable**, because the values $v_i$ and $M_i$ are both determined by what we learn up to step $i - 1$. They are used in the proof to define **predictable stopping times**. Compare with Pollard (2001, Section 6.5).

**Proof.** As with many martingale arguments, the main idea is to multiply the increments $\xi_i$ by predictable weights, which sets up an appeal to inequality <4>. The weights come from the stopping time

$$
\tau := \inf\{i \geq 1 : M_i > b \text{ or } V_i > W\}.
$$

As usual, the infimum of an empty set equals $+\infty$; that is, $\tau = +\infty$ if $\max_i M_i \leq b$ and $V_n \leq W$. The stopping time is predictable because

$$
\{\tau \leq i\} = \{\max_{k \leq i} M_k > b \text{ or } V_i > W\} \in \mathcal{F}_{i-1} \quad \text{for } 1 \leq i \leq n.
$$

The reweighted increment $\eta_i := \xi_i\{i < \tau\}$ also has zero $P_{\mathcal{F}_{i-1}}$ conditional expectation, because $\{i < \tau\}$ is $\mathcal{F}_{i-1}$-measurable:

$$
P_{i-1} \eta_i = P_{i-1} (\xi_i\{i < \tau\}) = \{i < \tau\} P_{i-1} \xi_i = 0.
$$

Moreover $\eta_i \leq b$, because $M_i \leq b$ when $\tau > i$; and

$$
P_{i-1} \eta_i^2 = \{i < \tau\} P_{i-1} \xi_i^2 \leq v_i.
$$

For each $\theta \geq 0$ inequality <4> implies $P_{i-1} e^{\theta \eta_i} \leq \exp\left(\frac{1}{2} \theta^2 v_i \Delta(\theta b)\right)$ so that

$$
\{i < \tau\} P_{i-1} \exp\left(\theta \xi_i - \frac{1}{2} \theta^2 v_i \Delta(\theta b)\right) \leq 1.
$$

Now argue as in Section 3.2, this time for the process

$$
D_i = \exp\left(\theta S_i - \frac{1}{2} \theta^2 V_i \Delta(\theta b)\right),
$$

Draft: 8Oct2015 © David Pollard
suitably stopped. For $1 \leq i \leq n$,
\[
\mathbb{P} D_i \{i < \tau\} = \mathbb{P} \left( D_{i-1} \{i < \tau\} \mathbb{P}_{i-1} e^{\theta \xi_i - \theta^2 v_i \Delta(\theta b)/2} \right) \\
\leq \mathbb{P} D_{i-1} \{i - 1 < \tau\}
\]
By repeated substitution
\[
\exp \left( -\frac{1}{2} \theta^2 W \Delta(\theta b) \right) \mathbb{P} e^{\theta S_n} \{n < \tau\} \leq \mathbb{P} D_n \{n < \tau\} \leq \mathbb{P} D_0 \{0 < \tau\} = e^{\theta \mu}.
\]
Conclude that
\[
\mathbb{P} \{S_n \geq x + \mu\} \leq \mathbb{P} \{\tau \leq n\} + \inf_{\theta \geq 0} \mathbb{P} \left( \{\tau > n\} e^{\theta (S_n - x - \mu)} \right) \\
\leq \mathbb{P} \{\tau \leq n\} + \inf_{\theta \geq 0} \exp \left( -\theta x + \frac{1}{2} \theta^2 W \Delta(\theta b) \right).
\]
Calculate the final infimum in the same way as in Section 2.6.

\section*{3.4 Concentration of random polynomials}

The main ideas for the following material come from the papers by Kim and Vu (2000) and Vu (2002). See the Notes for an explanation of the liberties I am taking in referring to the arguments as the Kim-Vu method.

Their method deals with random variables expressible as polynomials $f(X)$ in independent random variables $X = (X_1, \ldots, X_n)$, each taking values in $[0, 1]$. For example, we might have $f(x) = 3x^7_1 x^5_2 + 2x^7_1 x^2_3 x_4 + 7x^4_1 x^5_2 x_9 + 6x^2_1 x^4_3$ for $n = 9$. The degree of $f$ is defined to be the largest of the degrees of its terms, 17 for the example just given. At the cost of an extra factor of 2 we may assume that all the coefficients in the polynomial are nonnegative.

The problem is to find useful exponential tail bounds for $|f(X) - \mathbb{E} f(X)|$. Even though the polynomials do satisfy bounded difference conditions, as in Example 8, the $c_i$ coefficients can be too big for the bounds from that Example to be helpful; the squared ranges can be too crude an upper bound for the variances. Instead, a recursive appeal to Theorem 14 will give a better result that involves two quantities, $\mathcal{E}_0(f)$ and $\mathcal{E}_1(f)$ (see 18 below), that play the scaling role of a variance. As before
\[
h(t) := (1 + t) \log(1 + t) - t \quad \text{for } t \geq -1.
\]

\begin{equation}
\mathbb{P} \{|f(X) - \mathbb{E} f(X)| \geq C_k \lambda^k \sqrt{\mathcal{E}_0(f) \mathcal{E}_1(f)} \} \leq C_{n,k} e^{-h(\lambda)} \quad \text{for all } \lambda \geq 0,
\end{equation}

Kim-Vu inequality. For every $k$th degree polynomial $f(X)$ (with nonnegative coefficients),
§3.4 Concentration of random polynomials

where $C_{n,k}(\lambda) \leq 4(n+1)^{k-1}$ and the constants $C_k$ are defined recursively by $C_1 = 1$ and $C_k = 2k(1 + C_{k-1})$.

For the sake of comparison, here is the two-sided version of the Bennett inequality for a sum $X = X_1 + \cdots + X_n$ of independent random variables with $|X_i| \leq b$ and $W \geq \sum_i \mathbb{P}X_i^2$: for $t \geq 0$,

$$\mathbb{P}|X - \mathbb{E}X| \geq t \leq \exp \left(-\frac{t^2}{2W} \psi_{\text{Benn}} \left(\frac{bt}{W}\right)\right) = 2 \exp \left(-\frac{Wb^2}{b^2} h \left(\frac{bt}{W}\right)\right).$$

It takes a bit of notation to describe the two scaling constants, $E_0(f)$ and $E_1(f)$. For me, digesting their definition was the hardest part of understanding the papers.

Once again I simplify our lives by assuming that $\mathbb{P} = \otimes_{i \leq n} P_i$ on $\mathcal{B}[0,1]^n$ and the $X_i$'s are the coordinate maps: $X_i(x) = x_i$ for $x = (x_1, \ldots, x_n) \in [0,1]^n$. Write $\mathbb{Z}_n$ for the set of all $n$-tuples of nonnegative integers and $|a|$ for $\sum_i a_i$ if $a \in \mathbb{Z}_n$. For $a, \nu \in \mathbb{Z}_n$ write $\nu \leq a$ to mean $\nu_i \leq a_i$ for all $i$ and define

$$x^a := \prod_{i \leq n} x_i^{a_i} \quad \text{AND} \quad \partial_\nu = \frac{\partial^{\nu_1}}{\partial x_1^{\nu_1}} \cdots \frac{\partial^{\nu_n}}{\partial x_n^{\nu_n}},$$

where $x^0 \equiv 1$ and $\partial^0 / \partial x_i^0$ factors are ignored. For example, if $x \in [0,1]^n$ then

$$\partial_\nu x^a = \begin{cases} \prod_{a_i > 0} (a_i)_{\nu_i} x_i^{a_i - \nu_i} & \text{if } \nu \leq a, \\ 0 & \text{otherwise} \end{cases}$$

where

$$(a_i)_{\nu_i} := a_i(a_i - 1) \cdots (a_i - \nu_i + 1) = \frac{a_i!}{(a_i - \nu_i)!} \quad \text{for } 0 \leq \nu_i \leq a_i.$$ Notice that the contribution from $x_i$ disappears (leaving only an $a_i!$) if $\nu_i = a_i$. This small fact plays a subtle role in the Kim-Vu method.

The expected value of various $\partial_\nu x^a$ terms will appear as coefficients in the polynomials that define the predictable $M_i$’s and the conditional variances $v_i$ that are needed for an appeal to Theorem <14>.

The $k$th degree polynomial can written as

$$f(x) = f_k(x) = \sum_{a \in \mathcal{A}} w_a x^a,$$

where $\mathcal{A}$ is a finite subset of $\mathbb{Z}_n$ with $|a| = k$ and $w_a \geq 0$. The polynomial has expected value

$$\mathbb{E}f_k(x) = \sum_{a \in \mathcal{A}} w_a \mu_a \quad \text{where } \mu_a := \prod_{a_i > 0} P_i x_i^{a_i}.$$
\[ E \]

\[ \begin{align*}
\mathcal{E}_j(f) & := \max_{t=j} \mathcal{D}_t(f) \quad \text{where } \mathcal{D}_t(f) := \max_{|\nu|=t} \mathbb{P} \partial_\nu f. \\
\end{align*} \]

The \( \mathcal{E}_j \)'s for \( j \geq 2 \) will appear during the proof (using induction on \( k \)) in Section 3.5. Note that \( \mathcal{D}_0(f) = \mathbb{P} f(x) \) and (Problem \([2]\)) \( \mathcal{D}_\ell(f) = 0 \) for \( \ell > k \) = the degree of \( f \). By construction, \( \mathcal{E}_0(f) \geq \mathbb{P} f(x) \) and \( \mathcal{E}_0(f) \geq \mathcal{E}_1(f) \geq \cdots \geq \mathcal{E}_k(f) \geq 0 = \mathcal{E}_j(f) \) for \( j > k \).

For the remainder of the current Section let me try to give some feel for what calculation of the \( \mathcal{E}_j \)'s involves.

\[ \begin{align*}
\textbf{Example.} \quad & \text{For } g(x) = 3x_1^2x_2^5 + 2x_1^3x_2^3x_3x_4 + 7x_1^3x_2^5x_5x_9 + 6x_2^8x_9^0 \text{ and } \nu = (1, 2, 0, \ldots, 0): \\
& \partial_\nu g(x) = 120x_1x_2^3 + 36x_1^2x_2x_3x_4 + 0 + 0 \\
& \text{and } \mathbb{P} \partial_\nu g(x) = 120(P_1x_1)(P_2x_2^3) + 36(P_1x_1^3)(P_2x_2)(P_3x_3)(P_4x_4).
\end{align*} \]

\[ \Box \]

\[ \begin{align*}
\textbf{Example.} \quad & \text{Calculate the } \mathcal{E}_j(f) \text{ for the simplest case where } f \text{ has degree } 1, \\
& \text{that is, } f(x) = \sum_{i \leq n} w_i x_i \text{ with } 0 \leq w_i \text{ for all } i. \text{ (Remember that I assume } \mathcal{A} \text{, so that there is no constant term.)} \\
& \text{Write } b \text{ for } \max_i w_i \text{ and (at the risk of minor notational confusion) } \mu_i \text{ for } P_i x_i. \text{ Calculate.}
\end{align*} \]

- \( \mathcal{E}_j(f) = 0 \) for \( j \geq 2 \).
- If \( |\nu| = 1 \) then there is a single \( i \) for which \( \nu_i = 1 \) and \( \partial_\nu f = w_i \). Thus \( \mathcal{E}_1(f) = b \).
- If \( |\nu| = 0 \) then \( \partial_\nu f(x) = f(x) \) so that \( \mathcal{E}_0(f) = \max(b, \sum_i w_i \mu_i) \).

Consequently,

\[ \gamma^2 := \mathcal{E}_0(f) \mathcal{E}_1(f) = \max \left( b^2, b \sum_i w_i \mu_i \right) \]

and the \( k = 1 \) case of the Kim-Vu inequality becomes

\[ \mathbb{P} \{ \sum_{i \leq n} w_i(x_i - \mu_i) \geq \lambda \gamma \} \leq 4 \exp(-h(\lambda)). \]
§3.4  Concentration of random polynomials

Compare with the two-sided version of the Bennett inequality \(<13>\) for independent summands \(w_i x_i\) with \(W := \gamma^2 \geq b \sum_i w_i \mu_i \geq \sum_i P(w_i x_i)^2\) and \(\max_i |w_i x_i| \leq b:\)

\[\star := P\left\{ \left| \sum_i (x_i - \mu_i) \right| \geq \lambda \gamma \right\} \leq 2 \exp \left( -\frac{\gamma^2}{b^2} h \left( \frac{b \lambda \gamma}{\gamma^2} \right) \right).\]

From the fact that \(h(t)/t^2\) is a decreasing function and \(b/\gamma \leq 1\) we have

\[h(\lambda b/\gamma \gamma^2) \geq h(\lambda \gamma^2) \text{ AND } \star \leq 2 \exp (-h(\lambda)).\]

That is, the Kim-Vu inequality for \(k = 1\) follows directly from the Bennett inequality.

Example. (Continued from Example \(<11>\)) Recall the challenge of proving concentration of the number of triangles

\[f(x) = \sum_{\{e_1, e_2, e_3\} \in \mathcal{T}} x_{e_1} x_{e_2} x_{e_3}\]

near its expected value \(N p^3\) for the Erdős-Rényi random graph on \(m\) vertices with \(X_e \sim \text{Ber}(p)\) for each of the \(n = \binom{m}{2}\) possible edges and \(N = \# \mathcal{T} = \binom{m}{3}\). Specifically, I wanted to show that

\[\mathbb{P}\{|f(x) - \mathbb{E} f(x)| \geq \frac{1}{2} \mathbb{E} f(x)| \rightarrow 0 \quad \text{if } mp \rightarrow \infty.\]

Calculate \(D_j(f)\) for \(j = 0, 1, 2, 3\). First notice that if \(\max_{e} \nu_{e} \geq 2\) then \(\partial_{\nu} f = 0\): in the sum that defines \(f(x)\) no \(x_e\) is raised to a power higher than 1. For \(D_j(f)\) we have only to consider \(\nu\) with exactly \(j\) components equal to 1, the rest equal to 0.

- For \(j = 3\) suppose \(\nu_{e} = \nu_{e'} = \nu_{e''} = 1\). Then \(\partial_{\nu} x_{e_1} x_{e_2} x_{e_3} = 0\) unless \(\{e, e', e''\} = \{e_1, e_2, e_3\} \in \mathcal{T}\). Thus \(D_3(f) = 1\).

- For \(j = 2\) suppose \(\nu_{e} = \nu_{e'} = 1\). Then \(\partial_{\nu} x_{e_1} x_{e_2} x_{e_3} = 0\) unless \(e\) and \(e'\) share an endpoint, say \(e = \{i, \alpha\}\) and \(e' = \{i, \beta\}\), and the triangle \(\{e_1, e_2, e_3\}\) has vertices \(i, \alpha, \beta\). Thus \(D_2(f) = p\).

- For \(j = 1\) suppose \(\nu_{e} = 1\), for \(e = \{\alpha, \beta\}\). Then \(\partial_{\nu} x_{e_1} x_{e_2} x_{e_3} = 0\) unless the triangle has vertices \(\alpha, \beta, \gamma\), for \(\gamma\) one of the \(m-2\) vertices not in \(e\). Thus \(D_1(f) = (m-2)p^2\).
For large enough values of $mp$,

\[
\begin{align*}
\mathcal{E}_2(f) &= \mathcal{E}_3(f) = 1 \\
\mathcal{E}_1(f) &= \max \left( 1, (m - 2)p^2 \right) \leq 1 + mp^2 \\
\mathcal{E}_0(f) &= \max \left( \mathbb{P}f(x), \mathcal{E}_1(f) \right) = \left( \frac{1}{6} + o(1) \right) (mp)^3
\end{align*}
\]

How close to $<22>$ can we get using $<17>$. We certainly need $h(\lambda)$ to grow faster than $\log^4 m$ to ensure that the upper bound in

\[
\mathbb{P}\{|f(X) - \mathbb{P}f(X)| \geq C_3 \lambda^3 \sqrt{\mathcal{E}_0(f)\mathcal{E}_1(f)} \} \leq 4(n + 1)^2 e^{-h(\lambda)}
\]

go to zero. And to also ensure that

\[
(C_3 \lambda^3)^2 (1 + mp^2) \leq \frac{1}{4} \mathbb{P}f(x) = \frac{1}{4} \left( \frac{m}{3} \right) p^3
\]

we need $mp$ greater than some suitably large constant multiple of $\log^2 m$. Pretty close to $<22>$.

\[\square\]

*3.5 \quad \textbf{Proof of the Kim-Vu inequality}

First note that inequality $<17>$ holds for trivial reasons if $\lambda \leq 3$ because $4e^{-h(3)} = 4e^3/64 > 1$. We may assume $\lambda > 3$.

\textbf{Remark.} I had hoped to write the proof in a way that would reveal the reasons for defining the $\mathcal{E}_j$’s as in $<18>$ and to show how much flexibility we have in the choice of constants, such as the $b_k$ and $W_k$ for inequality $<23>$. Unfortunately I succumbed to the temptation of just checking that chosen values lead to the desired conclusion. See Problem [4].

Argue almost as in Example $<8>$, except that the martingale version of Bennett’s inequality takes over the role played by the martingale version of Hoeffding’s inequality. That is, express $S_n = f(x)$ as $\mathbb{P}f(x)$ plus a sum of martingale differences,

\[
\xi_i = f_i(x_1, \ldots, x_i) - f_{i-1}(x_1, \ldots, x_{i-1})
\]
where
\[ f_i(x_1, \ldots, x_i) = \mathbb{Q}_i f(x_1, \ldots, x_i, x_{i+1}, \ldots, x_n) \]
\[ = \sum_{a \in A} w_a \left( \prod_{\ell \leq i} x_{a_\ell}^a \right) \left( \prod_{\ell \leq i} P_i x_{a_\ell}^a \right) \]
\[ = \sum_{a \in A} w_a x^b[a,i] \mu_{t[a,i]} \]
with \( b[a,i] := (a_1, \ldots, a_i, 0, \ldots, 0) \) and \( t[a,i] := (0, \ldots, 0, a_{i+1}, \ldots, a_n) \).

The ‘head’, \( b[a,i] \), and the ‘tail’, \( t[a,i] \), of the vector \( a \) play different roles.

The increment \( \xi_i \) is bounded in absolute value by the predictable function
\[ M_i(x_1, \ldots, x_{i-1}) := \sum_{a \in A} w_a \{ a_i > 0 \} x^b[a,i-1] \mu_{t[a,i]} \]
because
\[ |\xi_i| = \left| \sum_{a \in A} w_a x^b[a,i-1] (x^a_i - P_i x^a_i) \mu_{t[a,i]} \right|. \]
The factor \( x^a_i - P_i x^a_i \) is zero when \( a_i = 0 \) and otherwise is bounded in absolute value by 1.

For \( \tau_k(f) = C_k \sqrt{E_0(f)E_1(f)} \) and \( V_n = \sum_{i \leq n} P_{i-1} \xi_i^2 \) and all choices of positive constants \( W_k \) and \( b_k \), Theorem <14> gives
\[ \text{PROB}_k(f, \lambda) := \mathbb{P}\{ |f(x) - \mathbb{P} f(x)| \geq \tau_k(f) \lambda k \} \leq \text{BENN}_k + \text{BAD}_k \]
where
\[ \text{BENN}_k := 2 \exp \left( -\frac{W_k}{b_k^2} h \left( \frac{\tau_k \lambda k b_k}{W_k} \right) \right) \]
\[ \text{BAD}_k := \mathbb{P}\{ \max_i M_i > b_k \text{ or } V_n > W_k \} \]
\[ \leq \sum_i \mathbb{P}\{ M_i > b_k \} + \mathbb{P}\{ \max_i M_i \leq b_k \text{ and } V_n > W_k \}. \]

Notice that \( M_i \) is a polynomial of degree at most \( k - 1 \). The upper bound for \( \text{BAD}_k \) suggests an induction on \( k \). The conditional variance term \( V_n \) is also a polynomial, but unfortunately with degree usually greater than \( k \). Luckily, on the set where all the \( M_i \) are bounded above by \( b_k \), we have
\[ V_n \leq \sum_{i \leq n} \mathbb{P}_{i-1} \xi_i^2 \{ M_i \leq b_k \} \leq \beta \sum_{i \leq n} \mathbb{P}_{i-1} |\xi_i| \leq b U_n, \]
where
\[ U_n := 2 \sum_{i \leq n} \sum_{a \in A} w_a \{ a_i > 0 \} x^b[a,i-1] \mu_{t[a,i-1]}, \]
a polynomial of degree at most \( k - 1 \).
Remark. The $\theta$th summand for $M_\theta$ equals $w_\theta \{ a_i > 0 \} x^{b_i} \mu_{\{x,i\}}$ whereas the $(i,a)$th summand for $U_n$ equals the same thing multiplied by $P_i x^n_i$. I was sorely tempted to discard that extra factor, replacing $U_n$ by $\sum_i M_i$. Not a good idea. Look at the proof of Lemma <28> to see why.

Temporarily write $\gamma$ for $\tau_k(f) \lambda^{k-1}$. If we choose $W_k = \gamma^2$ and $b_k = \gamma \theta_k$ for some $\theta_k$ in $(0,1]$, inequality <24> takes a much tidier form:

$$\text{BENN}_k \leq 2 e^{-h(\lambda \theta) \theta^2} \leq e^{-h(\lambda)},$$

again by virtue of the fact that $h(t)/t^2$ is a decreasing function. Inequality <23> then becomes

$$\text{PROB}_k(f, \lambda) \leq e^{-h(\lambda)} + \sum_i \mathbb{P}\{M_i > \gamma \theta_k\} + \mathbb{P}\{U_n > \gamma/\theta_k\}.$$  

If

$$\gamma \theta_k - \mathbb{P} M_i \geq \tau_{k-1}(M_i) \lambda^{k-1} \quad \text{and} \quad \gamma / \theta_k - \mathbb{P} U_n \geq \tau_{k-1}(U_n) \lambda^{k-1}$$

then the inductive hypothesis (for $k-1$) would give

$$\text{PROB}_k(f, \lambda) \leq 2 e^{-h(\lambda)} + n C_{n,k-1} e^{-h(\lambda)} + C_{n,k-1} e^{-h(\lambda)} = C_{n,k} e^{-h(\lambda)}$$

where

$$C_{n,k} = 2 + (n + 1) C_{n,k-1}.$$  

Example <20> gave $C_{n,1} = 2$. Repeated substitution followed by summation of a geometric progression (Problem [3]) gives

$$C_{n,k} = 2 \left( \frac{(n + 1)^k - 1}{n} \right) \leq 4 (n + 1)^{k-1},$$

as asserted.

It remains only to check inequalities <27> for an appropriate $\theta_k$. Because $\mathbb{P} M_i \leq \mathcal{E}_0(M_i)$ and $\mathbb{P} U_n \leq \mathcal{E}_0(U_n)$, it suffices to check that

$$C_k \theta_k \sqrt{\mathcal{E}_0(f) \mathcal{E}_1(f)} \geq \mathcal{E}_0(M_i) + C_{k-1} \sqrt{\mathcal{E}_0(M_i) \mathcal{E}_1(M_i)}$$

$$C_k \sqrt{\mathcal{E}_0(f) \mathcal{E}_1(f) / \theta_k} \geq \mathcal{E}_0(U_n) + C_{k-1} \sqrt{\mathcal{E}_0(U_n) \mathcal{E}_1(U_n)}$$

Choose $\theta_k = \sqrt{\mathcal{E}_1(f) / \mathcal{E}_0(f)}$ and note that $\mathcal{E}_1(M_i) \leq \mathcal{E}_0(M_i)$ and $\mathcal{E}_1(U_n) \leq \mathcal{E}_0(U_n)$ to reduce the task to showing that

$$C_k \mathcal{E}_1(f) \geq (1 + C_{k-1}) \mathcal{E}_0(M_i) \quad \text{and} \quad C_k \mathcal{E}_0(f) \geq (1 + C_{k-1}) \mathcal{E}_0(U_n).$$
The next Lemma should make clear to you why the constants are defined recursively by $C_1 = 1$ and $C_k = 2k(1 + C_{k-1})$. The Lemma exploits the fact that averaging out with respect to some $x_i$ coordinates or differentiating with respect to some $x_i$ both reduce the degree of the polynomial $f(x) = \sum_{a \in A} w_a x^a$.

**Lemma.** For all $j$ and $i$,

(i) $E_j(M_i) \leq 2E_{j+1}(f)$

(ii) $E_j(U_n) \leq 2kE_j(f)$

**Proof** For (i), for a fixed $i$ and $\nu \in \mathbb{Z}_n$,

$$\mathbb{P}\partial_\nu M_i = \sum_{\ell=1}^k \frac{1}{\ell!} \sum_{a \in A} w_a \{ a_i = \ell \} \{ \nu \leq h[a, i-1] \} \mathbb{P}\partial_\nu x^{|a[i-1]|} \mu_{[x,i]}.$$  
I inserted the indicator $\{ \nu \leq h[a, i-1] \}$ just to remind you that $\partial_\nu x^{|a[i-1]|} = 0$ if $\nu \not\geq h[a, i-1]$. Compare the $(\ell,a)$th term with

$$\mathbb{P}\partial_\nu x^a = \mathbb{P}\partial_\nu x^{|a[i-1]|} (P_{i|x_i}^\ell) \mu_{[x,i]}.$$  
We can eliminate the $x^a_i$ by augmenting $\nu$ to differentiate out the $x^a_i$. Let $\epsilon$ denote the $\mathbb{Z}_n$ vector with $\epsilon_i = 1$ and $\epsilon_r = 0$ for $r \neq i$. Then

$$\mathbb{P}\partial_\nu + \epsilon x^a = \mathbb{P}\partial_\nu x^{|a[i-1]|} (\ell!) \mu_{[x,i]}.$$  
Thus

$$\mathbb{P}\partial_\nu M_i = \sum_{\ell=1}^k \frac{1}{\ell!} \sum_{a \in A} w_a \{ a_i = \ell \} \{ \nu \leq h[a, i-1] \} \mathbb{P}\partial_\nu + \epsilon x^a$$

$$\leq \sum_{\ell=1}^k \frac{1}{\ell!} \mathbb{P}\sum_{a \in A} w_a \partial_\nu + \epsilon x^a$$

$$= \sum_{\ell=1}^k \frac{1}{\ell!} \mathbb{P}\partial_\nu + \epsilon f(x)$$

$$\leq \sum_{\ell=1}^k \frac{1}{\ell!} \mathcal{D}_{|\nu + \ell \epsilon|}(f)$$

If $|\nu| = j$ then $|\nu + \ell \epsilon| \geq j + 1$ and the last sum is less than $(e - 1)E_{j+1}(f)$.

Argue similarly for (ii), using the fact that $\{ a_i > 0 \} \leq a_i$ and $\sum_i a_i \leq k$.

This time there is no need to eliminate the $x^a_i$ term.

$$\mathbb{P}\partial_\nu U_n \leq 2 \sum_{i \leq n} \sum_{a \in A} w_a a_i \mathbb{P} x^{|a[i-1]|} \mu_{[a,i-1]}$$

$$= 2 \sum_{a \in A} w_a \left( \sum_{i \leq n} a_i \right) \mathbb{P}\partial_\nu x^a$$

$$\leq 2k\mathbb{P}\partial_\nu f(x).$$

The last term is bounded by $\mathcal{D}_{|\nu|}(f)$.

\[\square\]
### 3.6 Problems

1. Prove an analog of Theorem 14 under the Bernstein-like assumption: $\mathbb{P}_{i-1}\xi_i^2 = v_i$ and $\mathbb{P}_{i-1}|\xi_i|^k \leq \frac{1}{2}v_i k! M_i^{k-2}$ for $k \geq 3$, where each $M_i$ is $\mathcal{F}_{i-1}$-measurable.

2. In the notation of Section 3.4, prove that $\partial \nu x^a = 0$ if $|\nu| = \sum_i \nu_i > \sum_i a_i = |a|$. Hint: We must have $\nu_i > a_i$ for at least one $i$.

3. Suppose $m_1, m_2, \ldots$ is a sequence of real numbers with $m_1 = \alpha > 0$ and $m_k = \alpha + \beta m_{k-1}$, with $\beta > 1$. Show that

$$m_k = \alpha \left(1 + \beta + \beta^2 + \cdots + \beta^{k-1}\right) = \alpha \frac{\beta^k - 1}{\beta - 1}.$$

4. Rewrite Sections 3.4 and 3.5 to make clearer how much flexibility we have in the definition of $E_j$’s and in the choice of the constants $W_k$ and $b_k$. In particular, I am not particularly satisfied with the part of the argument following equation 17. Then send me your improved version.

### 3.7 Notes

The inequality for bounded differences in Example 8 is often attributed to McDiarmid. However McDiarmid (1989) seemed to be giving more credit to Hoeffding. See Boucheron et al. (2013, Chapter 6) for some powerful ways to extended the bounded difference idea.

Freedman (1975) proved analogs of the Bernstein inequality, for sums of bounded martingale differences, with a sum of conditional variances taking over the role played by the second moment bound $W$.

The Bennett inequality for independent summands was proved by Bennett (1962). The extension to martingales with increments that are bounded in absolute value by 1 is due to Freedman (1975).

The relaxation to predictable upper bounds is implicit in Sections 3.3 and 3.4 of the paper of Vu (2002), although he did not express it via stopping times. He did not appeal to Bennett’s inequality, but instead derived the necessary exponential bounds by direct arguments. Something similar to inequality 17 appeared in the Kim and Vu (2000) paper, but only for $X_i$’s taking values in $\{0, 1\}$. Vu (2002) developed more elaborate versions of the inequality. Most of my discussion in Sections 3.4 and 3.5 is a translation of
the Vu and Kim-Vu papers into the framework of Bennett’s inequality for martingales. Both papers used the numbers of triangles for the Erdős-Rényi random graph (my Example $<11>$ and $<21>$) to contrast their results with the results using the method of bounded differences.

References


