Remarks on Kneip’s linear smoothers

Sören R. Künzel
David Pollard
Dana Yang

Statistics Department, Yale University
18 April 2014

1 Introduction

We have been trying to understand the analysis provided by Kneip (1994). In particular we want to persuade ourselves that his results imply the oracle inequality stated by Tsybakov (2014, Lecture 8).

This note contains our reworking of Kneip’s ideas. We refer to page x of Kneip’s paper as Kx. For $n \times n$ symmetric matrices we write $A \preceq B$ to mean that $B - A$ is positive semi-definite. Also we write $|\cdot|$ for the usual Euclidean length in $\mathbb{R}^n$, that is, $|x|^2 = \sum_{i \leq n} x_i^2$.

Following Kneip, we consider an observed $n \times 1$ random vector $y = \mu + \xi$ with unknown $\mu$ and error $\xi$ (with independent components) with $P\xi = 0$ and $\operatorname{var}(\xi) = \sigma^2 I_n$. We assume that $\xi \sim N(0, \sigma^2 I_n)$. Kneip (K844, statement of Theorem 1) assumed subgaussianity. The possible estimators are of the form $Sy$, with $S$ in a specified set $S$ of $n \times n$ (symmetric) positive semi-definite smoothing matrices that is totally ordered under the semi-definite ordering $\preceq$, with $0 \preceq S \preceq I_n$ for all $S \in S$.

Kneip considered the estimator $\hat{S}y$ with

$$\hat{S} = \arg\min_{S \in S} \tilde{G}(S) \quad \text{where} \quad \tilde{G}(S) = |y - Sy|^2 + 2\sigma^2 \text{trace}(S).$$

Here and subsequently we omit multiplicative factors of $n^{-1}$ that Kneip used. This selection procedure is the well known Mallows’ $C_p$. 

The analysis and the statement of Kneip’s main result involve two related processes, which we define for all positive semi-definite matrices $S$:

$$G_\mu(S) := |\mu - Sy|^2$$
$$M_\mu(S) := \mathbb{P}G_\mu(S) = |\mu - S\mu|^2 + \sigma^2 \text{trace}(S^2).$$

Following Kneip, we assume that the minimum of $M_\mu$ over the set $S$ is achieved at the matrix $S_\mu$ in $S$ and define

$$m^* = M_\mu(S_\mu) = \min_{S \in S} M_\mu(S)$$

We ignore all questions of whether minima are achieved and whether $\hat{S}$ is measurable.

<1> **Theorem.** (K844) There exist constants $C_1$ and $C_2$ that depend only on $\sigma^2$ for which for all $\mu$ in $\mathbb{R}^n$,

$$\mathbb{P}\{|G_\mu(\hat{S}) - G_\mu(S_\mu)| \geq \max (x^2, x\sqrt{m^*})\} \leq C_1 e^{-C_2 x} \quad \text{for } x \geq 0.$$

<2> **Corollary.** There exist constants $C_3$ and $C_4$ that depend only on $\sigma^2$ for which for all $\mu$ in $\mathbb{R}^n$,

$$\mathbb{P}G_\mu(\hat{S}) \leq m^* + C_3 \sqrt{m^*} + C_4.$$

The Corollary is equivalent to

$$\mathbb{P}G_\mu(\hat{S}) \leq (1 + \epsilon)m^* + C_0/\epsilon + C_4 \quad \text{for all } \epsilon > 0 \text{ and } C_0 = C_3^2/4,$$

a minor modification of the oracle inequality stated by Tsybakov (2014, Lecture 8). For $\epsilon$ in a bounded range the $C_4$ can be absorbed into the previous term.

The proof of the Theorem makes extensive use of the properties of the metric $d$ defined on the set of all positive semi-definite matrices $S_1$ and $S_2$ by

$$d^2(S_1, S_2) = \mathbb{P}|S_1 y - S_2 y|^2 = |(S_1 - S_2)\mu|^2 + \sigma^2 \text{trace}(S_1 - S_2)^2$$

(Note that $d^2(S_1, S_2) = nq_\mu^2(S_1, S_2)$ for the $q_\mu$ defined near the bottom of K842.) In particular, the proof relies crucially on a bound (see Section 3.2) for the packing numbers of subsets of $\overline{S}$, a set of positive semi-definite matrices that contains $S$ as a subset. The arguments rely on the total ordering of $S$ to parametrize $S$ by a subset of the real line.
2 Outline of the Proofs

To prove Theorem <1> we first show that
\[
\hat{G}(S) \approx M_\mu(S) + \text{term not depending on } S
\]
\[
G_\mu(S) \approx M_\mu(S) + \text{term not depending on } S.
\]

More precisely, with
\[
D_\mu(S) := G_\mu(S) - M_\mu(S)
\]
\[
\hat{D}(S) := \hat{G}(S) - M_\mu(S),
\]
we show: There exist positive constants \(C_1, C_2\), depending only on \(\sigma^2\) for which, for every \(r > 0\),
\[
\{ \exists S \in S : |\hat{D}(S) - \hat{D}(S_\mu)| > L(S, x, r) \} \leq C_1 e^{-C_2 x}, \tag{3}
\]
\[
\{ \exists S \in S : |D_\mu(S) - D_\mu(S_\mu)| > L(S, x, r) \} \leq C_1 e^{-C_2 x}, \tag{4}
\]
\[
\text{where } L(S, x, r) = [d^2(S, S_\mu) + r^2] x / r. \tag{5}
\]

The proof of these inequalities (in Section 3) uses a chaining argument based on control of the increments of both the \(\hat{D}\) and \(D_\mu\) processes, together with a bound on the packing numbers that derives from the total ordering of \(S\).

We also make use of an inequality (cf. K843, Proposition 1) related to the growth of \(M_\mu(S) - M_\mu(S_\mu)\) as \(d(S, S_\mu)\) increases. For that we need the matrix analog of the inequality \(\alpha^2 + \beta^2 \geq (\alpha - \beta)^2\) for nonnegative real numbers.

<6> **Lemma.** If \(S_1\) and \(S_2\) are symmetric, positive semi-definite matrices that commute then \((S_1 - S_2)^2 \prec S_1^2 + S_2^2\).

**Proof** We want to show that the matrix
\[
(S_1^2 + S_2^2) - (S_1 - S_2)^2 = 2S_1S_2
\]
is positive semi-definite. Let \(U\) be an orthogonal matrix that simultaneously diagonalizes \(S_1\) and \(S_2\) to \(\Lambda_1\) and \(\Lambda_2\). Then for any vector \(\alpha\) in \(\mathbb{R}^n\), we have
\[
\alpha' S_1S_2\alpha = (U\alpha)'\Lambda_1\Lambda_2(U\alpha),
\]
which is nonnegative because the elements of the diagonal matrix \(\Lambda_1\Lambda_2\) are all nonnegative.


As a direct consequence of the Lemma,
\[ M_\mu(S_1) + M_\mu(S_2) \]
\[ = \mu' \left[ (I_n - S_1)^2 + (I_n - S_2)^2 \right] \mu + \sigma^2 \text{trace} \left[ S_1^2 + S_2^2 \right] \]
\[ \geq \mu'(S_1 - S_2)^2 \mu + \sigma^2 \text{trace}(S_1 - S_2)^2 = d^2(S_1, S_2). \]
In particular, if \( d^2(S, S_\mu) \geq 3m^* \) then \( d^2(S, S_\mu) \geq d^2(S, S_\mu)/3 + 2m^* \), so that \( <7> \) implies
\[ M_\mu(S) - m^* \geq \frac{1}{3} d^2(S, S_\mu) \{ d(S, S_\mu) \geq \sqrt{3m^*} \}. \]

**Proof (of Theorem <1>)** With \( L \) as defined in \( <5> \), define
\[ L(S, x) := L(S, x, r_x) \quad \text{where} \quad r_x = \max(\sqrt{3m^*}, 7x). \]

By inequalities \( <3> \) and \( <4> \), we can find a set \( \Omega_x \) with probability at least \( 1 - 2C_1 e^{-C_2 x} \), on which we have
\[ <9> \max \left( |\tilde{D}(S) - \tilde{D}(S_\mu)|, |D_\mu(S) - D_\mu(S_\mu)| \right) \leq L(S, x) \quad \text{for all} \quad S \in \mathbb{S}. \]

The rest of the proof is just a deterministic argument on the set \( \Omega_x \).
Define \( \hat{d} = d(\hat{S}, S_\mu) \). Then
\[ \frac{1}{7}(d^2 + r_x^2) \geq L(\hat{S}, x) \quad \text{because} \quad x/r_x < 1/7 \]
\[ \geq \tilde{D}(S_\mu) - \tilde{D}(\hat{S}) \quad \text{by} \quad <9> \]
\[ = \tilde{G}(S_\mu) - \tilde{G}(\hat{S}) + M_\mu(\hat{S}) - M_\mu(S_\mu) \]
\[ \geq M_\mu(\hat{S}) - m^* \quad \text{because} \quad \hat{S} \text{ minimizes} \quad \tilde{G} \]
\[ \geq \frac{1}{7} d^2 \{ \hat{d} \geq \sqrt{3m^*} \} \quad \text{by} \quad <8>. \]

If \( \hat{d} \) were larger than \( r_x \) the last inequality would give \( \frac{2}{7} d^2 \geq \frac{1}{3} d^2 \), which clearly cannot be true. Thus \( \hat{d} < r_x \) on \( \Omega_x \), implying
\[ 2r_x^2/x \geq L(\hat{S}, x) \geq M_\mu(\hat{S}) - m^*. \]

In summary,
\[ <10> \quad \hat{d} := d(\hat{S}, S_\mu) < r_x \quad \text{AND} \quad M_\mu(\hat{S}) \leq m^* + 2x r_x \quad \text{on} \quad \Omega_x. \]

Combine this inequality with the bound for \( |D_\mu(S) - D_\mu(S_\mu)| \) from \( <9> \) to deduce that, again on \( \Omega_x \),
\[ |G_\mu(\hat{S}) - G_\mu(S_\mu)| \leq \left( M_\mu(\hat{S}) - M_\mu(S_\mu) \right) + |D_\mu(\hat{S}) - D_\mu(S_\mu)| \]
\[ \leq 2x r_x + L(\hat{S}, x) \]
\[ \leq 4x r_x. \]
Thus
\[ \mathbb{P}\{|G_\mu(\hat{S}) - G_\mu(S_\mu)| > 4xr_x\} \leq \Omega_x \leq 2k_1 e^{-k_2x}. \]

This inequality is not quite the result announced in Theorem <1>. However,
\[ 4xr_x = 4x \max(\sqrt{3m^*}, 7x) \geq \max(4\sqrt{3}, 28) \max(x\sqrt{m^*}, x^2) \]
so we get the announced result, for \( Z = |G_\mu(\hat{S}) - G_\mu(S_\mu)| \):
\[ \mathbb{P}\{Z \geq \max(x^2, x\sqrt{m^*})\} \leq C_1 e^{-C_2x} \quad \text{for } x \geq 0. \]

by adjusting the constants.

The oracle inequality stated as Corollary <2> is an integrated version of the tail bound from Theorem <1>.

**Proof** From inequality <11> we have \( \mathbb{P}\{Z \geq f(x)\} \leq C_1 e^{-C_2x} \) for \( x \geq 0 \), where \( f(x) = \max(x^2, x\sqrt{m^*}) \), which gives
\[
|\mathbb{P}G_\mu(\hat{S}) - m^*| \leq \mathbb{P}Z = \int_0^\infty \mathbb{P}\{Z > t\} dt = \int_0^\infty \mathbb{P}\{Z \geq f(x)\} f'(x) dx \\
\leq \int_0^\infty \max(2x, \sqrt{m^*}) C_1 e^{-C_2x} dx \\
\leq \frac{C_1}{C_2} \sqrt{m^*} + 2 \frac{C_1}{C_2} e^{-C_2\sqrt{m^*}} + \frac{C_1}{C_2} \sqrt{m^*} e^{-C_2\sqrt{m^*}} \\
\leq C_3 \sqrt{m^*} + C_4
\]
for new constants \( C_3 = 2C_1/C_2 \) and \( C_4 = 2C_1/C_2^2 \).

3 Technical Stuff

This section proves the inequalities <3> and <4>,
\[
\mathbb{P}\{\exists S \in S : |\hat{D}(S) - \hat{D}(S_\mu)| > L(S, x, r)\} \leq C_1 e^{-C_2x}, \\
\mathbb{P}\{\exists S \in S : |D_\mu(S) - D_\mu(S_\mu)| > L(S, x, r)\} \leq C_1 e^{-C_2x},
\]
where \( L(S, x, r) = \left[d^2(S, S_\mu) + r^2\right] x/r \),

by means of a chaining argument with stratification. The necessary ingredients are the control of increments of the \( \hat{D} \) and \( D_\mu \) processes and bounds on packing numbers.
3.1 Exponential bounds for increments

The next Lemma is all we need to control the increments of the \( \hat{D} \) and \( D_\mu \) processes under the assumption of gaussian errors. First we expand each process into sums of simpler processes.

\[
D_\mu(S) = G_\mu(S) - M_\mu(S) = X_1(S) + X_2(S)
\]
\[
\hat{D}(S) = \hat{G}(S) - M_\mu(S) = X_3(S) + X_4(S) + n\sigma^2.
\]

where

\[
X_1(S) = \xi' S^2 \xi - \sigma^2 \text{trace}(S^2)
\]
\[
X_2(S) = -2\mu'(S - S^2) \xi
\]
\[
X_3(S) = \xi' (I - S)^2 \xi - \sigma^2 \text{trace}(I - S)^2
\]
\[
X_4(S) = 2\mu'(I - S)^2 \xi
\]

Notice that each \( X_i(S) \) is either a linear or quadratic function of \( \xi \).

\begin{align*}
\text{Lemma.} \quad &\text{(Compare with Kneip’s Lemma 2, K852)} \quad \text{Suppose } z \sim N(0, I_n). \\
&\text{For each vector of constants } a \text{ and each symmetric matrix } A, \\
&P\{z' a \geq w | a| \} \leq \exp(-w^2/2) \\
&P\{z' Az - \text{trace}(A) \geq w \sqrt{\text{trace}(A^2)} \} \leq 2e^{-w/4}
\end{align*}

for each \( w \geq 0 \).

\begin{proof}
The first inequality is just the usual bound for \( N(0,1) \) tails. (It extends easily to the subgaussian case.) For the second inequality write \( A = L' \text{diag}(\lambda_1, \ldots, \lambda_n) L \), with \( L \) orthogonal. Write \( \kappa \) for \( \sqrt{\text{trace}(A^2)} = |\lambda| \). Then \( x = Lz \sim N(0, I_n) \). With \( t = 1/(4\kappa) \),

\[
P\{z' Az - \text{trace}(A) \geq w \sqrt{\text{trace}(A^2)} \} = P\{\sum_i \lambda_i (x_i^2 - 1) \geq w\kappa\} \\
\leq e^{-tw\kappa} \prod_i P\exp\left( -t\lambda_i + t\lambda_i x_i^2 \right) \\
= e^{-w/4} \exp\left( \sum_i \left( -t\lambda_i - \frac{1}{2} \log(1 - 2t\lambda_i) \right) \right).
\]

As \( \max_i |2t\lambda_i| \leq 1/2 \), we have \(-\frac{1}{2} \log(1 - 2t\lambda_i) \leq t\lambda_i + \frac{1}{2}(2t\lambda_i)^2 \), which leaves \( 2t^2 \sum_i \lambda_i^2 = 1/8 < \log 2 \) in the exponent.
\end{proof}
The argument for the quadratic form comes from Nolan and Pollard (1987, Lemma 3). For subgaussian errors Kneip calculated moments, resulting in a bound similar to an earlier result of Hanson and Wright (1971). The Rudelson and Vershynin (2013) method provides a simpler derivation.

The \( \exp(-w^2/2) \) bound for \( z'a \) is more than we need. The inequality

\[
\min\left(1, 2e^{-w^2/2}\right) \leq 4\exp(-w/4) \quad \text{for all } w \geq 0.
\]

shows that all the increments of the \( X_i \) processes from \(<12>\) satisfy inequalities of the form

\[
\begin{align*}
\mathbb{P}\{|X_i(S_1) - X_i(S_2)| > d(\theta_1, \theta_2)x\} & \leq C_1e^{-C_2x} \quad \text{for } x \geq 0,
\end{align*}
\]

for constants \( C_1 \) and \( C_2 \).

### 3.2 Packing bounds

The assumption on \( S \) ensures the matrices can be diagonalized by a fixed rotation: \( S = U'\Lambda(S)U \) with \( U \) orthogonal and

\[
\Lambda(S) = \text{diag} (\lambda_1(S), \ldots, \lambda_n(S))
\]

The total ordering ensures that each \( S \in \mathcal{S} \) is uniquely determined by its trace. The set \( \mathcal{S} \) can be parametrized as \( S_\theta \), with \( \theta \in \Theta \subseteq [0, n] \), where

\[
\Lambda(S_\theta) = \Lambda(\theta) = \text{diag} (\lambda_1(\theta), \ldots, \lambda_n(\theta)) \quad \text{and } \theta = \sum_{i \leq n} \lambda_i(\theta).
\]

The maps \( \theta \mapsto \lambda_i(\theta) \) are increasing, for each \( i \). As Kneip showed (by interpolation, K857), \( \mathcal{S} \) can be embedded into a larger family of positive semi-definite matrices \( \overline{\mathcal{S}} = \{S_\theta : \theta \in \overline{\Theta}\} \) with \( S_\theta = U'\Lambda(\theta)U \) and \( \theta \mapsto \lambda_i(\theta) \) continuous and nondecreasing from \( \overline{\Theta} = [0, n] \) onto \([0, 1] \). The monotonicity of \( \theta \mapsto \lambda_i(\theta) \) simplifies calculation of packing/covering numbers for subsets of \( \overline{\Theta} = [0, n] \) under the metric \( d \). Recall that

\[
d^2(\theta_1, \theta_2) = \sum_i (\rho_i^2 + \sigma^2)|\lambda_i(\theta_1) - \lambda_i(\theta_2)|^2
\]

and \( \theta \mapsto \lambda_i(\theta) \) is nondecreasing. If \( a \leq t_1 < t_2 < \cdots < t_N \leq b \) then

\[
|\lambda_i(b) - \lambda_i(a)|^2 \geq \left(\sum_{j=2}^N \lambda_i(t_j) - \lambda_i(t_{j-1})\right)^2 \geq \sum_j |\lambda_i(t_j) - \lambda_i(t_{j-1})|^2
\]

which implies

\[
<15> \quad d^2(a, b) \geq \sum_{j=2}^N d^2(t_j, t_{j-1}).
\]
If \( d(a, b) \leq r \) and \( d(t_j, t_{j-1}) > \delta \) for each \( j \) then \((N-1)\delta^2 \leq r^2\). Thus

\[
\text{pack}(\delta, [a, b], d) \leq 1 + (r/\delta)^2 \quad \text{for } 0 < \delta \leq r.
\]

To avoid mess, we simplify the bound to \(2(r/\delta)^2\).

### 3.3 Chaining bounds

In this section we consider a generic stochastic process \( \{X(\theta) : \theta \in \Theta\} \) whose increments are controlled by the metric \( d \) in the sense that

\[
\mathbb{P}\{|X(\theta_1) - X(\theta_2)| > d(\theta_1, \theta_2)x\} \leq C_1 e^{-C_2x} \quad \text{for } x \geq 0,
\]

for constants \( C_1, \) and \( C_2 \). We establish a one-sided analog of \(\langle 3 \rangle\) and \(\langle 4 \rangle\),

\[
\mathbb{P}\{\exists \theta \geq \theta_\mu : |X(\theta) - X(\theta_\mu)| > 2L(\theta, x, r)\} \leq C_1 e^{-C_2x} \quad \text{for } x, r > 0
\]

where \( L(\theta, x, r) = [d^2(\theta, \theta_\mu) + r^2] x/r \)

We omit the argument for \( \theta < \theta_\mu \), which is similar.

As explained in Section 2, we actually only need the inequality for \( r \) equal to \( \max(\sqrt{3m\epsilon}, 7x) \), but that choice plays no role in the derivation of \(\langle 17 \rangle\).

The method works by cutting the index set into regions where \( L(\theta, x, r) \) is approximately constant. For a given \( r > 0 \) cover \([\theta_\mu, n]\) by \( \cup_{k=1}^m I_k \) where \( I_k = [a_{k-1}, a_k] \) and \( d^2(a_k, \theta_\mu) = kr^2 \) for \( k = 1, \ldots, m-1 \) and \( d^2(a_m, \theta_\mu) \leq kr^2 \). By \(\langle 15 \rangle\), each \( I_k \) is of \( d \)-diameter at most \( r \). Bound the left-hand side of \(\langle 17 \rangle\) by

\[
\sum_k \mathbb{P}\{\exists \theta \in I_k : |X(\theta) - X(\theta_\mu)| > 2krx\}.
\]

Here we have used the fact that \( d^2(\theta, \theta_\mu) + r^2 \geq kr^2 \) for all \( \theta \) in \( I_k \), with equality at \( \theta = a_{k-1} \). The \( k \)th term in the sum is less than

\[
\mathbb{P}\{|X(a_{k-1}) - X(\theta_\mu)| > kr x\} + \mathbb{P}\{\exists \theta \in I_k : |X(\theta) - X(a_{k-1})| > kr x\}
\]

By inequality \(\langle 16 \rangle\), the first term is less than \( C_1 e^{-C_2\sqrt{krx}} \). The next lemma handles the other contribution. Taken together they give a bound of the form \( \sum_{k \geq 1} C_3 \exp(-C_4 kx) \) for the left-hand side of \(\langle 17 \rangle\). If \( C_4 x \geq 1 \) the sum is bounded by a constant times \( \exp(-C_4 x) \). An increase in the constant \( C_1 \), if necessary, extends the bound to values of \( x \) for which \( C_4 x < 1 \).

**Lemma.** Suppose \( \{Z(t) : t \in T\} \) is a process with continuous sample paths indexed by a set \( T \) equipped with a metric \( d \). Suppose also that

\[
\langle 18 \rangle
\]
(i) The diameter of $T$ is $r$ and the packing numbers satisfy

$$\text{pack}(\delta, T, d) \leq C \left(\frac{r}{\delta}\right)^m \quad \text{for } 0 < \delta \leq r,$$

where $C$ and $m$ are constants.

(ii) The increments of $Z$ are controlled by $d$, in the sense that

$$\mathbb{P}\{|Z(t_1) - Z(t_2)| > \delta d(t_1, t_2)\} \leq C_1 \exp(-C_2 \delta) \quad \text{for all } \delta \geq 0.$$

Then

$$\mathbb{P}\{\sup_{t \in T} |Z(t) - Z(t_0)| > c_i x\} \leq c_2 e^{-x} \quad \text{for all } x \geq 0,$$

for constants $c_i$ depending on $C$ and $m$.

**Proof** Define $T_0 = \{t_0\}$ and construct packing sets $T_1, T_2, \ldots$ with

$$N_i = \#T_i \leq \text{pack}(\delta_i, T, d) \leq C 2^m \quad \text{where } \delta_i = r/2^i.$$

By construction,

$$\min_{t' \in T_i} d(t, t') \leq \delta_i \quad \text{for each } t \in T.$$

Let $\{\gamma_i\}_{i \geq 1}$ be a sequence of positive numbers whose value we will later choose. For simplicity of notation write $R_i = \sum_{j \leq i} \gamma_j$ and $R_\infty = \sum_{j=1}^{\infty} \gamma_j$. Denote $\Delta_i := \sup_{t_i \in T_i} |Z(t_i) - Z(t_0)|$. By continuity of sample paths,

$$\Delta_i \to \Delta := \sup_{t \in T} |Z(t) - Z(t_0)| \quad \text{as } i \to \infty.$$

so that $M_i \to \mathbb{P}\{\Delta > R_\infty\}$. It suffices to bound $M_i := \mathbb{P}\{\Delta_i > R_i\}$.

Define $\psi_i : T_i \to T_{i-1}$ as the function that maps $t_i$ to the element in $T_{i-1}$ that is the closest to $t_i$. Then $\Delta_i \leq \Delta_{i-1} + S_i$ for each $i$, where $S_i = \max_{t \in T_i} |Z(t) - Z(\psi_i t)|$, which implies the recursive bound

$$\mathbb{P}\{\Delta_i > R_i\} \leq \mathbb{P}\{\Delta_{i-1} > R_{i-1}\} + \mathbb{P}\{S_i > \gamma_i\}.$$

Use a union bound to control the second term.

$$\mathbb{P}\{S_i > \gamma_i\} \leq \sum_{t_i \in T_i} \mathbb{P}\{|Z(t_i) - Z(\psi_i t_i)| > \gamma_i\}$$

$$\leq C_1 N_i \exp\left(-C_2 \gamma_i / \delta_i\right)$$

$$\leq CC_1 \exp\left(im \log 2 - C_2 \gamma_i 2^i / r\right)$$
Since we eventually want $\sum_{i \geq 1} P\{ S_i > \gamma_i \}$ to be exponentially small, we choose $\gamma_i$ so that $\exp(im \log 2 - C_2 \gamma_i^2 / r) = \exp(-x)/2^i$, i.e.,

$$
\gamma_i = \frac{r}{C_2} 2^{-i} (i(m + 1) \log 2 + x).
$$

This choice of $\gamma_i$ ensures that the tail probability is small enough, but still we do not want $R_i = \sum_{j \leq i} \gamma_j$ to diverge as $i$ grows. Check

$$
R_i = \sum_{j \leq i} \gamma_j = \frac{r}{C_2} \sum_{j \leq i} [2^{-j} (j(m + 1) \log 2 + x)] \leq C_3 + C_4 x.
$$

Here $C_4$ is a universal constant, and $C_3$ only depends on $m$. When $x \geq 1$, we can absorb $C_3$ into the $C_4 x$ term. In summary,

$$
M_i = P\{ \Delta_i > R_i \} \leq \sum_{j \geq 1} e^{-x} / 2^j = e^{-x}.
$$

If $c_2 = e$ then the upper bound $c_2 e^{-x}$ also covers the $0 < x < 1$ case. Let $i$ go to infinity to complete the proof.

References

Hanson, D. L. and F. T. Wright (1971). A bound on tail probabilities for quadratic forms in independent random variables. *The Annals of Mathematical Statistics* 42(3), pp. 1079–1083.

Kneip, A. (1994). Ordered linear smoothers. *Annals of Statistics* 22(2), 835–866.

Nolan, D. and D. Pollard (1987). U-processes: rates of convergence. *Annals of Statistics* 15, 780–799.

Rudelson, M. and R. Vershynin (2013). Hanson-Wright inequality and sub-gaussian concentration. Technical report, arXiv:1306.2872v3.

Tsybakov, A. (2014, January–April). Stat 681: Nonparametric estimation and statistical learning. Graduate course given at Yale University.