III - Probability Tail Bounds

Math 728 D - Machine Learning & Data Science - Spring 2019

Contents

What is this about? ...

Statements quantifying the probability that certain events exceed some threshhold are called "tail bounds"; examples are Markov's or Chebyshev's inequality or the Law of Large Numbers (see Lecture II, Theorems 11, 12 14.)

While these results hold almost without or with only very little additional structural assumptions on the underlying probability measure, refined estimates can be obtained when more is known about the probability distribution. There are different categories of such structural assumptions important representatives of which will be discussed.

More comments:

- The Central Limit Theorem, (Lecture II, Theorem 20) gives an asymptotic statement that large sums of random variables behave like a Gaussian in the limit $N \to \infty$. The results to be discussed in this Lecture offer quantified bounds for all N.
- The Law of Large Numbers, which in turn is based on Chebyshev's and Markov's inequalities, are the most basic examples of probability tail bounds which, however, ensure only algebraic decay rates of excess probabilities.
- In this lecture we discuss more refined such bounds
  - for (large) sums of random variables;
  - under additional structural assumptions on the underlying probability distribution;
- which exhibit even exponential decay rates.

What are these Bounds Good for?

- In Machine Learning tail bounds help quantifying the extraction of information from large data sets by estimating the probability for a learning algorithm to be approximately correct. Typical bounds quantify the deviation of sample means from the exact expectation.
- They help understanding Concentration of Measure Phenomena in high-dimensional geometry underlying unsupervised learning concepts.
- They are important tools in statistical estimation.
- Countless applications involve sums of random variables, e.g., X_i is the amount of good the ith consumer buys; X_i is the length of the ith message sent over a network; X_i is the indicator random variable of whether the ith record in a large data base has a certain property; etc. Each X_i is modeled by a simple probability distribution like a Gaussian, an exponential ($p(t) = e^{-t}, t > 0$), or a Bernoulli distribution.
- A central issue is therefore to bound the probability of deviations of a (large) sum of random variables (e.g. i.i.d samples) from the expectation of this sum.

Some Simple Consequences of Chebyshev's Inequality

Deviation from the mean: let $X_1, \ldots, X_N$ be independent random variables and $S_n = \sum_{j=1}^{N} X_j$

Chebyshev's inequality (Lecture II, Theorem 14, (9.3)) and Lecture II, Lemma 7, (8.9) imply

$$\mathrm{Prob}\big(|S_N - \mathbb{E}[S_N]| \geq t\big) \leq \frac{\mathrm{var}[S_N]}{t^2} = \frac{\sum_{j=1}^{N} \mathrm{var}[X_j]}{t^2}. \tag{2.1}$$

Setting $\sigma^2 := \frac{1}{N} \sum_{j=1}^{N} \mathrm{var}[X_j]$, replacing t/N by ϵ, this can be rephrased as

$$\mathrm{Prob}\Big(\Big|\frac{1}{N}\sum_{j=1}^{N}(X_j - \mathbb{E}[X_j])\Big| \geq \epsilon\Big) \leq \frac{\sigma^2}{N\epsilon^2}. \tag{2.2}$$

Hence, the average deviation of the random variables from their expectation is controlled by the average variance, the number of variables, and the deviation threshhold.

This can be further quantified once one has bounds for the $\mathrm{var}[X_j]$.

Basic Chernoff Bounds

Chernoff bounds actually represent a method for deriving upper bounds for the probabilities of deviations from the mean.

Since for any $s > 0$, $\phi(x) = e^{sx}$ is positive strictly increasing we have by Markov's inequality (Lecture II, Theorem 12, (9.2))

$$\text{Prob}(X \geq t) = \text{Prob}(e^{sX} \geq e^{st}) \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}} \tag{3.1}$$

see also Lecture II, Exercise 15, (2). Idea: choose s so that the right hand side becomes small.

Application: $X_1, \ldots, X_N$ be independent random variables and $S_N = \sum_{j=1}^{N} X_j$, (3.1) $\rightsquigarrow$

$$\text{Prob}\left(S_N - \mathbb{E}[S_N] \geq t\right) \leq e^{-st}\mathbb{E}\left[e^{s\sum_{j=1}^{N}(X_j - \mathbb{E}[X_j])}\right] = e^{-st}\prod_{j=1}^{N}\mathbb{E}\left[e^{s(X_j - \mathbb{E}[X_j])}\right], \tag{3.2}$$

where we have used independence in the last step, see Lecture II, Lemma 7, (8.8).

To exploit these inequalities further, one needs bounds for expressions like $\mathbb{E}[e^{sX}]$ (see the notion of moment generating function (4.5) introduced later).

Hoeffding's Inequality

(3.2) $\rightsquigarrow$ one needs a bound for $\mathbb{E}\left[e^{sX}\right]$ when $\mathbb{E}[X] = 0$: this works, for instance, when X has a bounded range.

Lemma 1

Let X be a random variable with mean zero $\mathbb{E}[X] = 0$, taking values $a \leq X \leq b$. Then for $s > 0$ one has

$$\mathbb{E}\left[e^{sX}\right] \leq e^{s^2(b-a)^2/8}. \tag{3.3}$$

Proof: Since the exponential function is convex, we have

$$e^{sX} = e^{\frac{b-X}{b-a}sa + \frac{X-a}{b-a}sb} \leq \frac{b-X}{b-a}e^{sa} + \frac{X-a}{b-a}e^{sb}$$

Using $\mathbb{E}[X] = 0$, this gives

$$\mathbb{E}\left[e^{sX}\right] \leq \frac{b}{b-a}e^{sa} + \frac{-a}{b-a}e^{sb} = e^{-u\lambda}((1-\lambda) + \lambda e^{u}), \quad \text{where} \quad \lambda := \frac{-a}{b-a},\ u := s(b-a).$$

One can show (Taylor expansion) that $\log\left(e^{-u\lambda}((1-\lambda) + \lambda e^{u})\right) \leq \frac{u^2}{8}$, monotonicity of the exponential yields (3.3). $\square$

Hoeffding's Inequality

Combining Lemma 1 with (3.2) and choosing $s := \frac{4t}{\sum_{j=1}^{N}(b_j-a_j)^2}$, yields:

Theorem 2

Let $X_1, X_2, \ldots, X_N$ be independent random variables where X_j takes values in the interval $[a_j, b_j]$ with probability one ($P(X_j \notin [a_j, b_j]) = 0$), $S_N := X_1 + \cdots + X_N$. Then one has for any $t > 0$

$$\text{Prob}\big(S_N - \mathbb{E}[S_N] \geq t\big) \leq e^{-\frac{2t^2}{\sum_{j=1}^{N}(b_j-a_j)^2}} \quad \textit{and} \quad \text{Prob}\big(S_N - \mathbb{E}[S_N] \leq -t\big) \leq e^{-\frac{2t^2}{\sum_{j=1}^{N}(b_j-a_j)^2}}. \quad (3.4)$$

Hoeffing's inequality provides tail bounds when the range of the random variables is bounded, so it applies, for instance, to Binomial random variables. A bounded range is also a typical assumption in machine learning contexts and "non-parametric statistical estimation". In this context one is interested in relating (not the sum S_N but) the empirical average $\frac{1}{N}S_N$ to its expectation. Setting $\epsilon := t/N$ in (3.4), yields

$$\text{Prob}\Big(\Big|\frac{S_N}{N} - \mathbb{E}\Big[\frac{S_N}{N}\Big]\Big| \geq \epsilon\Big) \leq e^{-\frac{2\epsilon^2 N^2}{\sum_{j=1}^{N}(b_j-a_j)^2}}. \quad (3.5)$$

A shortcoming of Hoeffding's inequality is that it does not use any information about the variance of the random variables. The following concentration inequalities improve on this aspect and play therefore a central role in analyzing the performance of machine learning algorithms.

Bennett's Inequality

Theorem 3

Let $X_1, \ldots, X_N$ be independent real-valued random variables with zero mean $\mathbb{E}[X_j] = 0$, $j = 1, \ldots, N$. Assume that $X_j \leq 1$ with probability one and set

$$\sigma^2 := \frac{1}{N} \sum_{j=1} \mathbb{E}\big[X_j^2\big]. \tag{3.6}$$

Then for any $t > 0$,

$$\mathrm{Prob}\Big(\sum_{j=1}^{N} X_j > t \Big) \leq \exp\Big\{ - N\sigma^2 h\Big(\frac{t}{N\sigma^2} \Big) \Big\}, \tag{3.7}$$

where $h(u) := (1+u)\log(1+u) - u$ for $u \geq 0$.

Rewriting this again in terms of $S_N := \sum_{j=1}^{N} X_j$, this reads tor non-centered variables $X_j - \mathbb{E}[X_j] \leq 1$

$$\mathrm{Prob}\Big(\frac{S_N}{N} - \mathbb{E}\Big[\frac{S_N}{N} \Big] > \epsilon \Big) \leq \exp\Big\{ - N\sigma^2 h\Big(\frac{\epsilon}{\sigma^2} \Big) \Big\}. \tag{3.8}$$

Proof of Theorem 3: The starting point is again the Chernoff bound (3.2) which requires further estimating $\mathbb{E}\left[e^{sX_j}\right]$. The idea is to write the exponential as a sum of several parts each of which can be estimated well under the given assumptions on the X_j. To that end, introduce the function

$$\psi(x) := e^x - x - 1 = \sum_{k=2}^{\infty} \frac{x^k}{k!}.$$

Observe that

$$\begin{array}{lcll} \psi(x) & \leq & x^2/2 & \text{for} \quad x \leq 0, \\ \psi(sx) & \leq & x^2\psi(s) & \text{for} \quad s \geq 0 \text{ and } x \in [0,1], \\ \psi(s) & \geq & s^2/2 & \text{for} \quad s \geq 0. \end{array} \tag{3.9}$$

Defining $x_+ := \max\{x, 0\}$, $x_- := \max\{-x, 0\}$ and noticing that $\psi(0) = 0$, one can write $\psi(sx) = \psi(sx_+) + \psi(-sx_-)$. Since $e^x = \psi(x) + x + 1$ and $\mathbb{E}$ is a linear functional one obtains

$$\begin{aligned} \mathbb{E}\left[e^{sX_j}\right] &= 1 + s\mathbb{E}[X_j] + \mathbb{E}\left[\psi(sX_j)\right] = 1 + \mathbb{E}\left[\psi(sX_j)\right] \quad (\text{since } \mathbb{E}[X_j] = 0) \\ &= 1 + \mathbb{E}\left[\psi(s(X_j)_+) + \psi(-s(X_j)_-)\right] \\ &\leq 1 + \mathbb{E}\left[\psi(s(X_j)_+) + \frac{s^2}{2}(X_j)_-^2\right] \quad \text{(by the first inequality in (3.9))} \end{aligned}$$

Now use that the $X_j \leq 1$ and invoke the second inequality in (3.9) to conclude

$$\begin{aligned} \mathbb{E}\left[e^{sX_j}\right] &\leq 1 + \mathbb{E}\left[\psi(s)(X_j)_+^2 + \frac{s^2}{2}(X_j)_-^2\right] \leq 1 + \psi(s)\mathbb{E}\left[(X_j)_+^2 + (X_j)_-^2\right] \text{ (by the third inequalityin (3.9))} \\ &= 1 + \psi(s)\mathbb{E}\left[X_j^2\right] \leq \exp\left\{\psi(s)\mathbb{E}\left[X_j^2\right]\right\}. \end{aligned} \tag{3.10}$$

Proof of Theorem 3 continued: Recall the definition of σ^2 in (3.6) and insert (3.10) into (3.2) to obtain (here $\mathbb{E}[X_j] = 0$)

$$\text{Prob}\Big(\sum_{j=1}^{N} X_j > t\Big) \leq e^{-st} \prod_{j=1}^{N} e^{\psi(s)\mathbb{E}\left[X_j^2\right]} = e^{-st+\psi(s)\sum_{j=1}^{N}\mathbb{E}\left[X_j^2\right]} = e^{\psi(s)N\sigma^2 - st}. \tag{3.11}$$

Now we choose s so as to minimize the upper bound, namely

$$s = \log\left(1 + \frac{t}{N\sigma^2}\right). \tag{3.12}$$

Substituting this value in (3.11), yields (3.7). □

Bernstein's Inequality

Further bounding the function h in Bennett's inequality yields another very important inequality.

Theorem 4

Let $X_1, \ldots, X_N$ be independent real-valued random variables with zero mean $\mathbb{E}[X_j] = 0$, $j = 1, \ldots, N$. Assume that $X_j \leq 1$ with probability one. Then for σ^2 as in (3.6) and any $\epsilon > 0$ ($\epsilon \leftrightarrow t/N$)

$$\text{Prob}\Big(\frac{1}{N}\sum_{j=1}^{N} X_j > \epsilon\Big) \leq \exp\Big\{-\frac{N\epsilon^2}{2(\sigma^2 + \epsilon/3)}\Big\}. \tag{3.13}$$

Proof: Verify that $h(u) \geq \frac{u^2}{2+2u/3}$ for $u \geq 0$, by comparing the derivatives of both sides. □

Remark: When $\sigma^2 \leq \epsilon$ the decay is like $e^{-cN\epsilon}$ instead of $e^{-cN\epsilon^3}$ as predicted by Hoeffding's inequality (3.5).

Useful Reformulations of Bernstein's Inequality

Corollary 5

Let $X_1, \ldots, X_N$ be independent real-valued random variables with common mean $\mu = \mathbb{E}[X_j]$ and variance $\sigma^2 = \mathrm{var}[X_j]$, $j = 1, \ldots, N$. Assume that the X_j take values $|X_j - \mu| \leq M$ with probability one. Then

$$\mathrm{Prob}\Big(\Big|\frac{1}{N}\Big(\sum_{j=1}^{N} X_j\Big) - \mu\Big| > \epsilon\Big) \leq 2\exp\Big\{-\frac{N\epsilon^2}{2(\sigma^2 + \epsilon M/3)}\Big\}. \tag{3.14}$$

Proof: Consider the random variable $Y_j := \frac{X_j - \mu}{M}$, $j = 1, \ldots, N$. Then $\mathbb{E}[Y_j] = 0$, $Y_j \leq 1$, $j = 1, \ldots, N$, with probability one. Notice that now

$$\tilde{\sigma}^2 := \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}[Y_j^2] = \frac{1}{M^2 N}\sum_{j=1}^{N}\mathbb{E}[(X_j - \mu)^2] = \frac{1}{M^2 N}\sum_{j=1}^{N}\mathrm{var}[X_j] = \frac{1}{M^2}\sigma^2.$$

Theorem 4 gives then

$$\begin{aligned}\mathrm{Prob}\Big(\Big(\frac{1}{N}\sum_{j=1}^{N} X_j\Big) - \mu > \delta\Big) &= \mathrm{Prob}\Big(\frac{1}{N}\sum_{j=1}^{N}\frac{X_j - \mu}{M} > \frac{\delta}{M}\Big) = \mathrm{Prob}\Big(\frac{1}{N}\sum_{j=1}^{N} Y_j > \frac{\delta}{M}\Big)\\ &\leq \exp\Big\{-\frac{N(\delta/M)^2}{2(\tilde{\sigma}^2 + \delta/(M3))}\Big\} = \exp\Big\{-\frac{N\delta^2}{2(\sigma^2 + \frac{M\delta}{3})}\Big\}.\end{aligned}$$

Since $\mathrm{Prob}(|A| > \delta) = \mathrm{Prob}(A > \delta) + \mathrm{Prob}(-A > \delta)$ the assertion follows. □

Moment Based Tail Bounds

Theorem 6

Let $X_1, X_2, \ldots, X_N$ be mutually independent random variables with zero mean ($\mathbb{E}[X_j] = 0$) and variance at most σ^2 ($\mathrm{var}[X_j] \le \sigma^2$), and let $0 \le a \le \sqrt{2}N\sigma^2$. Assume that for a positive integer m one has

$$|\mathbb{E}[X_j^r]| \le \sigma^2 r!, \quad r = 3, 4, \ldots, m. \tag{4.1}$$

Then one has

$$\mathrm{Prob}\big(|X_1 + \cdots + X_N| \ge a\big) \le \begin{cases} \left(\frac{2mN\sigma^2}{a^2}\right)^{m/2}, & \text{if} \quad m \le N\sigma^2/2, \\ 3e^{-\frac{a^2}{12N\sigma^2}} & \text{if} \quad m \le \lfloor (a^2/6N\sigma^2) \rfloor. \end{cases} \tag{4.2}$$

Suppose the X_j are i.i.d. jointly distributed with mean μ so that the $Y_j := X_j - \mu$ are centered ($\mathbb{E}[Y_j] = 0$). Then, if the Y_j satisfy (4.1), the second relation in (4.2) becomes

$$\mathrm{Prob}\left(\left|\frac{X_1 + \cdots + X_N}{N} - \mu\right| \ge \frac{a}{N}\right) \le 3e^{-\frac{a^2}{12N\sigma^2}} \quad \text{if} \quad m \le \lfloor (a^2/6N\sigma^2) \rfloor. \tag{4.3}$$

or with $a/N =: \delta$

$$\mathrm{Prob}\left(\left|\frac{X_1 + \cdots + X_N}{N}\right| \ge \delta\right) \le 3e^{-\frac{N\delta^2}{12\sigma^2}} \quad \text{if} \quad m \le \lfloor (N\delta^2/6\sigma^2) \rfloor. \tag{4.4}$$

Comments

- Gaussians satisfy these moment bounds, see Lecture II, p. 40, (10.3).
- A systematic way of calculating (estimating) moments of probability distributions is to employ the so called moment generating function which is very similar to the characteristic function (replacing Fourier transforms by Laplace transforms, see Lecture II, (8.12)):

$$M(t) = M(t;p) = \sum_{x\in\mathcal{X}} e^{tx} p(x) \quad \text{or} \quad M(t) = M(t;p) = \int_{\mathcal{X}} e^{tx}\, dP(x). \tag{4.5}$$

In fact, the rth derivative of $M(t;p)$ is

$$M^{(r)}(t;p) = \sum_{x\in\mathcal{X}} x^r e^{tx} p(x) \quad \Rightarrow \quad M^{(r)}(0;p) = \mathbb{E}\big[X^k\big]. \tag{4.6}$$

Example: $X \sim B(1,p)$ (Bernoulli variable) for some $p \in [0,1]$; Lecture II, (10.17) $\rightsquigarrow$

$$M^{(r)}(t) = \sum_{k\in\{0,1\}} k^r e^{tk} p^k (1-p)^{1-k} \quad \rightsquigarrow$$

$$M'(0) = \mathbb{E}[X] = 0 + p = p, \quad \mathbb{E}\big[X^2\big] = M''(0) = 0^2 p^0 (1-p)^1 + 1^2 p (1-p)^0 = p \quad \rightsquigarrow$$

$$\operatorname{var}[X] = \mathbb{E}\big[X^2\big] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

see Lecture II, (10.18).

Proof of Theorem 6: Step 1: derive upper bound for $\mathbb{E}[X^r]$ any even $r \in \mathbb{N}$:

Notation: $\mathbf{r} = (r_1, \ldots, r_N)^\top \in \mathbb{N}_0^N$, $\mathbf{r}! := r_1! r_2! \cdots r_N!$, $|\mathbf{r}| = \|\mathbf{r}\|_1 = r_1 + \cdots + r_N$,
$\mathbf{x}^{\mathbf{r}} := x_1^{r_1} x_2^{r_2} \cdots x_N^{r_N}$;
Multinomial Theorem $\Rightarrow$

$$\Big(\sum_{j=1}^{N} x_j\Big)^r = \sum_{\mathbf{r}\in\mathbb{N}_0^N, |\mathbf{r}|=r} \binom{r}{r_1, r_2, \ldots, r_N} x_1^{r_1} x_2^{r_2} \cdots x_N^{r_N} = \sum_{\mathbf{r}\in\mathbb{N}_0^N, |\mathbf{r}|=r} \frac{r!}{\mathbf{r}!} \mathbf{x}^{\mathbf{r}}.$$

By independence of the X_i (see Lecture II, (7.5) and Lemma 7) for $X := X_1 + X_2 + \cdots + X_N$

$$\mathbb{E}[X^r] = \sum_{\mathbf{r}\in\mathbb{N}_0^N, |\mathbf{r}|=r} \frac{r!}{\mathbf{r}!} \mathbb{E}[X_1^{r_1}] \mathbb{E}[X_2^{r_2}] \cdots \mathbb{E}[X_N^{r_N}]$$

Note: $r_j = 1 \Rightarrow \mathbb{E}[X^{r_1}] = \mathbb{E}[X] = 0$, by assumption. Consider

$$R := \{\mathbf{r} : |\mathbf{r}| = r, \text{ all non-zero } r_j \text{ are greater or equal to } 2\} \quad \rightsquigarrow$$

$$\mathbb{E}[X^r] = \sum_{\mathbf{r}\in R} \frac{r!}{\mathbf{r}!} \mathbb{E}[X_1^{r_1}] \mathbb{E}[X_2^{r_2}] \cdots \mathbb{E}[X_N^{r_N}] \overset{(4.1)}{\leq} r! \sum_{\mathbf{r}\in R} \sigma^{2\#(\text{non-zero entries in } \mathbf{r})}, \tag{4.7}$$

where we have used the moment bound (4.1) in the last step.

Proof of Theorem 6 continued: Since $|\mathbf{r}| = r$ there are at most $r/2$ non-zero entries in each $\mathbf{r} \in R$. We now rearrange the summation properly, collecting all $\mathbf{r} \in R$ with q non-zero entries in the batch B_q.Since q can run from 1 to $r/2$, (4.7) can be reqritten as

$$\mathbb{E}[X^r] \leq r! \sum_{q=1}^{r/2} \sum_{\mathbf{r} \in B_q} \sigma^{2q} = r! \sum_{q=1}^{r/2} \#(B_q)\sigma^{2q}. \tag{4.8}$$

Now we need to count $\#(B_q)$. To that end, fix a subset S_q of q slots from $\{1, 2, \ldots, N\}$ occupied by the non-zero entries of an element $\mathbf{r} \in B_q$. To count the number of such possible candidates, we ask in how many ways we can distribute the total budget r by assigning a sub-budget r_j to each slot j in S_q. Since each $r_j \geq 2$, we can assign to each slot the value 2 which leaves the remaining budget $r - 2q$ yet to bet distributed. In other words in how many ways can we write $r - 2q$ as a sum of q integers (counting positions). This number is known to be $\binom{r-2q+q-1}{q-1}$. Since there are $\binom{N}{q}$ subsets S_q we conclude

$$\sigma^{2q}\#(B_q) = \sigma^{2q}\binom{N}{q}\binom{r-2q+q-1}{q-1} = \sigma^{2q}\binom{N}{q}\binom{r-q-1}{q-1} =: n(q),$$

and hence from (4.8)

$$\mathbb{E}[X^r] \leq r! \sum_{q=1}^{r/2} n(q). \tag{4.9}$$

Proof of Theorem 6 continued: To bound next $n(q)$, notice first that $\binom{N}{q} \leq N^q/q!$. Since $\binom{r-q-1}{q-1}$ is the number of subsets of cardinality $q-1$ of a set of cardinality $r-q-1$ and hence less than or equal to the number of all subsets of a set of cardinality $r-q-1$ which is 2^{r-q-1}. Hence,

$$n(q) \leq \frac{(\sigma^2 N)^q}{q!} 2^{r-q-1} =: u(q). \tag{4.10}$$

When $r \leq m \leq N\sigma^2/2$ and hence $q \leq r/2 \leq N\sigma^2/4$ (see the first inequality in (4.2)) we have

$$\frac{u(q)}{u(q-1)} = \frac{\sigma^2 N}{2q} \geq 2 \quad \Rightarrow \quad u(q-1) \leq \frac{u(q)}{2}.$$

Now we infer from (4.9) and (4.10) that

$$\mathbb{E}[X^r] \leq r! \sum_{q=1}^{r/2} u(r/2)\Big(1 + \frac{1}{2} + \cdots + \frac{1}{4} + \cdots\Big) \leq 2r!u(r/2) \overset{(4.10)}{=} \frac{r!}{(r/2)!} 2^{r/2} (N\sigma^2)^{r/2} \tag{4.11}$$

Step 2: Now we can apply Markov's inequality (Lecture II, Theorem 12, (9.2)) and the fact that r is even to conclude

$$\begin{aligned} \text{Prob}(|X| \geq a) &= \text{Prob}(X^r \geq a^r) \leq \frac{\mathbb{E}[X^r]}{a^r} \leq \frac{r!}{(r/2)!}\Big(\frac{2N\sigma^2}{a^2}\Big)^{r/2} := g(r) \leq r^{r/2}\Big(\frac{2N\sigma^2}{a^2}\Big)^{r/2} \\ &= \Big(\frac{2rN\sigma^2}{a^2}\Big)^{r/2}. \end{aligned} \tag{4.12}$$

Applying this for $r = m$, yields the first inequality in (4.2).

Proof of Theorem 6 continued: Regarding he second inequality in (4.2), notice that for $g(r)$ defined in (4.12),

$$\frac{g(r)}{g(r-2)} = \frac{4(r-1)N\sigma^2}{a^2},$$

i.e., as long as this quantity is less than one, which means $r - 1 \leq \frac{a^2}{4N\sigma^2}$, $g(r)$ decreases. This holds, in particular, for $r = m = \lfloor a^2/(6N\sigma^2) \rfloor$. Substituting this value into the right hand side of (4.12), yields

$$\text{Prob}(|X| \geq a) \leq 3^{-m/2} \leq e^{-m/2} \leq e \cdot e^{-a^2/(12N\sigma^2)} \leq 3e^{-a^2/(12N\sigma^2)},$$

as claimed. □