A Brief Survey On Concentration Inequalities

Nischal Bhattarai

March 2025

1 Introduction

Concentration inequalities provide bounds on the tail probability of a real-valued random variable, typically illustrating how quickly the variable converges toward its mean. The rate at which this tail probability decreases depends on factors such as the random variable's degree of integrability. This study begins with a fundamental inequality and progressively refines it by incorporating additional assumptions.

2 Basic Inequalities

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and X a real-valued random variable defined on this space. A natural question arises: How do we estimate $\mathbb{P}(|X| > t)$ to an acceptable degree? This inquiry leads us to one of the cornerstone results in probability theory.

Theorem 2.1 (**Markov's Inequality**). *Let $X \geq 0$ be a non-negative random variable. Then, for any $t > 0$,*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

Proof. Let $E = \{X \geq t\}$ be the set where X exceeds t. Observe that, for all $\omega \in \Omega$,

$$t \cdot 1_E \leq X.$$

By the monotonicity and linearity of the expectation, we have

$$t \cdot \mathbb{E}[1_E] \leq \mathbb{E}[X],$$

noting that the expectation may be infinite. Since $\mathbb{E}[1_E] = \mathbb{P}(X \geq t)$, rearranging yields

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}.$$

□

Corollary 2.1 (**Markov's Inequality for p-Moments**). *Let X be a real-valued random variable such that $\mathbb{E}[|X|^p] < \infty$ for some $p \geq 1$. Then, for any $t > 0$,*

$$\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}[|X|^p]}{t^p}.$$

Proof. Since the function $x \mapsto x^p$ is increasing on $[0, \infty)$ for $p \geq 1$, if $|X| \geq t$, then $|X|^p \geq t^p$. Thus, $\mathbb{P}(|X| \geq t) \leq \mathbb{P}(|X|^p \geq t^p)$. Define $Y = |X|^p$, which is non-negative. Applying Markov's Inequality (Theorem 2.1) to Y, we get

$$\mathbb{P}(Y \geq t^p) \leq \frac{\mathbb{E}[Y]}{t^p} = \frac{\mathbb{E}[|X|^p]}{t^p}.$$

Hence, $\mathbb{P}(|X| \geq t) \leq \frac{\mathbb{E}[|X|^p]}{t^p}$, as required. □

When $p = 1$, this inequality provides the coarsest tail probability estimate. A limitation of Markov's Inequality becomes evident when $t \leq \mathbb{E}[X]$, yielding a trivial bound. One should also note that assuming higher integrability (e.g., $p > 1$) results in a polynomial decay rate of t^p. Before proceeding to tighter bounds, we introduce a key concept that will be instrumental in later sections.

Definition 2.1 (**Moment Generating Function**). *The moment generating function (m.g.f) of a random variable X, denoted $M_X(t)$, is defined as*

$$M_X(t) = \mathbb{E}[e^{tX}],$$

where $t \in \mathbb{R}$, provided the expectation exists (i.e., $M_X(t) < \infty$) for some interval containing 0.

Now, we derive the another important corollary of Markov inequality for exponentially fast bounds.

Corollary 2.2 (**Exponential Tail Bounds**). *Let X be a real-valued random variable such that $M_X(\lambda) < \infty$ for some $\lambda > 0$. Then, we have that for any $t \in \mathbb{R}$,*

$$\mathbb{P}(X \geq t) \leq \frac{M_X(\lambda)}{e^{\lambda t}}.$$

Note: X needn't be positive and t can be any real number.

Proof. We know $\phi(x) = e^{\lambda x}$ is a family of increasing non-zero positive function for $\lambda > 0$. So, we have that,

$$\begin{aligned}\mathbb{P}(X \geq t) &\leq \mathbb{P}(\phi(X) \geq \phi(t)) \\ &\leq \frac{\mathbb{E}[\phi(X)]}{\phi(t)} \\ &= \frac{M_X(\lambda)}{e^{\lambda t}}\end{aligned}$$

where first step follows from Markov's Inequality (Theorem 2.1) to $Y = \phi(X)$ which is positive. Note that ϕ can be any nonzero increasing family of function. □

A further improvement for the exponential bound can be obtained from Cramér–Chernoff method which further optimizes the exponential bound

Corollary 2.3 (Cramér–Chernoff Bounds). *Let X be a real-valued random variable such that $M_X(\lambda) < \infty$ for some $\lambda > 0$. Then, for any $t \in \mathbb{R}$,*

$$\begin{aligned}\mathbb{P}(X \geq t) &\leq \inf_{\lambda \geq 0} \frac{M_X(\lambda)}{e^{\lambda t}} \\ &\leq e^{-\psi_X^*(t)},\end{aligned}$$

where $\psi_X^*(t) = \sup_{\lambda \geq 0}(\lambda t - \ln(M_X(\lambda))$ is called the Cramér transform of X.

Proof. The most natural choice for optimizing the exponential bound is to optimize over λ for each t on the right hand since the inequality holds for all $\lambda \geq 0$ where $M_X(t)$ exists. Now, we observe that if,

$$\begin{aligned}\frac{M_X(\lambda)}{e^{\lambda t}} &= e^{\ln(\frac{M_X(\lambda)}{e^{\lambda t}})} \\ &= e^{-(\lambda t - \ln(M_X(\lambda))}\end{aligned}$$

since $x \mapsto e^{-x}$ is an decreasing function, we have that,

$$\begin{aligned}\inf_{\lambda \geq 0} \frac{M_X(\lambda)}{e^{\lambda t}} &= e^{-\sup_{\lambda \geq 0}(\lambda t - \ln(M_X(\lambda))} \\ &= e^{-\psi_X^*(t)}\end{aligned}$$

which gives the desired result. □

Example 2.1 (Chernoff Bounds for Standard Normal Distribution). *Let X be a normal random variable with $\mathbb{E}[X] = 0$ and $\mathrm{Var}(X) = \sigma^2$, then we the following optimal exponential bound as,*

$$\mathbb{P}(X \geq t) \leq e^{\frac{-t^2}{2\sigma^2}}$$

Proof. We have that the m.g.f is given by,

$$M_X(t) = e^{\frac{\sigma^2 t^2}{2}}$$

then we can define the following function,

$$\begin{aligned}f_t(\lambda) &= \lambda t - \ln(M_X(t)) \\ &= \lambda t - \frac{\sigma^2 \lambda^2}{2}\end{aligned}$$

which is smooth function. Hence, the minimum occurs for each at,

$$\lambda_0 = \frac{t}{\sigma^2}$$

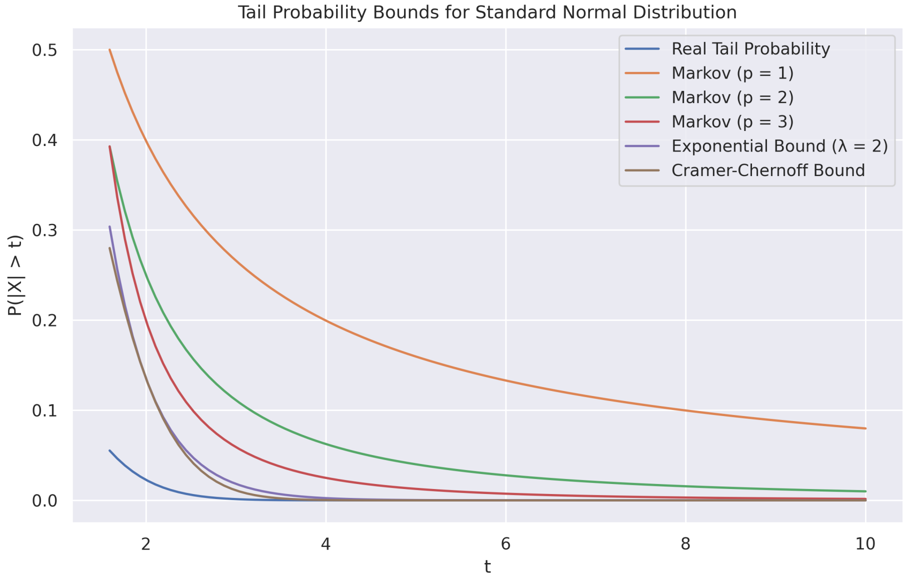

Figure 1: Tail bounds for standard normal using different inequalities.

So,

$$f_t(\lambda_0) = \frac{t^2}{2\sigma^2}$$

substituting in the Cramér-Chernoff (2.3), we have the desired result. □

Example 2.2 (**Chernoff Bounds for Sums of Independent Random Variables**)**.** *Let X be a real-valued random variable such that $M_X(\lambda) < \infty$ for some $\lambda > 0$ and $Z = \sum_{i=1}^n X_i$ be the sum. The the tail bound is given by,*

$$\mathbb{P}(\sum_{i=1}^n X_i \geq t) \leq e^{-n\psi_X^*(\frac{t}{n})}$$

where $\psi_X^*(t)$ is the Cramér transform of X.

Proof. The key is to observe that the m.g.f for the sum is given by:

$$M_Z(t) = \prod_{i=1}^n M_{X_i}(t) = M_X(t)^n$$

So the Cramér transform is given by,

$$\begin{aligned}\psi_Z^*(t) &= \sup_{\lambda\geq 0}(\lambda t - \ln(M_Z(\lambda)) \\ &= \sup_{\lambda\geq 0}(\lambda t - \ln(M_X(t)^n) \\ &= \sup_{\lambda\geq 0}(\lambda t - n\ln(M_X(t)) \\ &= n\sup_{\lambda\geq 0}(\lambda \frac{t}{n} - \ln(M_X(t)) \\ &= n\psi_X^*(\frac{t}{n})\end{aligned}$$

The results follows by substituting above in Cramér-Chernoff (2.3). This result shows the relative ease to calculate bounds for the sum of i.i.d random variables which is the main use case of Chernoff's transform although it is always weaker than moment bound as shown by the next theorem. □

Theorem 2.2 (Moment Bounds Are Always Better than Chernoff Bounds). *Let $X \geq 0$ be a non-negative random variable. Then, for any $t > 0$,*

$$\inf_n \mathbb{E}[X^n]t^{-n} \leq \inf_{\lambda>0} \mathbb{E}[e^{\lambda(X-t)}]$$

where the right hand side is the best Cramér-Chernoff Bound on X.

Proof. Observe that if $a_i \geq 0$ and $b_i > 0$ such that $\sum_i a_i < \infty$ and $\sum_i b_i < \infty$ and $c \leq \frac{a_i}{b_i}$, then we have that,

$$b_i c \leq a_i$$

summing over i we obtain,

$$c \leq \frac{\sum_i a_i}{\sum_i b_i}$$

Now we observe that,

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-t)}] &= \mathbb{E}[\frac{e^{\lambda X}}{e^{\lambda t}}] \\ &= \frac{\mathbb{E}[\sum_i \frac{\lambda^i X^i}{i!}]}{\sum_i \frac{\lambda^i t^i}{i!}} \\ &= \frac{\sum_i \frac{\lambda^i \mathbb{E}[X^i]}{i!}}{\sum_i \frac{\lambda^i t^i}{i!}}\end{aligned}$$

where the third line follows from monotone convergence theorem. Now let $c = \inf_i \mathbb{E}[X^i]t^{-i}$. Hence, the result follows from the above theorem. □

3 Hoeffding's Inequality

The next result provides tail bounds for class of bounded random variables and their independent sums.

Lemma 3.1 (**Hoeffding's Lemma**). *Let X be a random variable with $a \leq X \leq b$ almost surely, then*

$$\mathbb{E}[e^{t(X-\mathbb{E}[X])}] \leq e^{\frac{t^2(b-a)^2}{8}}$$

Proof. With loss of generality, $\mathbb{E}[X] = 0$ and $a \leq 0 \leq b$. First note that $x \mapsto e^{tx}$ is a convex function in $[a, b]$, so we have that for any $x \in [a, b]$,

$$e^{tx} \leq \frac{b-x}{b-a}e^{ta} + \frac{x-a}{b-a}e^{tb}$$

Since X is supported in $[a, b]$, we have by monotonicity,

$$\begin{aligned} \mathbb{E}[e^{tX}] &\leq \frac{b-\mathbb{E}[X]}{b-a}e^{ta} + \frac{\mathbb{E}[X]-a}{b-a}e^{tb} \\ &= \frac{b}{b-a}e^{ta} - \frac{a}{b-a}e^{tb} \\ &= e^{L(t(b-a))} \end{aligned}$$

where,

$$L(h) = \frac{ha}{b-a} + \ln\left(1 + \frac{a - e^{ha}}{b-a}\right)$$

The function L has the following derivatives,

$$L(0) = L^{'}(0) = 0$$

and

$$L^{''}(h) = -\frac{abe^h}{(b-ae^h)^2}$$

Now applying AM-GM inequality with $x = -ae^h \geq 0$ and $y = b \geq 0$, we see that the second derivative is bounded by,

$$L^{''}(h) = -\frac{abe^h}{(b-ae^h)^2} \leq \frac{1}{4}$$

Now the Taylor's theorem on the second derivative and above bound yields,

$$L(h) = L(0) + hL^{'}(0) + \frac{1}{2}h^2L^{''}(h\theta) \leq \frac{1}{8}h^2$$

where $\theta \in [0, 1]$. Now using this on the first result yields,

$$\mathbb{E}[e^{tX}] = e^{L(t(b-a))} \leq e^{\frac{t^2(b-a)^2}{8}}$$

which proves the lemma. □

In simple terms, this lemma states for any bounded random variable, the m.g.f of X is bounded by the m.g.f of a centered normal variable with Var($\mathcal{N}$) = $\frac{(b-a)^2}{4}$. This has the direct influence in the Chernoff bound since,

$$\begin{aligned}\psi_X^*(t) &= \sup_{\lambda\geq 0}(\lambda t - \ln(M_X(\lambda)) \\ &\geq \sup_{\lambda\geq 0}(\lambda t - \ln(M_{\mathcal{N}}(\lambda)) \\ &\geq \psi_{\mathcal{N}}^*(t)\end{aligned}$$

So from Exercise (2.1) and Corollary (2.3), it follows that,

$$\mathbb{P}(X \geq t) \leq e^{\frac{-2t^2}{(b-a)^2}}$$

Actually due to symmetry of m.g.f of normal variable we get,

$$\max\{\mathbb{P}(X \geq t)\,\mathbb{P}(X \leq t)\} \leq e^{\frac{-2t^2}{(b-a)^2}}$$

Such variables whose tail probability is bounded by tail probability of normal distribution is known as sub-gaussian random variable. In the above case we, say that the variable X belongs to the family $\mathcal{G}(\frac{(b-a)^2}{4})$ i.e sub-gaussian family with specified variance. Now the next theorem generalizes the above to sums of independent bounded random variables.

Theorem 3.1 (**Hoeffding's Inequality**). *Let $X_1, \cdots, X_n$ be independent bounded random variable such that $\forall i\ X_i \in [a_i, b_i]$ almost surely. Let,*

$$S = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$$

Then for every $t > 0$, we have

$$\mathbb{P}[S \geq t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i-a_i)^2}}$$

Proof. Now using the Hoeffding's Lemma (3.1) and independence, we observe that the m.g.f is bounded by,

$$\begin{aligned}\ln M_S(t) &= \sum_{i=1}^{n}\ln M_{X_i-\mathbb{E}[X_i]}(t) \\ &\leq \sum_{i=1}^{n}\frac{t^2(b_i-a_i)^2}{8} \\ &= \frac{t^2\sum_{i=1}^{n}(b_i-a_i)^2}{8}\end{aligned}$$

So, the m.g.f of S is bounded by the m.g.f of normal with variance $\sigma^2 = \frac{\sum_{i=1}^{n}(b_i - a_i)^2}{4}$. Hence, the random variable S belongs to the sub-gaussian family $\mathcal{G}(\frac{\sum_{i=1}^{n}(b_i - a_i)^2}{4})$, which has the following tail bound,

$$\mathbb{P}[S \geq t] \leq e^{\frac{-2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

hence the result follows. □

Example 3.1 (**Hoeffding's Inequality to Rademacher Random Variables**). *Let X_i be a i.i.d random variable such that $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$ and $S = \sum_{i=1}^{n} \alpha_i X_i$ where $\alpha_i \in \mathbb{R}$. Then we have for $t > 0$,*

$$\mathbb{P}(S \geq t) \leq e^{\frac{-t^2}{2\,\mathrm{Var}(S)}}$$

Proof. We note that in this case we have $\mathrm{Var}(S) = \sum_{i=1}^{n} \alpha_i^2$ and applying Hoeffding's Inequality (3.1) to S gives the desired result. □

Note that in general case, the variance may be much smaller that the sum in the denominator - take centered uniform distribution on $[0, 1]$. Hence, the example is a special case where we obtain normal like tail bounds using the variance of the random variable instead of using the crude bounds. A natural question is that under similar conditions can one get sharper bounds using variance.

4 Bennett's Inequality

Before we get to the next improvement, we calculate the tail bounds for Poisson random variable.

Lemma 4.1 (**Cramér transform of Poisson Variable**). *Let X be Poisson (μ) random variable and define $Y = X - \mu$, the we have that*

$$\ln M_Y(t) = \mu\phi(t)$$

where $\phi(t) = e^t - 1 - t$. Moreover, the Cramér transform (ψ_Y^*) is given by,

$$\psi_Y^*(t) = \mu h(\frac{t}{\mu})$$

where $h(t) = (1 + t)\ln(1 + t) - t$.

Proof. Now from direct calculation, we observe that,

$$\begin{aligned}\ln M_Y(t) = \ln \mathbb{E}[e^{tY}] = \ln \sum_{k=0}^{\infty} \frac{e^{tk-t\mu}\mu^k e^{-\mu}}{k!} \\ &= -t\mu - \mu + \ln \sum_{k=0}^{\infty} \frac{(e^t\mu)^k}{k!} \\ &= -t\mu - \mu + \ln e^{e^t\mu} \\ &= \mu(e^t - 1 - t) \\ &= \mu\phi(t)\end{aligned}$$

Now, define $\Phi(t, \lambda) = \lambda t - \ln M_Y(\lambda)$, then we observe,

$$\begin{aligned}\frac{\partial}{\partial \lambda}\Phi(t,\lambda) &= \frac{\partial}{\partial \lambda}(\lambda t - \ln M_Y(\lambda)) \\ &= \frac{\partial}{\partial \lambda}(\lambda t - \mu\phi(\lambda)) \\ &= t - \mu(e^{\lambda} - 1)\end{aligned}$$

So the minimum occurs at $\lambda_m = \ln(\frac{t}{u} + 1)$. Hence, we have,

$$\begin{aligned}\psi_Y^*(t) = \Phi(t, \lambda_m) &= \ln(\frac{t}{\mu} + 1)t - \mu(\frac{t}{\mu} + 1 - 1 - \ln(\frac{t}{\mu} + 1)) \\ &= \mu h(\frac{t}{\mu})\end{aligned}$$

where $h(t) = (1+t)\ln(1+t) - t$. Now using Cramér-Chernoff (2.3) bound, we also see that the for $t > -1$,

$$\mathbb{P}(Y \geq t) \leq e^{-\mu h(\frac{t}{\mu})}$$

□

Theorem 4.1 (**Bennett's Inequality**). *Let X_i be independent random variables such that $|X_i - \mathbb{E}[X_i]| \leq b$ for some $b > 0$ almost surely. Let*

$$S_n = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$$

and $\sigma^2 = \sum_{i=1}^{n} \mathrm{Var}(X_i)$. Then, for $t > 0$, we have that,

$$\mathbb{P}(S_n \geq t) \leq e^{\frac{-\sigma^2}{b^2}h(\frac{bt}{\sigma^2})}$$

where $h(t) = (1+t)\ln(1+t) - t$.

Proof. Without loss of generality, assume $\mathbb{E}[X_i] = 0$. Then we note the following,

$$\begin{aligned}
M_{X_i}(\lambda) = \mathbb{E}[e^{\lambda X_i}] &= \mathbb{E}[\sum_{i=1}^{\infty} \frac{(\lambda X_i)^n}{n!}] \\
&= \sum_{i=1}^{\infty} \frac{\mathbb{E}[(\lambda X_i)^n]}{n!} \\
&= 1 + \sum_{i=2}^{\infty} \frac{\mathbb{E}[(\lambda X_i)^n]}{n!} \\
&\leq 1 + \sum_{i=2}^{\infty} \frac{(\lambda b)^{n-2}\, \mathbb{E}[(\lambda X_i)^2]}{n!} \\
&\leq 1 + \lambda^2 \operatorname{Var}(X_i) \sum_{i=2}^{\infty} \frac{(\lambda b)^{n-2}}{n!}
\end{aligned}$$

where the expectation is taking inside using dominated convergence theorem since $|e^{\lambda X_i}| \leq e^{\lambda b}$. Now, we note the log-m.g.f of the S_n is given by,

$$\begin{aligned}
\ln M_{S_n}(\lambda) &= \sum_{j=1}^{n} \ln M_{X_i}(\lambda) \\
&\leq \sum_{j=1}^{n} \ln(1 + \lambda^2 \operatorname{Var}(X_i) \sum_{i=2}^{\infty} \frac{(\lambda b)^{n-2}}{n!}) \\
&\leq \sum_{j=1}^{n} \lambda^2 \operatorname{Var}(X_i) \sum_{i=2}^{\infty} \frac{(\lambda b)^{n-2}}{n!} \\
&= \frac{\sigma^2}{b^2} \sum_{i=2}^{\infty} \frac{(\lambda b)^n}{n!} \\
&= \frac{\sigma^2}{b^2} (e^{b\lambda} - 1 - b\lambda) \\
&= \frac{\sigma^2}{b^2} \phi(\lambda b)
\end{aligned}$$

where we used the fact that $\ln(1 + x) \leq x$ and $\phi(t) = e^t - 1 - t$. Now observe that the Cramér transform of S_n is related by,

$$\begin{aligned}
\psi^*_{S_n}(t) &= \sup_{\lambda \geq 0}(\lambda t - \ln(M_{S_n}(\lambda)) \\
&\geq \sup_{\lambda \geq 0}(\lambda t - \frac{\sigma^2}{b^2}\phi(\lambda b)) \\
&= \frac{\sigma^2}{b^2}\sup_{\lambda \geq 0}(\lambda b \frac{tb}{\sigma^2} - \phi(\lambda b)) \\
&= \frac{\sigma^2}{b^2}\sup_{\lambda b \geq 0}(\lambda b \frac{tb}{\sigma^2} - \phi(\lambda b)) \\
&= \frac{\sigma^2}{b^2} h(\frac{tb}{\sigma^2})
\end{aligned}$$

hence the result follows from Cramér-Chernoff (2.3) and Lemma (4.1) with $\mu = 1$. □

5 Bernstein's Inequality

Using the following lemma, we can get a slightly worst estimate at the cost of compact expression.

Lemma 5.1. *Let $h(t) = (1+t)\ln(1+t) - t$, then for $t \geq 0$, we have that*

$$h(t) \geq \frac{t^2}{2(1+\frac{t}{3})}$$

Proof. Rearranging the terms, it equivalent to showing for $t \geq 0$,

$$s(t) = \ln(t+1) - \frac{5t^2+6t}{2(t^2+4t+3)} \geq 0$$

Now taking the derivative, we have

$$s^{'}(t) = \frac{t^3}{(t^2+4t+3)^2}$$

Hence, the derivative is positive for all $t \geq 0$. Hence, by fundamental theorem of calculus, we have,

$$s(t) \geq s(0) = 0$$

□

Theorem 5.1 (**Bernstein Inequality**). *Let X_i be independent random variables such that $|X_i - \mathbb{E}[X_i]| \leq b$ for some $b > 0$ almost surely. Let*

$$S_n = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$$

and $\sigma^2 = \sum_{i=1}^n \operatorname{Var}(X_i)$*. Then, for* $t > 0$*, we have that,*

$$\mathbb{P}(S_n \geq t) \leq e^{\frac{-t^2}{2(\sigma^2 + \frac{bt}{3})}}$$

Proof. This immediately follows from Lemma (5.1) and Bennett's Inequality (4.1). □

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 02 2013.

[Dur19] Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.