Chernoff Inequalities

Andreas Klappenecker

Texas A&M University

Recall the following basic tail inequality.

Theorem (Markov's Inequality)

If X is a nonnegative random variable and t a positive real number, then

$$\Pr[X \geqslant t] \leqslant \frac{\mathrm{E}[X]}{t}.$$

Proof.

Let Y denote the indicator random variable of the event $X \geqslant t$, so

$$Y(\omega) = \begin{cases} 1 & \text{if } X(\omega) \geqslant t, \\ 0 & \text{if } X(\omega) < t. \end{cases}$$

The expectation value of X satisfies

$$\mathrm{E}[X] \geqslant \mathrm{E}[tY] = t\,\mathrm{E}[Y] = t\,\mathrm{Pr}[X \geqslant t],$$

which proves the claim. □

Definition

The **variance** $\mathrm{Var}[X]$ of a discrete random variable X is defined by

$$\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2,$$

whenever this expression is well-defined. The variance measures the squared deviation from the expected value $\mathrm{E}[X]$.

As a consequence, we obtain the following concentration inequality.

Theorem (Chebychev's inequality)

If X is a random variable, then

$$\Pr[(X - E[X])^2 \geqslant \beta] \leqslant \frac{\mathrm{Var}[X]}{\beta}. \tag{1}$$

Proof.

Given the random variable X, we can define the new random variables $Y = X - \mathrm{E}[X]$ and Y^2. Since Y^2 is a nonnegative random variable, Markov's inequality shows that

$$\Pr[Y^2 \geqslant \beta] \leqslant \frac{\mathrm{E}[Y^2]}{\beta}.$$

Since $\mathrm{E}[Y^2] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{Var}(X)$, we have

$$\Pr[(X - \mathrm{E}[X])^2 \geqslant \beta] = \Pr[Y^2 \geqslant \beta] \leqslant \frac{\mathrm{E}[Y^2]}{\beta} = \frac{\mathrm{Var}[X]}{\beta}. \quad \square$$

We can reformulate this in the following way.

Corollary (Chebychev's inequality)

If X is a random variable, then

$$\Pr[|X - E[X]| \geqslant t] \leqslant \frac{\mathrm{Var}[X]}{t^2}. \tag{2}$$

Proof.

$$\Pr[|X - E[X]| \geqslant t] = \Pr[(X - \mathrm{E}[X])^2 \geqslant t^2] \leqslant \frac{\mathrm{Var}[X]}{t^2}. \quad \square$$

Consider now the sum of n independent random variables X_k,

$$X = X_1 + X_2 + \cdots + X_n.$$

What kind of bound can we get in this case?

Can we improve on Chebychev's inequality in this case?

Proposition

Let $X_1, X_2, \ldots, X_n$ be independent random variables. Then for real numbers $t > 0$, we get

$$\Pr\left[\left|\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mathrm{E}[X_k]\right| \geqslant t\right] \leqslant \frac{\sum_{k=1}^{n} \mathrm{Var}[X_k]}{t^2}.$$

Proof.

We can apply Chebychev's inequality to $\sum_{k=1}^{n} X_k$. Since the random variables X_k are independent, the variance satisfies

$$\mathrm{Var}\left[\sum_{k=1}^{n} X_k\right] = \sum_{k=1}^{n} \mathrm{Var}[X_k]. \qquad \square$$

Corollary

Let $X_1, X_2, \ldots, X_n$ be independent identically distributed random variables with $\mu = \mathrm{E}[X_k]$. Then the average of these random variables satisfies

$$\Pr\left[\left|\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\right| \geqslant t\right] \leqslant \frac{\mathrm{Var}[X_1]}{nt^2}.$$

Corollary (Khinchin)

Let $X_1, X_2, \ldots$ be independent identically distributed random variables with $\mu = \mathrm{E}[X_k]$. Then the average of these random variables satisfies

$$\lim_{n \to \infty} \Pr \left[\left| \frac{1}{n} \sum_{k=1}^{n} X_k - \mu \right| \geqslant t \right] = 0.$$

In other words, the sample average converges in probability towards the expected value μ.

Chernoff Bound

Consider now the sum of n independent bounded random variables X_k,

$$X = X_1 + X_2 + \cdots + X_n,$$

What kind of bound can we get in this case?

Murphy's Law

Anything that can go wrong, does.

Let $A_1, A_2, \ldots, A_n$ denote mutually independent "bad" events. Murphy's law is the folk-wisdom that some bad event is likely to happen.

Let X_k denote the indicator random variable for the event A_k. So

$$\Pr[X_k = 1] = p_k = \Pr[A_k].$$

Let $X = \sum_{k=1}^{n} X_k$ denote number of bad events happening.

How likely is it that no "bad" event will happen?

Proposition

$$\Pr[X = 0] \leqslant e^{-\mathrm{E}[X]}.$$

Proof.

$$\begin{aligned}
\Pr[X = 0] &= \Pr[\overline{A_1 \cup A_2 \cup \cdots \cup A_n}] \\
&= \Pr[\overline{A_1} \cap \overline{A_2} \cap \cdots \cap \overline{A_n}] \\
&= \prod_{k=1}^{n} (1 - \Pr[A_k]) \\
&\leqslant \prod_{k=1}^{n} e^{-\Pr[A_k]} = e^{-\sum_{k=1}^{n} \Pr[A_k]} = e^{-\sum_{k=1}^{n} \mathrm{E}[X_k]} = e^{-\mathrm{E}[X]}.
\end{aligned}$$

□

A production facility for micro processors boasts that the probability that a transistor will be faulty is merely 10^{-5}. You plan to produce a VLSI circuit with 10^6 transistors. Thus, the probability that you get a correctly working micro processor is less than

$$e^{-10} \approx 0.000045$$

Observation

The main reason why we obtained an improvement by going from Markov's inequality to Chebychev's inequality was that we considered a function of the random variable X.

We were able to use the higher moment X^2 to improve the accuracy in the bound.

If X is an arbitrary random variable, a and t real numbers with $t > 0$, then

$$\begin{aligned}\Pr[X \geqslant a] &= \Pr[e^{tX} \geqslant e^{ta}] \\ &\leqslant \frac{E[e^{tX}]}{e^{ta}}\end{aligned}$$

by Markov's inequality. We can now choose t to minimize the right-hand side.

The bounds depend on the random variables and on the value that we choose for t. There are many different Chernoff bounds as a result!

Proposition

Let $X_1, X_2, \ldots, X_n$ be independent random variables with values in the interval $[0, 1]$. If $X = X_1 + X_2 + \cdots + X_n$ and $\mathsf{E}[X] = \mu$, then for every $a > 0$ we get the bounds

1. $\Pr[X \geqslant \mu + a] \leqslant e^{-a^2/2n}$,
2. $\Pr[X \leqslant \mu - a] \leqslant e^{-a^2/2n}$.

The random variables do not need to be Bernoulli random variables, but they need to be independent.

Proof.

Let $Y_k = X_k - \mathrm{E}[X_k]$ for k in the range $1 \leqslant k \leqslant n$.

Then the mean $\mathrm{E}[Y_k] = 0$.

Set $Y = Y_1 + Y_2 + \cdots + Y_n$. Then $Y = X - \mu$.

We can now apply the Bernstein's idea. For some $t > 0$, we get

$$\Pr[X \geqslant \mu + a] = \Pr[Y \geqslant a] = \Pr[e^{tY} \geqslant e^{ta}] \leqslant \frac{\mathrm{E}[e^{tY}]}{e^{ta}}.$$

Proof. (Continued)

We have

$$\Pr[X \geqslant \mu + a] \leqslant \frac{\mathrm{E}[e^{tY}]}{e^{ta}} = \frac{\mathrm{E}[e^{\sum_{k=1}^{n} tY_k}]}{e^{ta}} = \frac{\mathrm{E}\left[\prod_{k=1}^{n} e^{tY_k}\right]}{e^{ta}}.$$

Since the random variables Y_k are independent, it follows that the random variables e^{tY_k} are independent. Therefore, we have

$$\Pr[X \geqslant \mu + a] \leqslant \frac{\mathrm{E}\left[\prod_{k=1}^{n} e^{tY_k}\right]}{e^{ta}} = e^{-at} \prod_{k=1}^{n} \mathrm{E}\left[e^{tY_k}\right]$$

Proof. (Continued)

It remains to bound $\mathrm{E}[e^{tY_k}]$.

The function $f(y) = e^{ty}$ is convex, since $f''(y) = t^2 e^{ty} > 0$.

Let $c + dy$ be the line through the points $(-1, e^{-t})$ and $(1, e^{t})$. So the coefficients c and d must satisfy

$$c = \frac{e^t + e^{-t}}{2} \quad \text{and} \quad d = \frac{e^t - e^{-t}}{2}.$$

By convexity of $f(y)$, we have

$$e^{ty} = f(y) \leqslant c + dy$$

for all y in $[-1, 1]$.

Proof. (Continued)

Therefore, we can conclude that

$$\mathrm{E}[e^{tY_k}] \leqslant \mathrm{E}[c + dY_k] = c + d \underbrace{\mathrm{E}[Y_k]}_{=0}$$

$$= c = \frac{e^t + e^{-t}}{2}.$$

Now we use the Taylor expansion of e^x to simplify the bound:

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Proof. (Continued)

$$
\begin{aligned}
\mathrm{E}[e^{tY_k}] &\leqslant \frac{e^t + e^{-t}}{2} \\
&= \frac{1}{2}\left(1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots\right) \\
&+ \frac{1}{2}\left(1 - t + \frac{t^2}{2!} - \frac{t^3}{3!} + - \cdots\right) \\
&= \left(1 + \frac{t^2}{2!} + \frac{t^4}{4!} + \frac{t^6}{6!} + \cdots\right) \\
&\leqslant \left(1 + \frac{t^2}{2 \cdot 1!} + \frac{t^4}{2^2 \cdot 2!} + \frac{t^6}{2^3 \cdot 3!} + \cdots\right) \quad \text{as } 2^k k! \leqslant (2k)! \\
&= e^{t^2/2}.
\end{aligned}
$$

Proof. (Continued)

$$\Pr[X \geqslant \mu + a] \leqslant e^{-at} \prod_{k=1}^{n} \mathrm{E}\left[e^{tY_k}\right] \leqslant e^{-at} \prod_{k=1}^{n} e^{t^2/2} = e^{nt^2/2 - at}$$

The function $h(t) = n\frac{t^2}{2} - at$ has a minimum at $t_0 = a/n$. So

$$\Pr[X \geqslant \mu + a] \leqslant e^{n(a/n)^2 - a(a/n)} = e^{-a^2/(2n)}.$$

(Continued.)

We can obtain the second Chernoff inequality as follows.

Set $X' = -X$. Then $X \leqslant \mu - a$ if and only if $X' \geqslant -\mu + a$. Therefore,

$$\Pr[X \leqslant \mu - a] \leqslant e^{-a^2/(2n)}.$$

This concludes the proof. □

Coin Toss

Suppose that we toss a fair coin 10,000 times. Let $X_k = 1$ denote the event that the k-th coin toss yields heads. Let $X = X_1 + X_2 + \cdots + X_{10,000}$. Then

$$\mathrm{E}[X] = 5000, \quad \mathrm{Var}[X] = np(1-p) = 2500.$$

1. **Markov**: $\Pr[X \geqslant 6000] \leqslant 5000/6000 = 5/6$.
2. **Chebychev**:
   $\Pr[X \geqslant 6000] = \Pr[X - \mathrm{E}[X] \geqslant 1000] \leqslant 2500/10^6 = 1/400$.
3. **Chernoff**:
   $\Pr[X \geqslant 6000] = \Pr[X \geqslant \mathrm{E}[X] + 1000] \leqslant e^{-10^6/(2 \cdot 10,000)} = e^{-50}$.

Proposition

Let $X = X_1 + X_2 + \cdots + X_n$ be the sum of n independent Bernoulli random variables with $\Pr[X_k = 1] = p_k$. Let

$$\mu = \mathrm{E}[X] = p_1 + p_2 + \cdots + p_n.$$

Then

$$\Pr[X \geqslant (1+\delta)\mu] \leqslant e^{\delta\mu}(1+\delta)^{-(1+\delta)\mu},$$
$$\Pr[X \leqslant (1-\delta)\mu] \leqslant e^{\delta\mu}(1-\delta)^{-(1-\delta)\mu}.$$

These are good bounds, but they are rarely used in this form. We will find more convenient, but looser bounds, as a consequence.

Proof.

For all $t > 1$, we have

$$\begin{aligned}
\Pr[X \geqslant (1+\delta)\mu] &= \Pr[t^X \geqslant t^{(1+\delta)\mu}] \\
&\leqslant \frac{\mathrm{E}[t^X]}{t^{(1+\delta)\mu}} \quad \text{by Markov's inequality} \\
&= \frac{\mathrm{E}[t^{X_1} t^{X_2} \cdots t^{X_n}]}{t^{(1+\delta)\mu}} \\
&= \frac{\mathrm{E}[t^{X_1}]\mathrm{E}[t^{X_2}] \cdots \mathrm{E}[t^{X_n}]}{t^{(1+\delta)\mu}} \quad \text{by independence}
\end{aligned}$$

Proof. (Continued)

For all $t > 1$, we have

$$\begin{aligned}
\Pr[X \geqslant (1+\delta)\mu] &= \frac{\prod_{k=1}^{n} \mathrm{E}[t^{X_k}]}{t^{(1+\delta)\mu}} \\
&= \frac{\prod_{k=1}^{n} ((1-p_k) + p_k t)}{t^{(1+\delta)\mu}} \\
&= \frac{\prod_{k=1}^{n} (1 + p_k(t-1))}{t^{(1+\delta)\mu}} \\
&\leqslant \frac{\prod_{k=1}^{n} e^{p_k(t-1)}}{t^{(1+\delta)\mu}} \\
&= \frac{e^{(t-1)\sum_{k=1}^{n} p_k}}{t^{(1+\delta)\mu}} = \frac{e^{(t-1)\mu}}{t^{(1+\delta)\mu}}
\end{aligned}$$

Proof. (Continued)

We showed that for all $t \geqslant 1$, we have

$$\Pr[X \geqslant (1+\delta)\mu] \leqslant \frac{e^{(t-1)\mu}}{t^{(1+\delta)\mu}}.$$

Substituting $t = 1 + \delta$ yields

$$\Pr[X \geqslant (1+\delta)\mu] \leqslant \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}},$$

which is our claim.

The proof of the second inequality is done in a similar way. □

Corollary

$$\Pr[X \geqslant (1+\delta)\mu] \leqslant e^{-\delta^2\mu/3} \quad \textit{for } 0 < \delta < 1,$$
$$\Pr[X \leqslant (1-\delta)\mu] \leqslant e^{-\delta^2\mu/2} \quad \textit{for } 0 < \delta < 1.$$

Proof.

For the first inequality, it suffices to show that

$$e^{\delta}(1+\delta)^{-(1+\delta)} \leqslant e^{-\delta^2/3}$$

for $0 < \delta < 1$. Taking logarithms on both sides yields the equivalent inequality

$$f(\delta) := \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leqslant 0$$

The latter inequality is not hard to show. Indeed,

$$f'(\delta) = 1 - \frac{1+\delta}{1+\delta} - \ln(1+\delta) + \frac{2}{3}\delta = -\ln(1+\delta) + \frac{2}{3}\delta.$$

Proof.

$$f'(\delta) = -\ln(1+\delta) + \frac{2}{3}\delta.$$

Since $f'(\delta) \leqslant 0$ for all δ in the range $0 \leqslant \delta \leqslant 1$, we can conclude that $f(\delta)$ is decreasing on $0 \leqslant \delta \leqslant 1$. Since $f(0) = 0$, it follows that

$$f(\delta) = \delta - (1+\delta)\ln(1+\delta) + \frac{\delta^2}{3} \leqslant 0.$$

Therefore, we can conclude that

$$\Pr[X \geqslant (1+\delta)\mu] \leqslant e^{-\delta^2\mu/3} \quad \text{for } 0 < \delta < 1$$

holds. The proof of the second inequality is similar. □

Probability Amplification

Recall that following definition.

Definition

Let ε be a constant in the range $0 \leqslant \varepsilon < 1/2$.

The class **BPP** consists of all languages L such that there exists a polynomial-time randomized algorithm A such that

1. $x \in L$ implies $\Pr[A(x) \text{ accepts}] \geqslant 1 - \varepsilon$,
2. $x \notin L$ implies $\Pr[A(x) \text{ rejects}] \geqslant 1 - \varepsilon$.

It is customary to choose $\epsilon = 1/3$.

Let A be a randomized algorithm that decides $L \in \mathbf{BPP}$.

Let us construct an algorithm A' that runs A on an input x precisely n times and returns the majority vote as an answer.

How likely is it that A' errs?

Probability Amplification

Let X_k denote the indicator random variable that the k-th run returns the correct result. Let

$$X = \sum_{k=1}^{n} X_k$$

denote the number of runs with correct answer. Then

$$\mathrm{E}[X] \geqslant \frac{2}{3}n.$$

Then A' gives the wrong answer with probability

$$\Pr\left[\sum_{k=1}^{n} X_k \leqslant \frac{n}{2}\right] \leqslant \Pr\left[\sum_{k=1}^{n} X_k \leqslant \mathrm{E}[X] - \frac{n}{6}\right] \leqslant e^{-(n/6)^2/(2n)} = e^{-n/72}.$$

Therefore, if we choose $n = 72\ln(1/\delta)$, then the algorithm errs with probability δ.

One can reduce the constant 72 by using other versions of the Chernoff bound.

However, the point is that the error δ of A' can be made as small as we please.

Conclusion

There are many different Chernoff bounds. The recipe to bound the tail of random variable X is as follows. If a and t real numbers with $t > 0$, then

$$\Pr[X \geqslant a] = \Pr[e^{tX} \geqslant e^{ta}]$$

$$\leqslant \frac{E[e^{tX}]}{e^{ta}}$$

by Markov's inequality. We can now choose t to minimize the right-hand side.

The right-hand side is often bounded in various ways so that the bound is easier to use. All the resulting bounds are referred to as Chernoff bounds. You can find somewhat tighter bounds and variations in our textbook.

M. Mitzenmacher, E. Upfal, Probability and Computing, 2nd edition, 2017, Chapter 4.

Stasys Jukna, *Crashkurs Mathematik für Informatiker*, B.G. Teubner, 2008.