Randomized Algorithms

From moments to exponential concentration

Alex Gittens
Department of Computer Science, Rensselaer Polytechnic Institute

Fall 2026

PART 1

Markov with better transforms

1

Variance alone gives a weak Gaussian tail

Let $G \sim N(0, \sigma^2)$. For every $x > 0$, Chebyshev gives

$$\Pr(|G| \geq x) \leq \frac{\sigma^2}{x^2}.$$

Variance alone gives a weak Gaussian tail

Let $G \sim N(0, \sigma^2)$. For every $x > 0$, Chebyshev gives

$$\Pr(|G| \geq x) \leq \frac{\sigma^2}{x^2}.$$

But a Gaussian actually satisfies

$$\Pr(|G| \geq x) \leq 2 \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Setting $x = t\sigma$ gives the following comparison (for any $\sigma > 0$):

	$x = 3\sigma$	$x = 4\sigma$	$x = 5\sigma$
Chebyshev	1.1×10^{-1}	6.3×10^{-2}	4.0×10^{-2}
Gaussian tail bound	2.2×10^{-2}	6.7×10^{-4}	7.5×10^{-6}

Variance alone gives a weak Gaussian tail

Let $G \sim N(0, \sigma^2)$. For every $x > 0$, Chebyshev gives

$$\Pr(|G| \geq x) \leq \frac{\sigma^2}{x^2}.$$

But a Gaussian actually satisfies

$$\Pr(|G| \geq x) \leq 2 \exp\left(-\frac{x^2}{2\sigma^2}\right).$$

Setting $x = t\sigma$ gives the following comparison (for any $\sigma > 0$):

	$x = 3\sigma$	$x = 4\sigma$	$x = 5\sigma$
Chebyshev	1.1×10^{-1}	6.3×10^{-2}	4.0×10^{-2}
Gaussian tail bound	2.2×10^{-2}	6.7×10^{-4}	7.5×10^{-6}

What did Chebyshev discard?

It used only $\mathbb{E}G^2 = \sigma^2$. For a Gaussian, we know much more than its variance.

Gaussian moments grow on the $\sqrt{p}$ scale

For a random variable Y, define

$$\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}.$$

Gaussian moments grow on the $\sqrt{p}$ scale

For a random variable Y, define

$$\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}.$$

For $G \sim N(0, \sigma^2)$, some exact values are

$$\|G\|_2 = \sigma, \qquad \|G\|_4 = 3^{1/4}\sigma, \qquad \|G\|_6 = 15^{1/6}\sigma.$$

Gaussian moments grow on the $\sqrt{p}$ scale

For a random variable Y, define

$$\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}.$$

For $G \sim N(0, \sigma^2)$, some exact values are

$$\|G\|_2 = \sigma, \qquad \|G\|_4 = 3^{1/4}\sigma, \qquad \|G\|_6 = 15^{1/6}\sigma.$$

More generally,

$$\|G\|_p \asymp \sigma\sqrt{p}, \qquad p \geq 1,$$

and in particular

$$\boxed{\|G\|_p \leq \sigma\sqrt{p}.}$$

Gaussian moments grow on the $\sqrt{p}$ scale

For a random variable Y, define

$$\|Y\|_p = (\mathbb{E}|Y|^p)^{1/p}.$$

For $G \sim N(0, \sigma^2)$, some exact values are

$$\|G\|_2 = \sigma, \qquad \|G\|_4 = 3^{1/4}\sigma, \qquad \|G\|_6 = 15^{1/6}\sigma.$$

More generally,

$$\|G\|_p \asymp \sigma\sqrt{p}, \qquad p \geq 1,$$

and in particular

$$\boxed{\|G\|_p \leq \sigma\sqrt{p}.}$$

> Can we feed these stronger moment estimates into Markov instead of stopping at $p = 2$?

Every known moment gives a Markov bound

For every $p \geq 1$,

$$|Y| \geq x \quad \Longleftrightarrow \quad |Y|^p \geq x^p.$$

Every known moment gives a Markov bound

For every $p \geq 1$,

$$|Y| \geq x \quad \Longleftrightarrow \quad |Y|^p \geq x^p.$$

Applying Markov to $|Y|^p$ gives

$$\boxed{\Pr(|Y| \geq x) \leq \left(\frac{\|Y\|_p}{x}\right)^p.}$$

Every known moment gives a Markov bound

For every $p \geq 1$,

$$|Y| \geq x \quad \Longleftrightarrow \quad |Y|^p \geq x^p.$$

Applying Markov to $|Y|^p$ gives

$$\boxed{\Pr(|Y| \geq x) \leq \left(\frac{\|Y\|_p}{x}\right)^p.}$$

Equivalently,

$$\Pr(|Y| \geq x) \leq \frac{\mathbb{E}|Y|^p}{x^p}.$$

Larger p gives a faster power of x in the denominator, but $\mathbb{E}|Y|^p$ also grows with p.

Every known moment gives a Markov bound

For every $p \geq 1$,

$$|Y| \geq x \quad \Longleftrightarrow \quad |Y|^p \geq x^p.$$

Applying Markov to $|Y|^p$ gives

$$\boxed{\Pr(|Y| \geq x) \leq \left(\frac{\|Y\|_p}{x}\right)^p.}$$

Equivalently,

$$\Pr(|Y| \geq x) \leq \frac{\mathbb{E}|Y|^p}{x^p}.$$

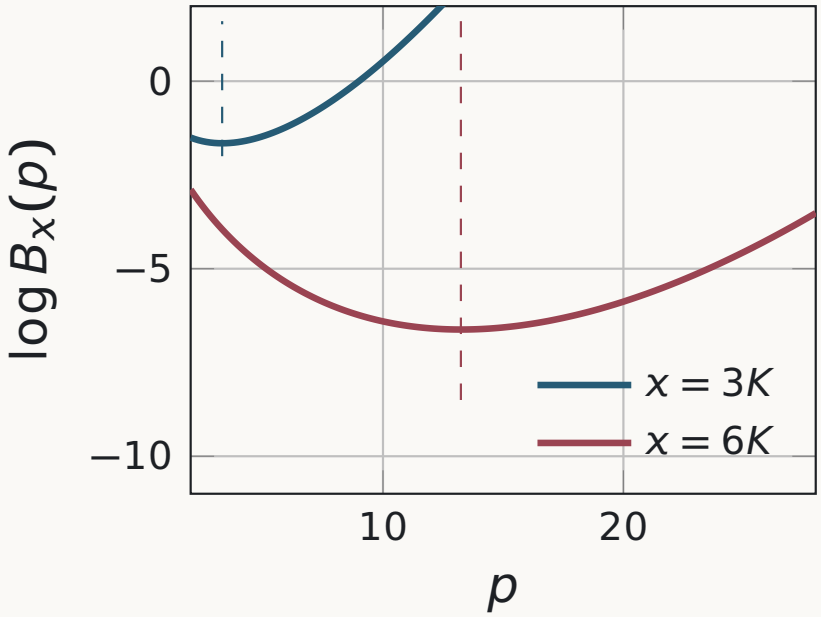
Larger p gives a faster power of x in the denominator, but $\mathbb{E}|Y|^p$ also grows with p.

> Markov itself was not the problem. A weak transform plus limited information produced the weak bound.

The best moment order depends on the deviation

Suppose $\|Y\|_p \le K\sqrt{p}$ for every $p \ge 2$. Markov gives

$$B_x(p) = \left(\frac{K\sqrt{p}}{x}\right)^p.$$

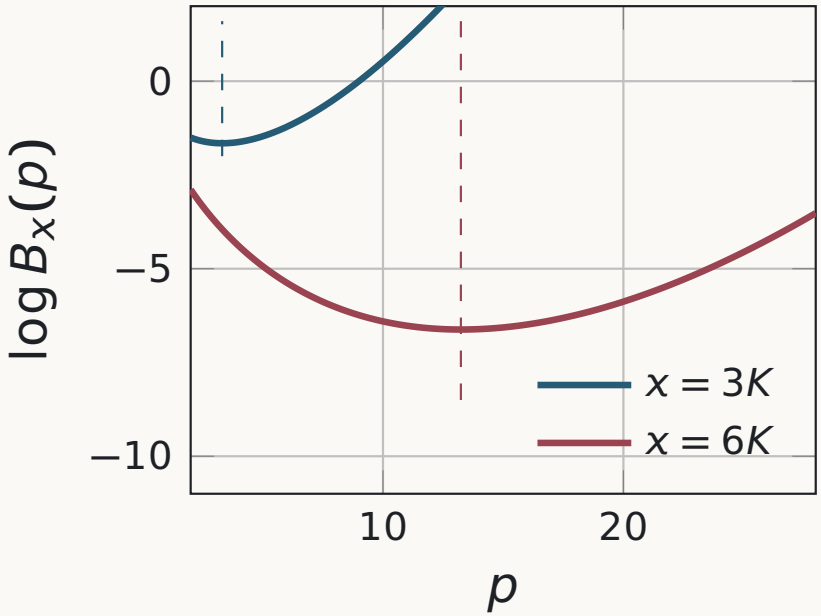


The best moment order depends on the deviation

Suppose $\|Y\|_p \le K\sqrt{p}$ for every $p \ge 2$. Markov gives

$$B_x(p) = \left(\frac{K\sqrt{p}}{x}\right)^p.$$

For a fixed deviation x, increasing p initially helps because the ratio $K\sqrt{p}/x$ is below one and is raised to the pth power.



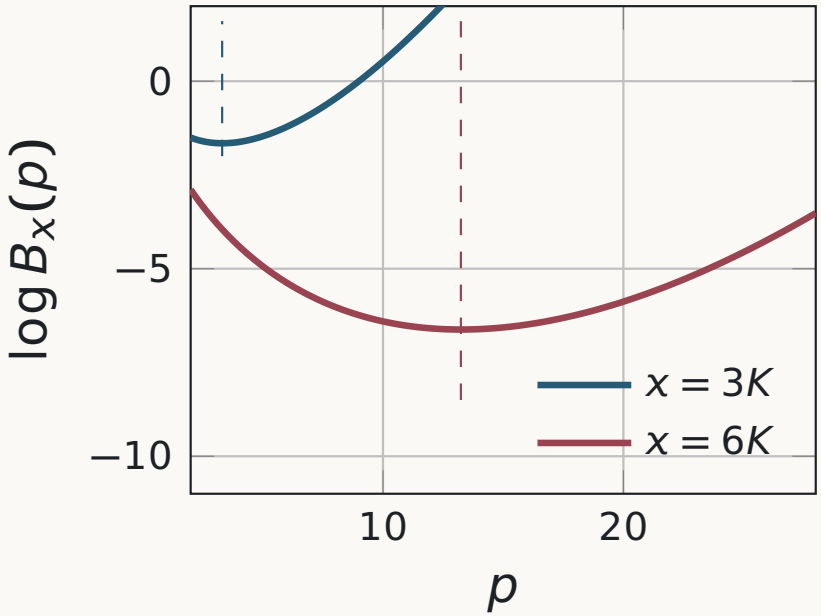
The best moment order depends on the deviation

Suppose $\|Y\|_p \le K\sqrt{p}$ for every $p \ge 2$. Markov gives

$$B_x(p) = \left(\frac{K\sqrt{p}}{x}\right)^p.$$

For a fixed deviation x, increasing p initially helps because the ratio $K\sqrt{p}/x$ is below one and is raised to the pth power.

But the numerator $K\sqrt{p}$ also grows with p, so eventually increasing p makes the bound worse.



The best moment order depends on the deviation

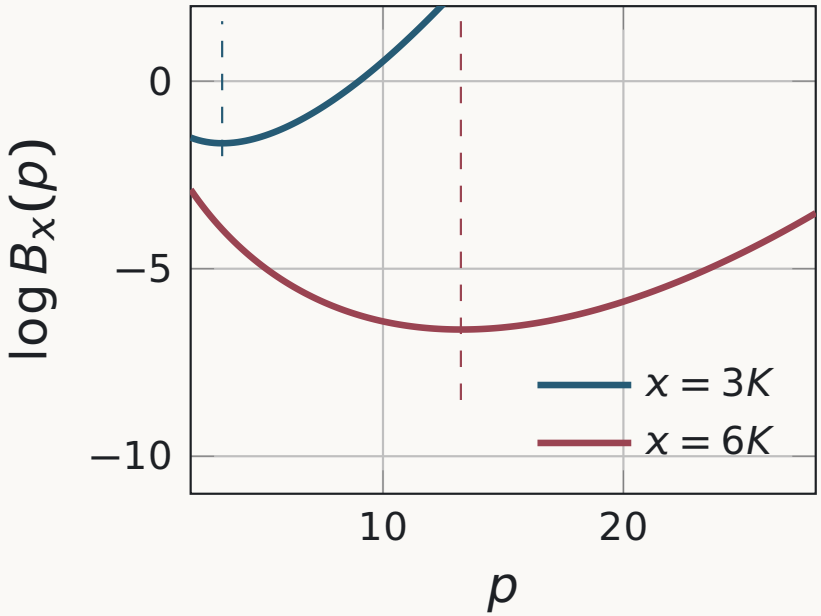
Suppose $\|Y\|_p \le K\sqrt{p}$ for every $p \ge 2$. Markov gives

$$B_x(p) = \left(\frac{K\sqrt{p}}{x}\right)^p.$$

For a fixed deviation x, increasing p initially helps because the ratio $K\sqrt{p}/x$ is below one and is raised to the pth power.

But the numerator $K\sqrt{p}$ also grows with p, so eventually increasing p makes the bound worse.

The best moment order therefore depends on the target deviation x.



Optimizing the moment recovers a Gaussian tail

Let

$$\phi(p) = p\left(\log\frac{K}{x} + \frac{1}{2}\log p\right).$$

Optimizing the moment recovers a Gaussian tail

Let

$$\phi(p) = p\left(\log\frac{K}{x} + \frac{1}{2}\log p\right).$$

Then

$$\phi'(p) = \log\frac{K}{x} + \frac{1}{2}\log p + \frac{1}{2}.$$

Optimizing the moment recovers a Gaussian tail

Let

$$\phi(p) = p\left(\log\frac{K}{x} + \frac{1}{2}\log p\right).$$

Then

$$\phi'(p) = \log\frac{K}{x} + \frac{1}{2}\log p + \frac{1}{2}.$$

Setting $\phi'(p) = 0$ gives

$$\boxed{p_\star = \frac{x^2}{eK^2}.}$$

Optimizing the moment recovers a Gaussian tail

Let

$$\phi(p) = p\left(\log\frac{K}{x} + \frac{1}{2}\log p\right).$$

Then

$$\phi'(p) = \log\frac{K}{x} + \frac{1}{2}\log p + \frac{1}{2}.$$

Setting $\phi'(p) = 0$ gives

$$\boxed{p_\star = \frac{x^2}{eK^2}.}$$

Substituting $p_\star$ yields, whenever $p_\star \geq 2$,

$$\boxed{\Pr(|Y| \geq x) \leq \exp\left(-\frac{x^2}{2eK^2}\right).}$$

Optimizing the moment recovers a Gaussian tail

Let

$$\phi(p) = p\left(\log\frac{K}{x} + \frac{1}{2}\log p\right).$$

Then

$$\phi'(p) = \log\frac{K}{x} + \frac{1}{2}\log p + \frac{1}{2}.$$

Setting $\phi'(p) = 0$ gives

$$\boxed{p_\star = \frac{x^2}{eK^2}.}$$

Substituting $p_\star$ yields, whenever $p_\star \geq 2$,

$$\boxed{\Pr(|Y| \geq x) \leq \exp\left(-\frac{x^2}{2eK^2}\right).}$$

For a Gaussian, $K = \sigma$: higher-moment growth alone already implies the correct e^{-cx^2/σ^2} tail shape.

Gaussian moment growth is not unique to Gaussia

The argument we just used only required the moment-growth condition

$$\|Y\|_p \le K\sqrt{p}, \qquad p \ge 1.$$

Gaussian moment growth is not unique to Gaussia

The argument we just used only required the moment-growth condition

$$\|Y\|_p \le K\sqrt{p}, \qquad p \ge 1.$$

A centered random variable whose moments grow on this scale is called **subgaussian**, up to the choice of numerical constants.

Gaussian moment growth is not unique to Gaussia

The argument we just used only required the moment-growth condition

$$\|Y\|_p \le K\sqrt{p}, \qquad p \ge 1.$$

A centered random variable whose moments grow on this scale is called **subgaussian**, up to the choice of numerical constants.
Simple examples include:

- Gaussian random variables;
- Rademacher variables, $Y \in \{-1, +1\}$ with equal probability;
- $Y \sim \mathrm{Unif}[-1, 1]$;
- more generally, centered bounded random variables (we will prove this later using Hoeffding's lemma).

Subgaussian moment growth means light tails

Optimizing the moment order showed that

$$\|Y\|_p \lesssim K\sqrt{p} \qquad \Longrightarrow \qquad \Pr(|Y| \geq x) \lesssim e^{-cx^2/K^2}.$$

Subgaussian moment growth means light tails

Optimizing the moment order showed that

$$\|Y\|_p \lesssim K\sqrt{p} \qquad \Longrightarrow \qquad \Pr(|Y| \geq x) \lesssim e^{-cx^2/K^2}.$$

Conversely, a tail bound of the form e^{-cx^2/K^2} implies $\|Y\|_p \lesssim K\sqrt{p}$. Thus, up to constants,

$$\boxed{\|Y\|_p \lesssim K\sqrt{p} \quad \Longleftrightarrow \quad \Pr(|Y| \geq x) \lesssim e^{-cx^2/K^2}.}$$

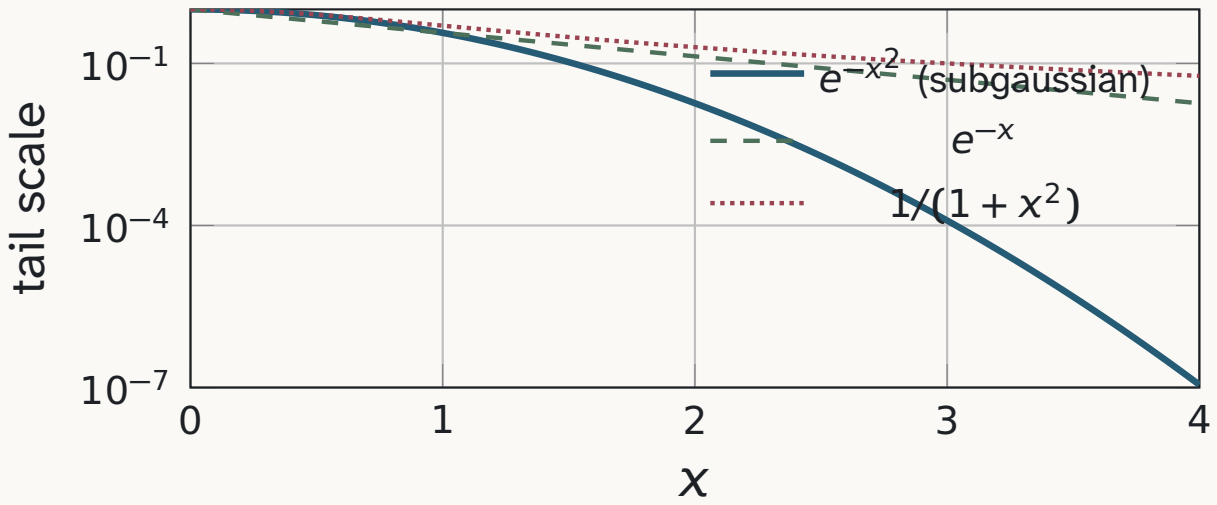
Subgaussian moment growth means light tails

Optimizing the moment order showed that

$$\|Y\|_p \lesssim K\sqrt{p} \qquad \Longrightarrow \qquad \Pr(|Y| \geq x) \lesssim e^{-cx^2/K^2}.$$

Conversely, a tail bound of the form e^{-cx^2/K^2} implies $\|Y\|_p \lesssim K\sqrt{p}$. Thus, up to constants,

$$\boxed{\|Y\|_p \lesssim K\sqrt{p} \quad \Longleftrightarrow \quad \Pr(|Y| \geq x) \lesssim e^{-cx^2/K^2}.}$$



The exponential transform gives another route

Instead of the family $|Y|^p$, use the exponential transform

$$Y \mapsto e^{\theta Y}.$$

Its expectation

$$M_Y(\theta) = \mathbb{E}e^{\theta Y}$$

is the **moment generating function (MGF)** of Y.

The exponential transform gives another route

Instead of the family $|Y|^p$, use the exponential transform

$$Y \mapsto e^{\theta Y}.$$

Its expectation

$$M_Y(\theta) = \mathbb{E}e^{\theta Y}$$

is the **moment generating function (MGF)** of Y.
For every $\theta > 0$,

$$Y \geq x \quad \Longrightarrow \quad e^{\theta Y} \geq e^{\theta x},$$

so Markov gives

$$\boxed{\Pr(Y \geq x) \leq e^{-\theta x}\mathbb{E}e^{\theta Y}.}$$

The exponential transform gives another route

Instead of the family $|Y|^p$, use the exponential transform

$$Y \mapsto e^{\theta Y}.$$

Its expectation

$$M_Y(\theta) = \mathbb{E}e^{\theta Y}$$

is the **moment generating function (MGF)** of Y.
For every $\theta > 0$,

$$Y \geq x \quad \Longrightarrow \quad e^{\theta Y} \geq e^{\theta x},$$

so Markov gives

$$\boxed{\Pr(Y \geq x) \leq e^{-\theta x}\mathbb{E}e^{\theta Y}.}$$

Because this holds for every $\theta > 0$,

$$\boxed{\Pr(Y \geq x) \leq \inf_{\theta>0} e^{-\theta x} M_Y(\theta)} = \inf_{\theta>0} \exp(\log M_Y(\theta) - \theta x).$$

The exponential transform gives another route

Instead of the family $|Y|^p$, use the exponential transform

$$Y \mapsto e^{\theta Y}.$$

Its expectation

$$M_Y(\theta) = \mathbb{E}e^{\theta Y}$$

is the **moment generating function (MGF)** of Y.
For every $\theta > 0$,

$$Y \geq x \quad \Longrightarrow \quad e^{\theta Y} \geq e^{\theta x},$$

so Markov gives

$$\boxed{\Pr(Y \geq x) \leq e^{-\theta x}\mathbb{E}e^{\theta Y}.}$$

Because this holds for every $\theta > 0$,

$$\boxed{\Pr(Y \geq x) \leq \inf_{\theta>0} e^{-\theta x} M_Y(\theta)} = \inf_{\theta>0} \exp(\log M_Y(\theta) - \theta x).$$

Applying Markov to an exponential transform and optimizing over θ is the **Laplace–Chernoff method**.

The Gaussian MGF gives the sharp exponent

For $G \sim N(0, \sigma^2)$,

$$\boxed{\mathbb{E}e^{\theta G} = e^{\sigma^2 \theta^2 / 2}.}$$

The Gaussian MGF gives the sharp exponent

For $G \sim N(0, \sigma^2)$,

$$\boxed{\mathbb{E}e^{\theta G} = e^{\sigma^2 \theta^2 / 2}.}$$

Hence

$$\Pr(G \geq x) \leq \exp\left(-\theta x + \frac{\sigma^2 \theta^2}{2} \right).$$

The Gaussian MGF gives the sharp exponent

For $G \sim N(0, \sigma^2)$,

$$\boxed{\mathbb{E}e^{\theta G} = e^{\sigma^2\theta^2/2}.}$$

Hence

$$\Pr(G \geq x) \leq \exp\left(-\theta x + \frac{\sigma^2\theta^2}{2}\right).$$

The exponent is minimized at

$$\theta_\star = \frac{x}{\sigma^2}.$$

The Gaussian MGF gives the sharp exponent

For $G \sim N(0, \sigma^2)$,

$$\boxed{\mathbb{E}e^{\theta G} = e^{\sigma^2\theta^2/2}.}$$

Hence

$$\Pr(G \geq x) \leq \exp\left(-\theta x + \frac{\sigma^2\theta^2}{2}\right).$$

The exponent is minimized at

$$\theta_\star = \frac{x}{\sigma^2}.$$

Therefore

$$\boxed{\Pr(G \geq x) \leq e^{-x^2/(2\sigma^2)}.}$$

The Gaussian MGF gives the sharp exponent

For $G \sim N(0, \sigma^2)$,

$$\boxed{\mathbb{E}e^{\theta G} = e^{\sigma^2\theta^2/2}.}$$

Hence

$$\Pr(G \geq x) \leq \exp\left(-\theta x + \frac{\sigma^2\theta^2}{2}\right).$$

The exponent is minimized at

$$\theta_\star = \frac{x}{\sigma^2}.$$

Therefore

$$\boxed{\Pr(G \geq x) \leq e^{-x^2/(2\sigma^2)}.}$$

> The exponential transform recovers the Gaussian $e^{-x^2/(2\sigma^2)}$ exponent exactly.

Power transforms and MGFs tell the same story

Power transform

$$|Y|^p$$

Exponential transform

$$e^{\theta Y}$$

Power transforms and MGFs tell the same story

Power transform

$$|Y|^p$$

Uses a bound on $\|Y\|_p$.

Exponential transform

$$e^{\theta Y}$$

Uses a bound on $\mathbb{E}e^{\theta Y}$.

Power transforms and MGFs tell the same story

Power transform

$$|Y|^p$$

Uses a bound on $\|Y\|_p$.
Optimize the moment order p.

Exponential transform

$$e^{\theta Y}$$

Uses a bound on $\mathbb{E}e^{\theta Y}$.
Optimize the tilt θ.

Power transforms and MGFs tell the same story

Power transform

$$|Y|^p$$

Uses a bound on $\|Y\|_p$.
Optimize the moment order p.
For $\|Y\|_p \lesssim K\sqrt{p}$,

$$\Pr(|Y| \geq x) \leq e^{-cx^2/K^2}.$$

Exponential transform

$$e^{\theta Y}$$

Uses a bound on $\mathbb{E}e^{\theta Y}$.
Optimize the tilt θ.
For $\mathbb{E}e^{\theta Y} \leq e^{K^2\theta^2/2}$,

$$\Pr(Y \geq x) \leq e^{-x^2/(2K^2)}.$$

Power transforms and MGFs tell the same story

Power transform

$$|Y|^p$$

Uses a bound on $\|Y\|_p$.
Optimize the moment order p.
For $\|Y\|_p \lesssim K\sqrt{p}$,

$$\Pr(|Y| \geq x) \leq e^{-cx^2/K^2}.$$

Exponential transform

$$e^{\theta Y}$$

Uses a bound on $\mathbb{E}e^{\theta Y}$.
Optimize the tilt θ.
For $\mathbb{E}e^{\theta Y} \leq e^{K^2\theta^2/2}$,

$$\Pr(Y \geq x) \leq e^{-x^2/(2K^2)}.$$

We now know how to prove exponential concentration for an individual random variable. In randomized algorithms, however, the quantity of interest is often a sum

$$S = X_1 + \cdots + X_n.$$

How can we show that such sums stay close to their means?

PART 2

Concentration for sums

2

Variance makes Chebyshev convenient for sums

Let

$$S = \sum_{i=1}^{n} X_i.$$

In general,

$$\boxed{\operatorname{Var}(S) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j).}$$

Variance makes Chebyshev convenient for sums

Let

$$S = \sum_{i=1}^{n} X_i.$$

In general,

$$\boxed{\operatorname{Var}(S) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j).}$$

If the variables are independent, then $\operatorname{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\operatorname{Var}(S) = \sum_{i} \operatorname{Var}(X_i).$$

Variance makes Chebyshev convenient for sums

Let

$$S = \sum_{i=1}^{n} X_i.$$

In general,

$$\boxed{\operatorname{Var}(S) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j).}$$

If the variables are independent, then $\operatorname{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\operatorname{Var}(S) = \sum_{i} \operatorname{Var}(X_i).$$

For i.i.d. variables with variance σ^2,

$$\operatorname{Var}(S) = n\sigma^2.$$

Variance makes Chebyshev convenient for sums

Let

$$S = \sum_{i=1}^{n} X_i.$$

In general,

$$\boxed{\operatorname{Var}(S) = \sum_{i=1}^{n} \operatorname{Var}(X_i) + 2 \sum_{i<j} \operatorname{Cov}(X_i, X_j).}$$

If the variables are independent, then $\operatorname{Cov}(X_i, X_j) = 0$ for $i \neq j$, so

$$\operatorname{Var}(S) = \sum_{i} \operatorname{Var}(X_i).$$

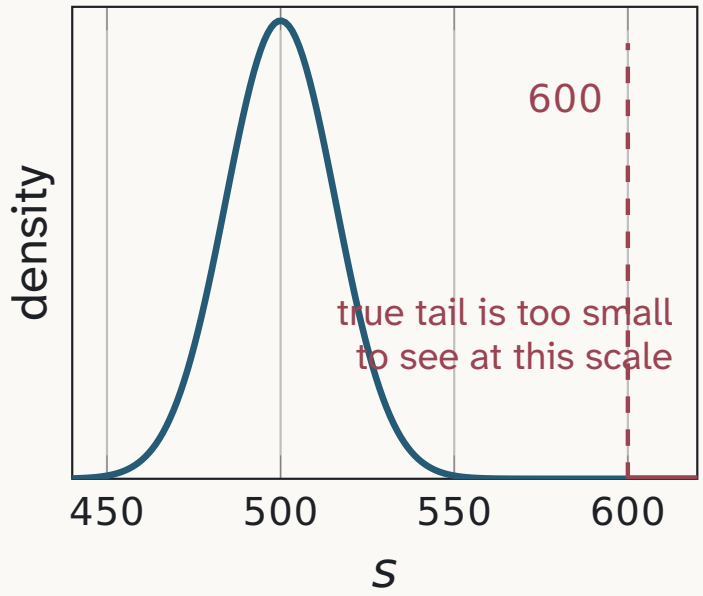
For i.i.d. variables with variance σ^2,

$$\operatorname{Var}(S) = n\sigma^2.$$

This property makes Chebyshev easy to apply to independent sums: the variance information it needs adds exactly.

Chebyshev is terrible for 1,000 fair coins

Let $S \sim \text{Binomial}(1000, 1/2)$. Then $\mathbb{E}S = 500$ and $\text{Var}(S) = 250$.

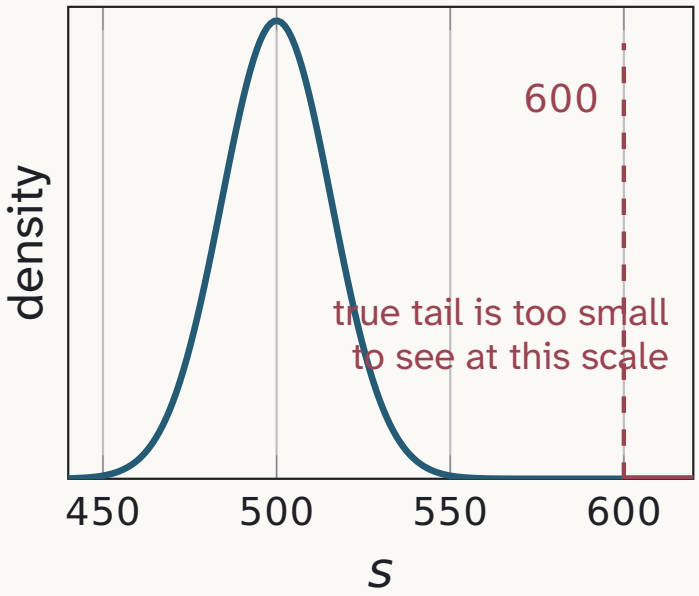


Chebyshev is terrible for 1,000 fair coins

Let $S \sim \text{Binomial}(1000, 1/2)$. Then $\mathbb{E}S = 500$ and $\text{Var}(S) = 250$.

For $S \geq 600$,

$$\Pr(S \geq 600) \leq \Pr(|S - 500| \geq 100)$$
$$\leq \frac{250}{100^2} = \boxed{0.025}.$$



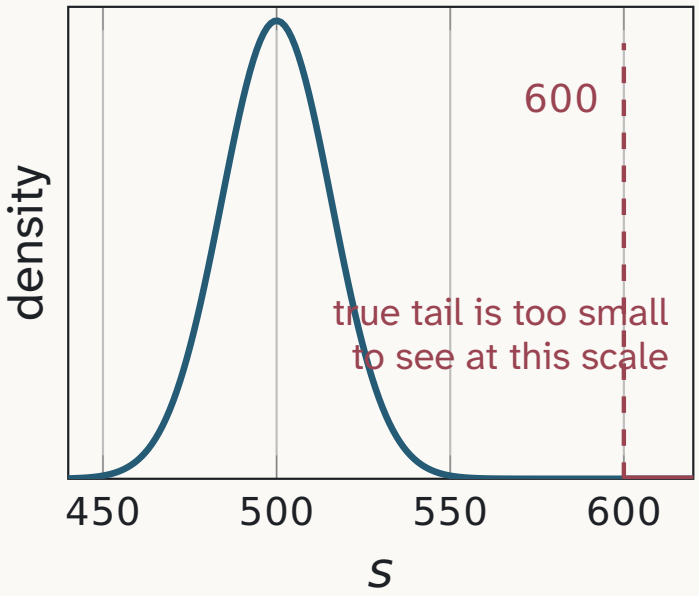
Chebyshev is terrible for 1,000 fair coins

Let $S \sim \text{Binomial}(1000, 1/2)$. Then $\mathbb{E}S = 500$ and $\text{Var}(S) = 250$.

For $S \geq 600$,

$$\Pr(S \geq 600) \leq \Pr(|S - 500| \geq 100)$$
$$\leq \frac{250}{100^2} = \boxed{0.025}.$$

600 is about 6.32 standard deviations above the mean.

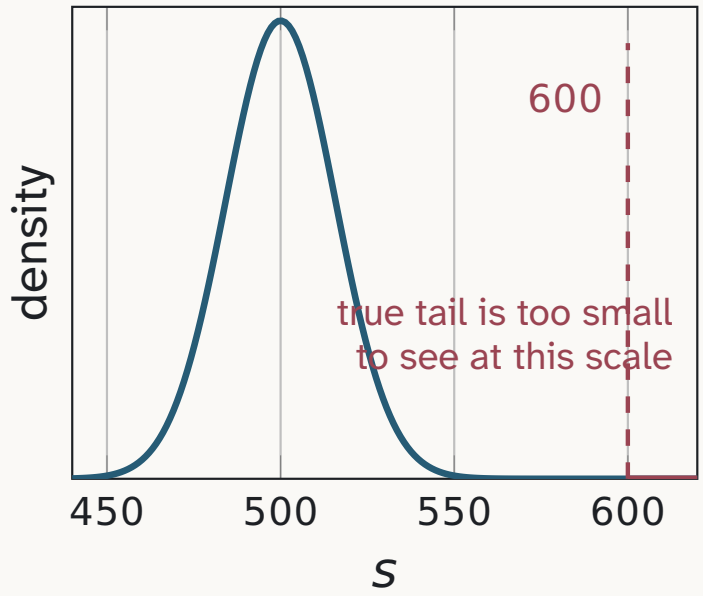


Chebyshev is terrible for 1,000 fair coins

Let $S \sim \text{Binomial}(1000, 1/2)$. Then $\mathbb{E}S = 500$ and $\text{Var}(S) = 250$.
For $S \geq 600$,

$$\Pr(S \geq 600) \leq \Pr(|S - 500| \geq 100)$$
$$\leq \frac{250}{100^2} = \boxed{0.025}.$$

600 is about 6.32 standard deviations above the mean. Chebyshev gives 0.025 even though the threshold is more than six standard deviations above the mean. Can we use the fact that every summand lies in $[0, 1]$ to do better?



Higher moments do not compose cleanly

Could we repeat the Gaussian strategy and use pth moments of $S - \mathbb{E}S$?

Higher moments do not compose cleanly

Could we repeat the Gaussian strategy and use pth moments of $S - \mathbb{E}S$? For independent centered variables,

$$\mathbb{E}(X_1 + X_2)^2 = \mathbb{E}X_1^2 + \underbrace{2\mathbb{E}X_1X_2}_{0} + \mathbb{E}X_2^2.$$

Higher moments do not compose cleanly

Could we repeat the Gaussian strategy and use pth moments of $S - \mathbb{E}S$? For independent centered variables,

$$\mathbb{E}(X_1 + X_2)^2 = \mathbb{E}X_1^2 + \underbrace{2\mathbb{E}X_1X_2}_{0} + \mathbb{E}X_2^2.$$

For the fourth moment, independence and centering kill the odd mixed terms, but

$$\mathbb{E}(X_1 + X_2)^4 = \mathbb{E}X_1^4 + \underbrace{6\mathbb{E}X_1^2\mathbb{E}X_2^2}_{\text{survives}} + \mathbb{E}X_2^4.$$

Higher moments do not compose cleanly

Could we repeat the Gaussian strategy and use pth moments of $S - \mathbb{E}S$? For independent centered variables,

$$\mathbb{E}(X_1 + X_2)^2 = \mathbb{E}X_1^2 + \underbrace{2\mathbb{E}X_1X_2}_{0} + \mathbb{E}X_2^2.$$

For the fourth moment, independence and centering kill the odd mixed terms, but

$$\mathbb{E}(X_1 + X_2)^4 = \mathbb{E}X_1^4 + \underbrace{6\mathbb{E}X_1^2\mathbb{E}X_2^2}_{\text{survives}} + \mathbb{E}X_2^4.$$

So the fourth moment of a sum is not the sum of the fourth moments.

Higher moments do not compose cleanly

Could we repeat the Gaussian strategy and use pth moments of $S - \mathbb{E}S$? For independent centered variables,

$$\mathbb{E}(X_1 + X_2)^2 = \mathbb{E}X_1^2 + \underbrace{2\mathbb{E}X_1X_2}_{0} + \mathbb{E}X_2^2.$$

For the fourth moment, independence and centering kill the odd mixed terms, but

$$\mathbb{E}(X_1 + X_2)^4 = \mathbb{E}X_1^4 + \underbrace{6\mathbb{E}X_1^2\mathbb{E}X_2^2}_{\text{survives}} + \mathbb{E}X_2^4.$$

So the fourth moment of a sum is not the sum of the fourth moments.

> Variance adds under independence. Higher moments contain mixed terms that survive.

Exponentials turn sums into products

Let $S = X_1 + \cdots + X_n$.

Exponentials turn sums into products

Let $S = X_1 + \cdots + X_n$.
The exponential transform turns addition of random variables into multiplication:

$$e^{\theta(X_1+\cdots+X_n)} = \prod_{i=1}^{n} e^{\theta X_i}.$$

Exponentials turn sums into products

Let $S = X_1 + \cdots + X_n$.
The exponential transform turns addition of random variables into multiplication:

$$e^{\theta(X_1+\cdots+X_n)} = \prod_{i=1}^{n} e^{\theta X_i}.$$

Independence then turns the expectation of that product into a product of expectations:

$$\mathbb{E}e^{\theta S} = \mathbb{E}\prod_{i=1}^{n} e^{\theta X_i} = \prod_{i=1}^{n} \mathbb{E}e^{\theta X_i}.$$

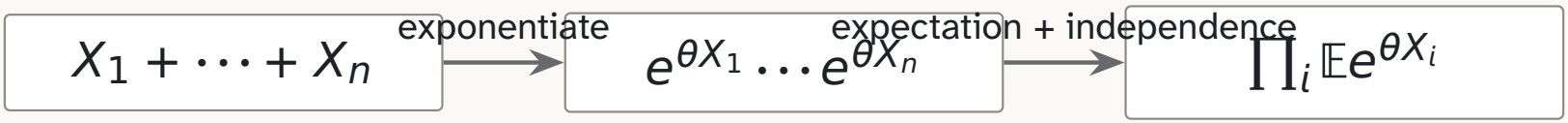
Exponentials turn sums into products

Let $S = X_1 + \cdots + X_n$.
The exponential transform turns addition of random variables into multiplication:

$$e^{\theta(X_1+\cdots+X_n)} = \prod_{i=1}^{n} e^{\theta X_i}.$$

Independence then turns the expectation of that product into a product of expectations:

$$\mathbb{E}e^{\theta S} = \mathbb{E}\prod_{i=1}^{n} e^{\theta X_i} = \prod_{i=1}^{n} \mathbb{E}e^{\theta X_i}.$$



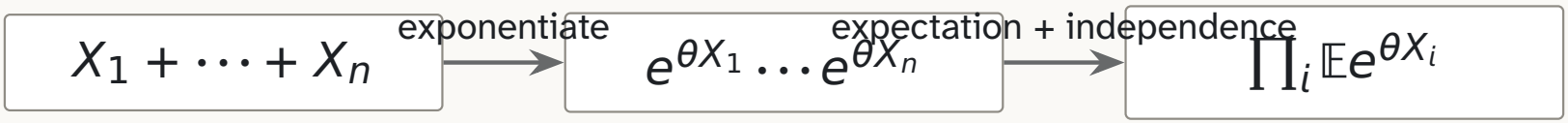
Exponentials turn sums into products

Let $S = X_1 + \cdots + X_n$.
The exponential transform turns addition of random variables into multiplication:

$$e^{\theta(X_1+\cdots+X_n)} = \prod_{i=1}^{n} e^{\theta X_i}.$$

Independence then turns the expectation of that product into a product of expectations:

$$\mathbb{E}e^{\theta S} = \mathbb{E}\prod_{i=1}^{n} e^{\theta X_i} = \prod_{i=1}^{n} \mathbb{E}e^{\theta X_i}.$$



For independent summands, the MGF of the sum is the product of the individual MGFs.

PART 3

Hoeffding's inequality

3

Use boundedness to control the MGF of the sum

In the fair-coin example, every summand satisfies

$$X_i \in [0, 1].$$

We will use that boundedness in three steps:

$X_i \in [a_i, b_i]$ $\Longrightarrow$ control each MGF $\Longrightarrow$ multiply the MGFs $\Longrightarrow$ optimize

Use boundedness to control the MGF of the sum

In the fair-coin example, every summand satisfies

$$X_i \in [0, 1].$$

We will use that boundedness in three steps:

$X_i \in [a_i, b_i]$ $\Longrightarrow$ control each MGF $\Longrightarrow$ multiply the MGFs $\Longrightarrow$ optimize

Hoeffding's lemma is the first step: it converts a range constraint into a quadratic bound on the centered MGF.

Boundedness controls each centered MGF

If $X \in [a, b]$, Hoeffding's lemma says that for every $\theta \in \mathbb{R}$,

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \leq \exp\left(\frac{\theta^2(b-a)^2}{8}\right).}$$

Boundedness controls each centered MGF

If $X \in [a, b]$, Hoeffding's lemma says that for every $\theta \in \mathbb{R}$,

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \leq \exp\left(\frac{\theta^2(b-a)^2}{8}\right).}$$

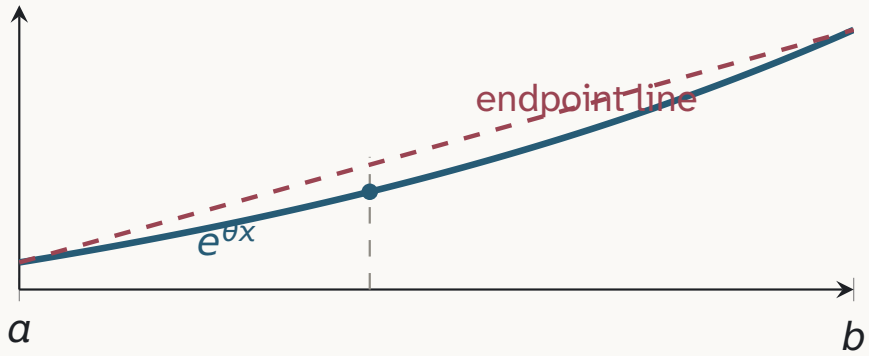
We will prove it from one elementary geometric fact: on $[a, b]$, the graph of $e^{\theta x}$ lies below the straight line joining its endpoint values. This property is called convexity.

Boundedness controls each centered MGF

If $X \in [a, b]$, Hoeffding's lemma says that for every $\theta \in \mathbb{R}$,

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le \exp\left(\frac{\theta^2(b-a)^2}{8}\right).}$$

We will prove it from one elementary geometric fact: on $[a, b]$, the graph of $e^{\theta x}$ lies below the straight line joining its endpoint values. This property is called convexity.



The endpoint line bounds the MGF

For $x \in [a, b]$, the endpoint line gives

$$e^{\theta x} \leq \frac{b-x}{b-a} e^{\theta a} + \frac{x-a}{b-a} e^{\theta b}.$$

The endpoint line bounds the MGF

For $x \in [a, b]$, the endpoint line gives

$$e^{\theta x} \le \frac{b-x}{b-a} e^{\theta a} + \frac{x-a}{b-a} e^{\theta b}.$$

Apply this with $x = X$ and take expectations. Define

$$q = \frac{\mathbb{E}X - a}{b-a}, \qquad 1-q = \frac{b-\mathbb{E}X}{b-a}.$$

The endpoint line bounds the MGF

For $x \in [a, b]$, the endpoint line gives

$$e^{\theta x} \le \frac{b-x}{b-a} e^{\theta a} + \frac{x-a}{b-a} e^{\theta b}.$$

Apply this with $x = X$ and take expectations. Define

$$q = \frac{\mathbb{E}X - a}{b-a}, \qquad 1 - q = \frac{b - \mathbb{E}X}{b-a}.$$

Then

$$\boxed{\mathbb{E}e^{\theta X} \le (1-q)e^{\theta a} + qe^{\theta b}.}$$

Centering leaves a one-parameter bound

Since $\mathbb{E}X = a + q(b-a)$,

$$\begin{aligned}\mathbb{E}e^{\theta(X-\mathbb{E}X)} &= e^{-\theta\mathbb{E}X}\mathbb{E}e^{\theta X} \\ &\le e^{-\theta[a+q(b-a)]}\left((1-q)e^{\theta a} + qe^{\theta b}\right).\end{aligned}$$

Centering leaves a one-parameter bound

Since $\mathbb{E}X = a + q(b-a)$,

$$\begin{aligned}\mathbb{E}e^{\theta(X-\mathbb{E}X)} &= e^{-\theta\mathbb{E}X}\mathbb{E}e^{\theta X} \\ &\le e^{-\theta[a+q(b-a)]}\left((1-q)e^{\theta a} + qe^{\theta b}\right).\end{aligned}$$

Cancel the common factor $e^{\theta a}$:

$$\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le e^{-q\theta(b-a)}\left((1-q) + qe^{\theta(b-a)}\right).$$

Centering leaves a one-parameter bound

Since $\mathbb{E}X = a + q(b-a)$,

$$\begin{aligned}\mathbb{E}e^{\theta(X-\mathbb{E}X)} &= e^{-\theta\mathbb{E}X}\mathbb{E}e^{\theta X} \\ &\le e^{-\theta[a+q(b-a)]}\left((1-q)e^{\theta a} + qe^{\theta b}\right).\end{aligned}$$

Cancel the common factor $e^{\theta a}$:

$$\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le e^{-q\theta(b-a)}\left((1-q) + qe^{\theta(b-a)}\right).$$

Setting $t = \theta(b-a)$ gives

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le e^{-tq}\big((1-q) + qe^{t}\big).}$$

Centering leaves a one-parameter bound

Since $\mathbb{E}X = a + q(b-a)$,

$$\begin{aligned}\mathbb{E}e^{\theta(X-\mathbb{E}X)} &= e^{-\theta\mathbb{E}X}\mathbb{E}e^{\theta X} \\ &\le e^{-\theta[a+q(b-a)]}\left((1-q)e^{\theta a} + qe^{\theta b}\right).\end{aligned}$$

Cancel the common factor $e^{\theta a}$:

$$\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le e^{-q\theta(b-a)}\left((1-q) + qe^{\theta(b-a)}\right).$$

Setting $t = \theta(b-a)$ gives

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le e^{-tq}\big((1-q) + qe^{t}\big).}$$

The centered MGF is therefore bounded using only the endpoints a, b and the mean $\mathbb{E}X$.

It remains to bound a one-variable function

From the previous slide, it remains to prove

$$e^{-tq}\big((1-q)+qe^{t}\big) \le e^{t^2/8}.$$

Take logs and define

$$g_q(t) = \log\big(e^{-tq}\big((1-q)+qe^{t}\big)\big).$$

It remains to bound a one-variable function

From the previous slide, it remains to prove

$$e^{-tq}\big((1-q)+qe^{t}\big) \le e^{t^2/8}.$$

Take logs and define

$$g_q(t) = \log\big(e^{-tq}\big((1-q)+qe^{t}\big)\big).$$

At the origin,

$$g_q(0) = 0, \qquad g_q'(0) = 0.$$

It remains to bound a one-variable function

From the previous slide, it remains to prove

$$e^{-tq}\big((1-q)+qe^{t}\big) \le e^{t^2/8}.$$

Take logs and define

$$g_q(t) = \log\big(e^{-tq}\big((1-q)+qe^{t}\big)\big).$$

At the origin,

$$g_q(0) = 0, \qquad g_q'(0) = 0.$$

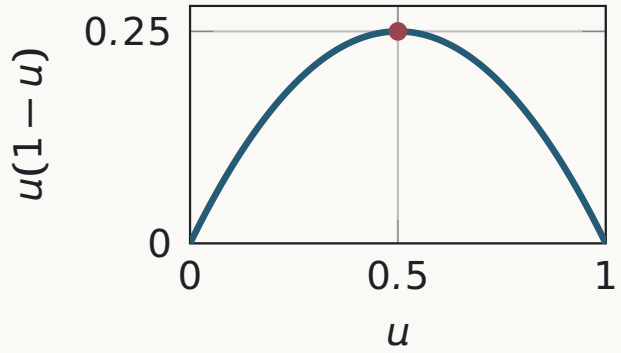
A direct differentiation gives

$$g_q''(t) = p_t(1-p_t), \qquad p_t = \frac{qe^{t}}{(1-q)+qe^{t}}.$$

The curvature is at most $1/4$

Since $0 \le p_t \le 1$,

$$g_q''(t) = p_t(1-p_t) \le \frac{1}{4}.$$



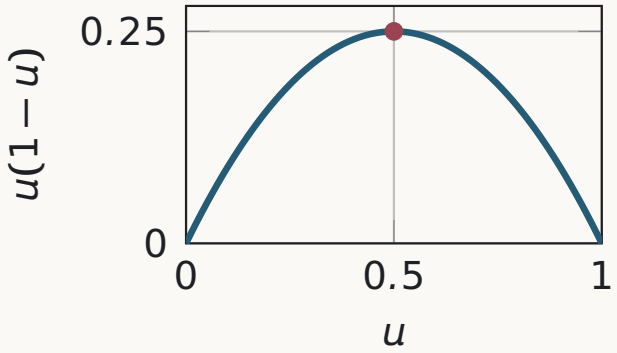
The curvature is at most $1/4$

Since $0 \le p_t \le 1$,

$$g_q''(t) = p_t(1-p_t) \le \frac{1}{4}.$$

Because $g_q(0) = g_q'(0) = 0$, Taylor's theorem gives, for some ξ between 0 and t,

$$g_q(t) = \frac{1}{2} g_q''(\xi) t^2 \le \frac{t^2}{8}.$$



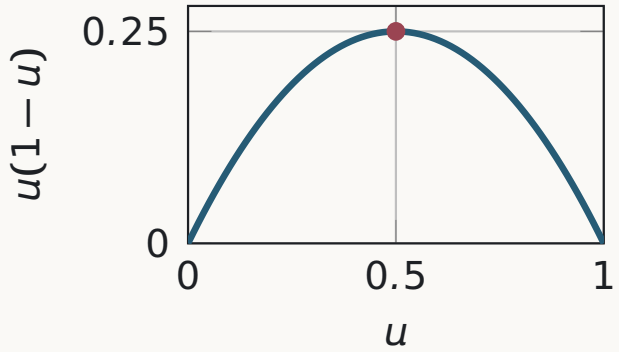
The curvature is at most $1/4$

Since $0 \le p_t \le 1$,

$$g_q''(t) = p_t(1-p_t) \le \frac{1}{4}.$$

Because $g_q(0) = g_q'(0) = 0$, Taylor's theorem gives, for some ξ between 0 and t,

$$g_q(t) = \frac{1}{2} g_q''(\xi) t^2 \le \frac{t^2}{8}.$$



Exponentiating and substituting $t = \theta(b-a)$ proves

$$\boxed{\mathbb{E}e^{\theta(X-\mathbb{E}X)} \le \exp\left(\frac{\theta^2(b-a)^2}{8}\right).}$$

Independence turns the lemma into a sum bound

For independent $X_i \in [a_i, b_i]$ and $S = \sum_i X_i$,

$$\begin{aligned}\mathbb{E}e^{\theta(S-\mathbb{E}S)} &= \prod_i \mathbb{E}e^{\theta(X_i-\mathbb{E}X_i)} \\ &\le \exp\left(\frac{\theta^2}{8}\sum_i (b_i-a_i)^2\right).\end{aligned}$$

Independence turns the lemma into a sum bound

For independent $X_i \in [a_i, b_i]$ and $S = \sum_i X_i$,

$$\begin{aligned}\mathbb{E}e^{\theta(S-\mathbb{E}S)} &= \prod_i \mathbb{E}e^{\theta(X_i-\mathbb{E}X_i)} \\ &\le \exp\left(\frac{\theta^2}{8}\sum_i (b_i-a_i)^2\right).\end{aligned}$$

Write

$$R^2 = \sum_i (b_i-a_i)^2.$$

Then the entire sum has the same quadratic MGF form as a subgaussian variable.

Optimizing the MGF gives Hoeffding

For every $\theta > 0$,

$$\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\theta x + \frac{\theta^2 R^2}{8}\right).$$

Optimizing the MGF gives Hoeffding

For every $\theta > 0$,

$$\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\theta x + \frac{\theta^2 R^2}{8}\right).$$

Differentiate the exponent:

$$-x + \frac{\theta R^2}{4} = 0 \qquad \Longrightarrow \qquad \theta_\star = \frac{4x}{R^2}.$$

Optimizing the MGF gives Hoeffding

For every $\theta > 0$,

$$\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\theta x + \frac{\theta^2 R^2}{8}\right).$$

Differentiate the exponent:

$$-x + \frac{\theta R^2}{4} = 0 \qquad \Longrightarrow \qquad \theta_\star = \frac{4x}{R^2}.$$

Substitution yields

$$\boxed{\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\frac{2x^2}{R^2}\right).}$$

Optimizing the MGF gives Hoeffding

For every $\theta > 0$,

$$\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\theta x + \frac{\theta^2 R^2}{8}\right).$$

Differentiate the exponent:

$$-x + \frac{\theta R^2}{4} = 0 \qquad \Longrightarrow \qquad \theta_\star = \frac{4x}{R^2}.$$

Substitution yields

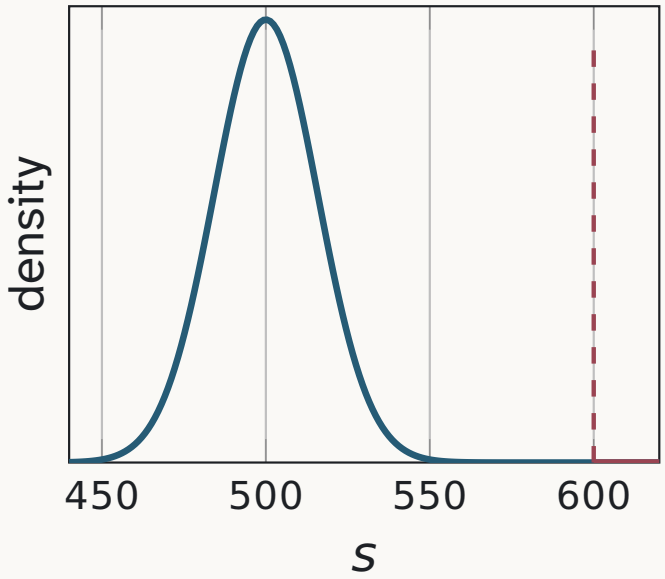
$$\boxed{\Pr(S - \mathbb{E}S \geq x) \leq \exp\left(-\frac{2x^2}{R^2}\right).}$$

Applying the same argument to $-S$ gives

$$\boxed{\Pr(|S - \mathbb{E}S| \geq x) \leq 2e^{-2x^2/R^2}.}$$

Hoeffding fixes the 1,000-coin example

For $S \sim \text{Binomial}(1000, 1/2)$, $\mathbb{E}S = 500$ and $R^2 = 1000$.

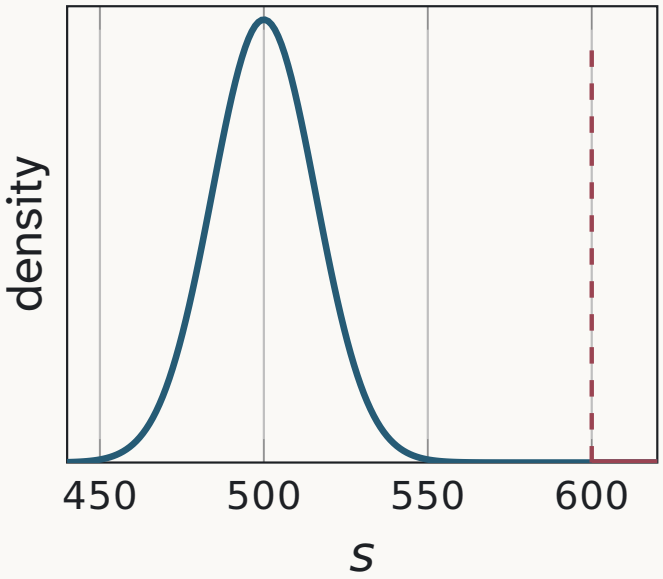


Hoeffding fixes the 1,000-coin example

For $S \sim \text{Binomial}(1000, 1/2)$, $\mathbb{E}S = 500$ and $R^2 = 1000$.

Hoeffding gives

$$\Pr(S \geq 600) \leq \exp\left(-\frac{2(100)^2}{1000}\right)$$
$$= e^{-20} < 2.1 \times 10^{-9}.$$

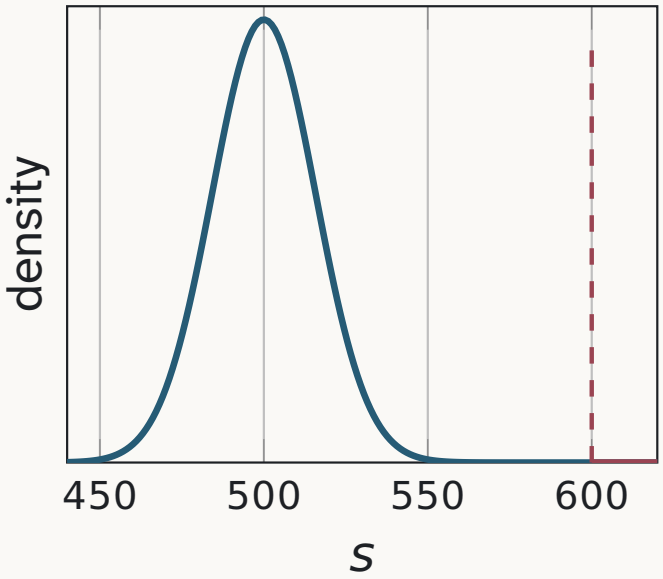


Hoeffding fixes the 1,000-coin example

For $S \sim \text{Binomial}(1000, 1/2)$, $\mathbb{E}S = 500$ and $R^2 = 1000$.

Hoeffding gives

$$\Pr(S \geq 600) \leq \exp\left(-\frac{2(100)^2}{1000}\right)$$
$$= e^{-20} < 2.1 \times 10^{-9}.$$



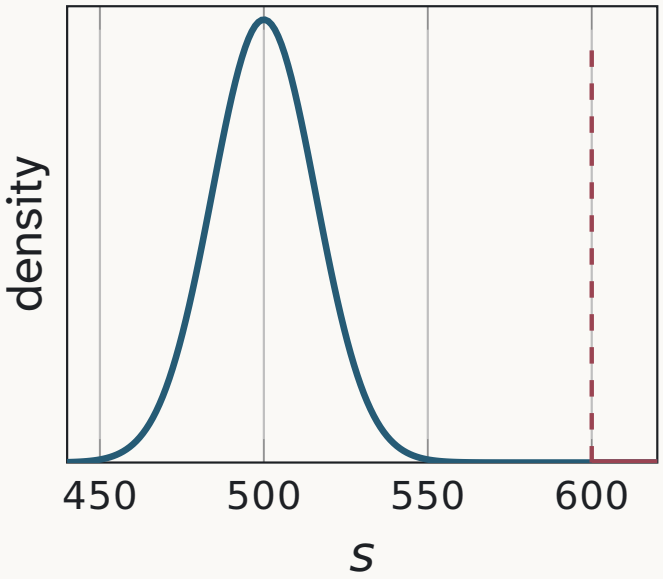
Chebyshev	2.5×10^{-2}
Hoeffding	2.1×10^{-9}
Exact	1.36×10^{-10}

Hoeffding fixes the 1,000-coin example

For $S \sim \text{Binomial}(1000, 1/2)$, $\mathbb{E}S = 500$ and $R^2 = 1000$.

Hoeffding gives

$$\Pr(S \geq 600) \leq \exp\left(-\frac{2(100)^2}{1000}\right)$$
$$= e^{-20} < 2.1 \times 10^{-9}.$$

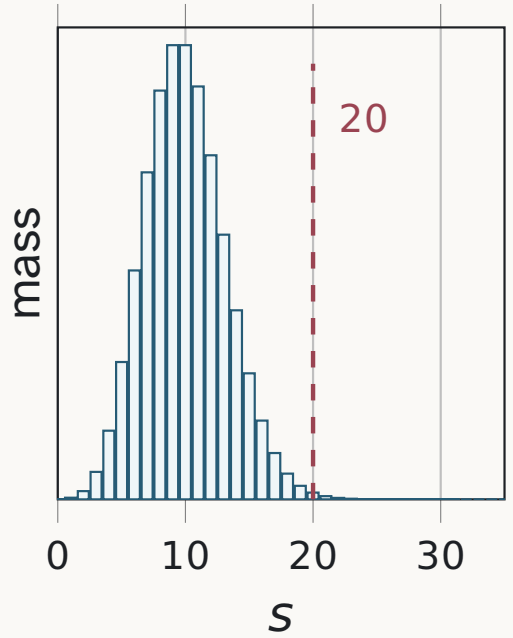


Chebyshev	2.5×10^{-2}
Hoeffding	2.1×10^{-9}
Exact	1.36×10^{-10}

Hoeffding is not exact, but it captures the correct exponential scale.

Hoeffding is weak when successes are rare

Let $S \sim \text{Binomial}(1000, 0.01)$, so $\mu = 10$, and consider $S \geq 20$.

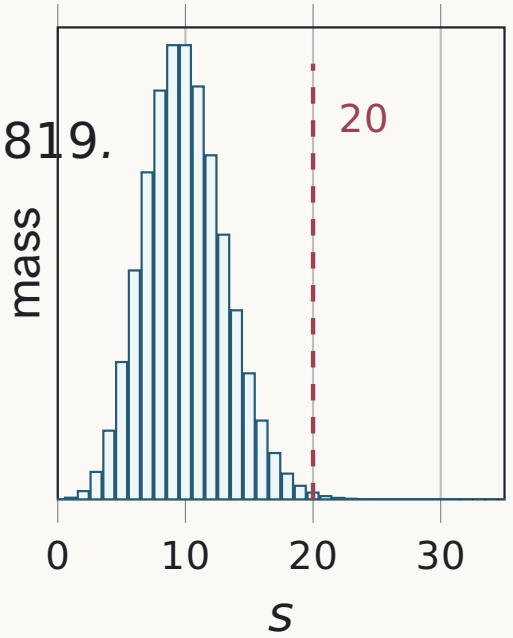


Hoeffding is weak when successes are rare

Let $S \sim \text{Binomial}(1000, 0.01)$, so $\mu = 10$, and consider $S \geq 20$.

Hoeffding uses only $X_i \in [0, 1]$:

$$\Pr(S - 10 \geq 10) \leq e^{-2(10)^2/1000} = e^{-0.2} \approx 0.819.$$



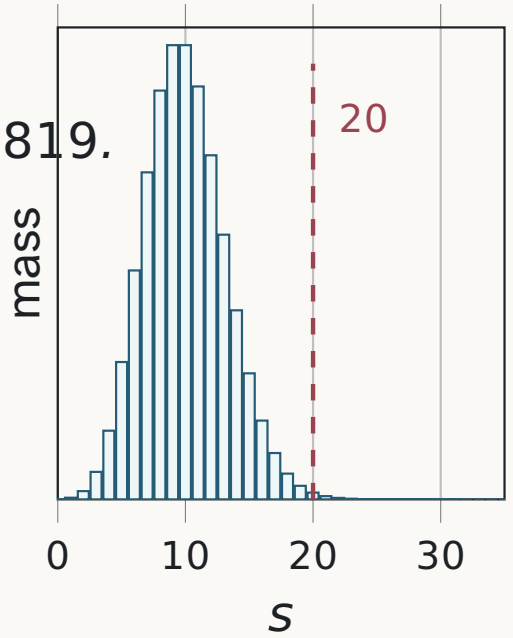
Hoeffding is weak when successes are rare

Let $S \sim \text{Binomial}(1000, 0.01)$, so $\mu = 10$, and consider $S \geq 20$.

Hoeffding uses only $X_i \in [0, 1]$:

$$\Pr(S - 10 \geq 10) \leq e^{-2(10)^2/1000} = e^{-0.2} \approx 0.819.$$

This is barely a useful upper bound, even though the threshold is twice the mean.



Hoeffding ignores the success probability

Why did the previous bound lose so much?

Hoeffding ignores the success probability

Why did the previous bound lose so much?
Hoeffding uses the same distribution-free information for every Bernoulli variable:

$$X_i \in [0, 1].$$

Hoeffding ignores the success probability

Why did the previous bound lose so much?
Hoeffding uses the same distribution-free information for every Bernoulli variable:

$$X_i \in [0, 1].$$

But for $X_i \sim \text{Bernoulli}(p_i)$ we know the full distribution,

$$\Pr(X_i = 1) = p_i, \qquad \Pr(X_i = 0) = 1 - p_i,$$

so we can compute its MGF directly and retain p_i.

Hoeffding ignores the success probability

Why did the previous bound lose so much?
Hoeffding uses the same distribution-free information for every Bernoulli variable:

$$X_i \in [0, 1].$$

But for $X_i \sim \text{Bernoulli}(p_i)$ we know the full distribution,

$$\Pr(X_i = 1) = p_i, \qquad \Pr(X_i = 0) = 1 - p_i,$$

so we can compute its MGF directly and retain p_i.
For a sum, these parameters combine through

$$\mu = \sum_i p_i.$$

In our example, $\mu = 10$ while $n = 1000$.

Hoeffding ignores the success probability

Why did the previous bound lose so much?
Hoeffding uses the same distribution-free information for every Bernoulli variable:

$$X_i \in [0, 1].$$

But for $X_i \sim \text{Bernoulli}(p_i)$ we know the full distribution,

$$\Pr(X_i = 1) = p_i, \qquad \Pr(X_i = 0) = 1 - p_i,$$

so we can compute its MGF directly and retain p_i.
For a sum, these parameters combine through

$$\mu = \sum_i p_i.$$

In our example, $\mu = 10$ while $n = 1000$.

Use the Bernoulli distribution

Instead of a distribution-free MGF bound based only on $X_i \in [0, 1]$, we will use the exact Bernoulli MGF.

PART 4

Multiplicative Chernoff bounds

4

The Bernoulli MGF retains the mean

Now use the full Bernoulli distribution rather than only the fact that $X_i \in [0, 1]$. If $X_i \sim \text{Bernoulli}(p_i)$, then

$$\begin{aligned} \mathbb{E}e^{\theta X_i} &= (1 - p_i) + p_i e^{\theta} \\ &= 1 + p_i(e^{\theta} - 1) \\ &\leq \exp\big(p_i(e^{\theta} - 1)\big). \end{aligned}$$

The Bernoulli MGF retains the mean

Now use the full Bernoulli distribution rather than only the fact that $X_i \in [0, 1]$. If $X_i \sim \text{Bernoulli}(p_i)$, then

$$\begin{aligned} \mathbb{E}e^{\theta X_i} &= (1 - p_i) + p_i e^{\theta} \\ &= 1 + p_i(e^{\theta} - 1) \\ &\le \exp\big(p_i(e^{\theta} - 1)\big). \end{aligned}$$

For independent Bernoulli variables,

$$S = \sum_i X_i, \qquad \mu = \mathbb{E}S = \sum_i p_i,$$

so

$$\boxed{\mathbb{E}e^{\theta S} \le \exp\big(\mu(e^{\theta} - 1)\big).}$$

Optimizing gives multiplicative Chernoff

For every $\theta > 0$,

$$\Pr(S \geq (1+\delta)\mu) \leq \exp\bigl(\mu(e^{\theta}-1) - \theta(1+\delta)\mu\bigr).$$

Optimizing gives multiplicative Chernoff

For every $\theta > 0$,

$$\Pr(S \geq (1+\delta)\mu) \leq \exp\big(\mu(e^{\theta}-1) - \theta(1+\delta)\mu\big).$$

The exponent is minimized when

$$e^{\theta_\star} = 1+\delta, \qquad \theta_\star = \log(1+\delta).$$

Optimizing gives multiplicative Chernoff

For every $\theta > 0$,

$$\Pr(S \geq (1+\delta)\mu) \leq \exp\big(\mu(e^{\theta}-1) - \theta(1+\delta)\mu\big).$$

The exponent is minimized when

$$e^{\theta_\star} = 1+\delta, \qquad \theta_\star = \log(1+\delta).$$

Substitution gives, for every $\delta > 0$,

$$\boxed{\Pr(S \geq (1+\delta)\mu) \leq e^{-\mu h(\delta)}, \qquad \delta > 0,}$$

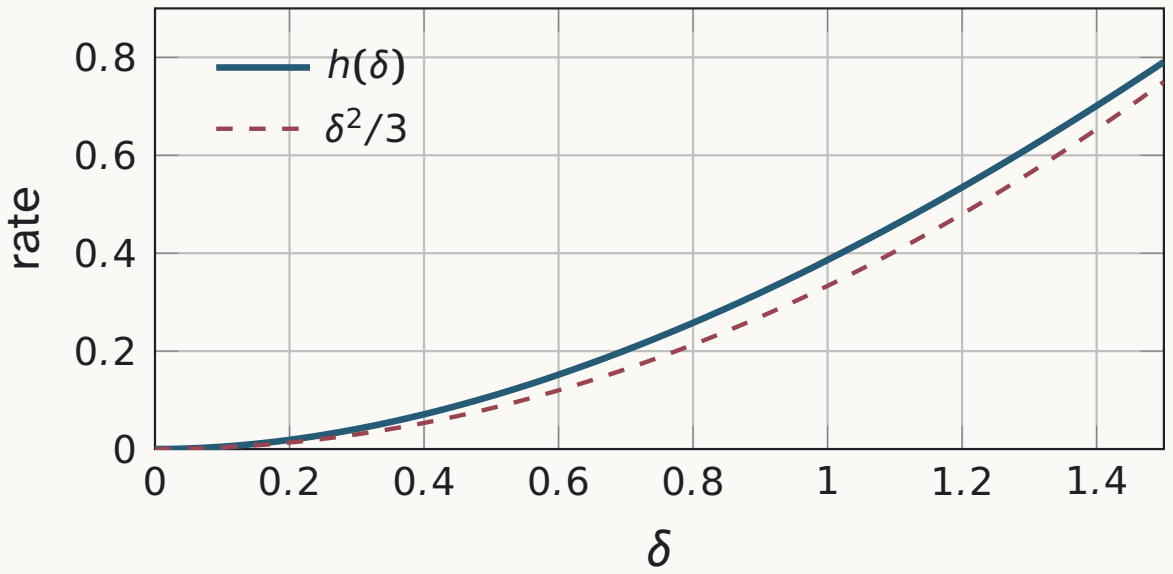
where

$$h(\delta) = (1+\delta)\log(1+\delta) - \delta.$$

For $0 \le \delta \le 1$, the exact rate dominates $\delta^2/3$

The exact exponent is

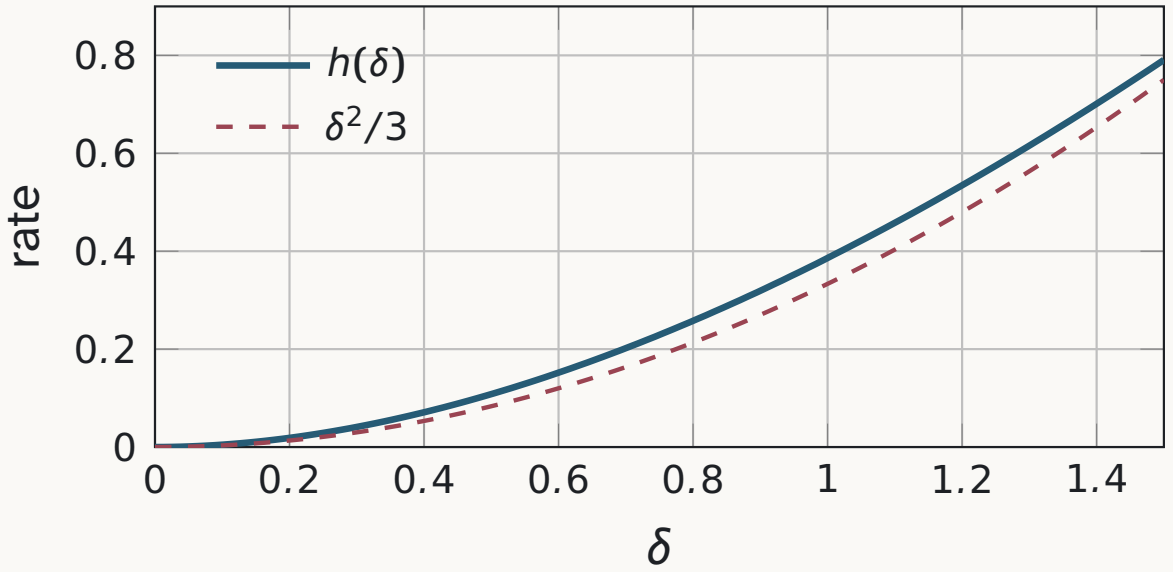
$$h(\delta) = (1+\delta)\log(1+\delta) - \delta.$$



For $0 \le \delta \le 1$, the exact rate dominates $\delta^2/3$

The exact exponent is

$$h(\delta) = (1+\delta)\log(1+\delta) - \delta.$$



The plot shows that, for $0 \le \delta \le 1$,

$$h(\delta) \ge \frac{\delta^2}{3}.$$

Therefore

$$\boxed{\Pr(S \ge (1+\delta)\mu) \le e^{-\delta^2\mu/3}, \qquad 0 \le \delta \le 1.}$$

Using the Bernoulli MGF fixes the $p = 0.01$ examp

Recall

$$S \sim \text{Binomial}(1000, 0.01), \qquad \mu = 10, \qquad S \geq 20 = (1+1)\mu.$$

Using the Bernoulli MGF fixes the $p = 0.01$ examp

Recall

$$S \sim \text{Binomial}(1000, 0.01), \qquad \mu = 10, \qquad S \geq 20 = (1+1)\mu.$$

Here $\delta = 1$ and

$$h(1) = 2\log 2 - 1 \approx 0.386.$$

Using the Bernoulli MGF fixes the $p = 0.01$ examp

Recall

$$S \sim \text{Binomial}(1000, 0.01), \qquad \mu = 10, \qquad S \geq 20 = (1+1)\mu.$$

Here $\delta = 1$ and

$$h(1) = 2\log 2 - 1 \approx 0.386.$$

Thus

$$\boxed{\Pr(S \geq 20) \leq e^{-10h(1)} \approx 0.021.}$$

Using the Bernoulli MGF fixes the $p = 0.01$ examp

Recall

$$S \sim \text{Binomial}(1000, 0.01), \qquad \mu = 10, \qquad S \geq 20 = (1+1)\mu.$$

Here $\delta = 1$ and

$$h(1) = 2\log 2 - 1 \approx 0.386.$$

Thus

$$\boxed{\Pr(S \geq 20) \leq e^{-10h(1)} \approx 0.021.}$$

Compare:

Hoeffding	0.819
Multiplicative Chernoff	0.021
Exact	0.00329

Chernoff is the method; multiplicative is the scale

The *Chernoff method* is the exponential-Markov argument

$$\Pr(Y \geq x) \leq \inf_{\theta > 0} e^{-\theta x} \mathbb{E} e^{\theta Y}.$$

Chernoff is the method; multiplicative is the scale

The *Chernoff method* is the exponential-Markov argument

$$\Pr(Y \geq x) \leq \inf_{\theta > 0} e^{-\theta x} \mathbb{E} e^{\theta Y}.$$

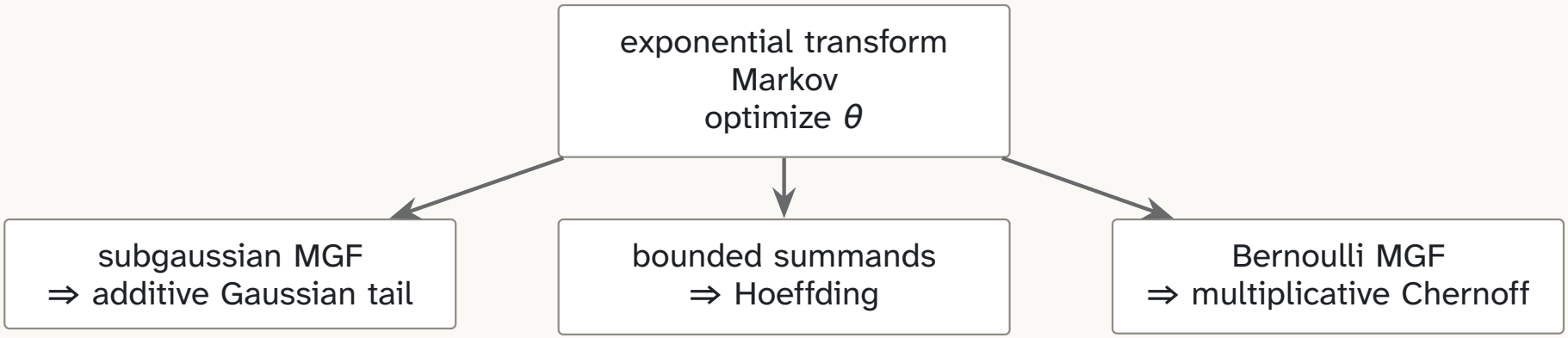
Different MGF information gives different concentration inequalities.

Chernoff is the method; multiplicative is the scale

The *Chernoff method* is the exponential-Markov argument

$$\Pr(Y \geq x) \leq \inf_{\theta > 0} e^{-\theta x} \mathbb{E} e^{\theta Y}.$$

Different MGF information gives different concentration inequalities.

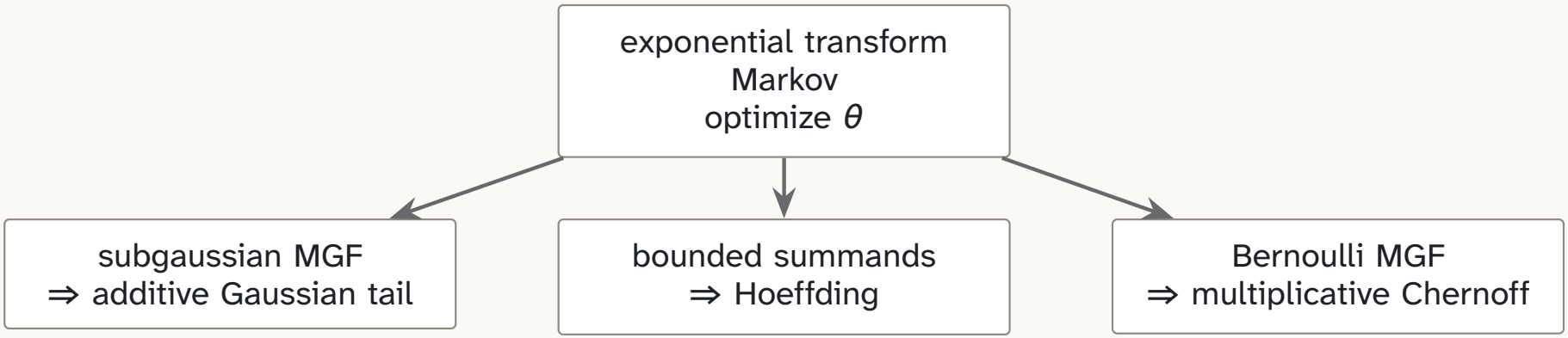


Chernoff is the method; multiplicative is the scale

The *Chernoff method* is the exponential-Markov argument

$$\Pr(Y \geq x) \leq \inf_{\theta > 0} e^{-\theta x} \mathbb{E} e^{\theta Y}.$$

Different MGF information gives different concentration inequalities.



"Multiplicative Chernoff" means the deviation is measured relative to the mean, $S \geq (1+\delta)\mu$. It is one application of the broader Laplace–Chernoff method.

Negative tilts give lower-tail Chernoff bounds

For a lower-tail threshold a, choose $\theta < 0$. Then $s \mapsto e^{\theta s}$ is decreasing, so

$$S \le a \quad \Longleftrightarrow \quad e^{\theta S} \ge e^{\theta a}.$$

Negative tilts give lower-tail Chernoff bounds

For a lower-tail threshold a, choose $\theta < 0$. Then $s \mapsto e^{\theta s}$ is decreasing, so

$$S \le a \quad \Longleftrightarrow \quad e^{\theta S} \ge e^{\theta a}.$$

Markov again gives

$$\Pr(S \le a) \le e^{-\theta a} \mathbb{E} e^{\theta S}.$$

Negative tilts give lower-tail Chernoff bounds

For a lower-tail threshold a, choose $\theta < 0$. Then $s \mapsto e^{\theta s}$ is decreasing, so

$$S \le a \quad \Longleftrightarrow \quad e^{\theta S} \ge e^{\theta a}.$$

Markov again gives

$$\Pr(S \le a) \le e^{-\theta a} \mathbb{E} e^{\theta S}.$$

For the Bernoulli MGF, the corresponding optimizer is

$$\theta_\star = \log(1-\delta) < 0.$$

For $0 \le \delta \le 1$, this gives

$$\boxed{\Pr(S \le (1-\delta)\mu) \le e^{-\delta^2 \mu/2}.}$$

Negative tilts give lower-tail Chernoff bounds

For a lower-tail threshold a, choose $\theta < 0$. Then $s \mapsto e^{\theta s}$ is decreasing, so

$$S \le a \quad \Longleftrightarrow \quad e^{\theta S} \ge e^{\theta a}.$$

Markov again gives

$$\Pr(S \le a) \le e^{-\theta a} \mathbb{E} e^{\theta S}.$$

For the Bernoulli MGF, the corresponding optimizer is

$$\theta_\star = \log(1-\delta) < 0.$$

For $0 \le \delta \le 1$, this gives

$$\boxed{\Pr(S \le (1-\delta)\mu) \le e^{-\delta^2 \mu/2}.}$$

Combining upper and lower tails gives the convenient form

$$\boxed{\Pr(|S-\mu| \ge \delta\mu) \le 2e^{-\delta^2 \mu/3}.}$$

PART 5

What is actually needed for concentration of sums

5

Independence is sufficient, but not necessary

For Chebyshev, independence gave

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

Independence is sufficient, but not necessary

For Chebyshev, independence gave

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

For the Laplace–Chernoff argument, independence was used at exactly one point:

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

Independence is sufficient, but not necessary

For Chebyshev, independence gave

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

For the Laplace–Chernoff argument, independence was used at exactly one point:

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

The concentration argument does not fundamentally require independence. It requires a useful upper bound on the MGF of the sum, for example

$$\mathbb{E}e^{\theta(S-\mathbb{E}S)} \leq e^{\psi(\theta)}.$$

Independence is sufficient, but not necessary

For Chebyshev, independence gave

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

For the Laplace–Chernoff argument, independence was used at exactly one point:

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

The concentration argument does not fundamentally require independence. It requires a useful upper bound on the MGF of the sum, for example

$$\mathbb{E}e^{\theta(S-\mathbb{E}S)} \leq e^{\psi(\theta)}.$$

Certain forms of negative dependence imply such MGF inequalities. Pairwise negative covariance alone generally does not.

Independence is sufficient, but not necessary

For Chebyshev, independence gave

$$\operatorname{Var}\left(\sum_i X_i\right) = \sum_i \operatorname{Var}(X_i).$$

For the Laplace–Chernoff argument, independence was used at exactly one point:

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

The concentration argument does not fundamentally require independence. It requires a useful upper bound on the MGF of the sum, for example

$$\mathbb{E}e^{\theta(S-\mathbb{E}S)} \leq e^{\psi(\theta)}.$$

Certain forms of negative dependence imply such MGF inequalities. Pairwise negative covariance alone generally does not.

> Any probabilistic structure that supplies a comparable MGF bound can replace independence in the concentration argument.

PART 6

Application: confidence amplification

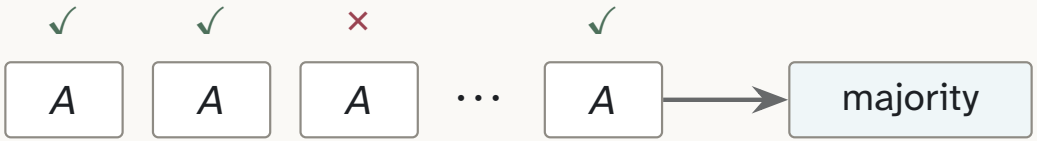
6

Majority vote amplifies a reliable algorithm

Goal: reduce the failure probability to an arbitrary $\varepsilon > 0$ by independently repeating an algorithm that succeeds with constant probability.
Suppose one run is correct with probability $q \geq 3/4$.

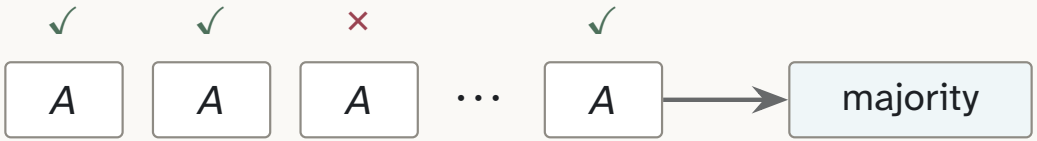
Majority vote amplifies a reliable algorithm

Goal: reduce the failure probability to an arbitrary $\varepsilon > 0$ by independently repeating an algorithm that succeeds with constant probability.
Suppose one run is correct with probability $q \geq 3/4$.



Majority vote amplifies a reliable algorithm

Goal: reduce the failure probability to an arbitrary $\varepsilon > 0$ by independently repeating an algorithm that succeeds with constant probability.
Suppose one run is correct with probability $q \geq 3/4$.

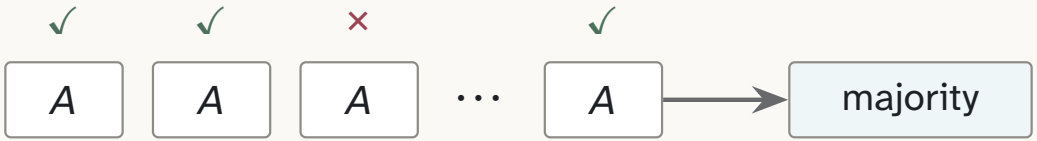


Run it independently r times and define

$$X_i = \mathbf{1}_{\{\text{run } i \text{ is correct}\}}, \qquad S = \sum_{i=1}^{r} X_i.$$

Majority vote amplifies a reliable algorithm

Goal: reduce the failure probability to an arbitrary $\varepsilon > 0$ by independently repeating an algorithm that succeeds with constant probability.
Suppose one run is correct with probability $q \geq 3/4$.



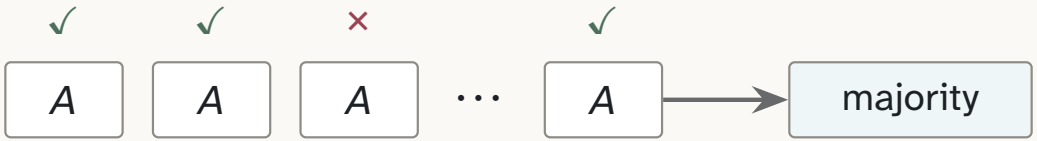
Run it independently r times and define

$$X_i = \mathbf{1}_{\{\text{run } i \text{ is correct}\}}, \qquad S = \sum_{i=1}^{r} X_i.$$

Then $\mathbb{E}S = qr \geq 3r/4$. A majority vote fails only if $S \leq r/2$, so failure is a constant-fraction lower-tail deviation from the mean.

Majority vote amplifies a reliable algorithm

Goal: reduce the failure probability to an arbitrary $\varepsilon > 0$ by independently repeating an algorithm that succeeds with constant probability.
Suppose one run is correct with probability $q \geq 3/4$.



Run it independently r times and define

$$X_i = \mathbf{1}_{\{\text{run } i \text{ is correct}\}}, \qquad S = \sum_{i=1}^{r} X_i.$$

Then $\mathbb{E}S = qr \geq 3r/4$. A majority vote fails only if $S \leq r/2$, so failure is a constant-fraction lower-tail deviation from the mean.
Writing $r/2 = (1-\delta)qr$ gives

$$\delta = 1 - \frac{1}{2q} \geq \frac{1}{3}.$$

Chernoff makes high confidence inexpensive

The lower-tail Chernoff bound gives

$$\Pr(\text{majority fails}) \leq \exp\left(-\frac{\delta^2 qr}{2}\right).$$

Chernoff makes high confidence inexpensive

The lower-tail Chernoff bound gives

$$\Pr(\text{majority fails}) \le \exp\left(-\frac{\delta^2 qr}{2}\right).$$

Since $q \ge 3/4$ and $\delta \ge 1/3$,

$$\boxed{\Pr(\text{majority fails}) \le e^{-r/24}.}$$

Chernoff makes high confidence inexpensive

The lower-tail Chernoff bound gives

$$\Pr(\text{majority fails}) \le \exp\left(-\frac{\delta^2 qr}{2}\right).$$

Since $q \ge 3/4$ and $\delta \ge 1/3$,

$$\boxed{\Pr(\text{majority fails}) \le e^{-r/24}.}$$

Therefore failure probability at most ε requires only

$$\boxed{r \ge 24 \log \frac{1}{\varepsilon}.}$$

Chernoff makes high confidence inexpensive

The lower-tail Chernoff bound gives

$$\Pr(\text{majority fails}) \le \exp\left(-\frac{\delta^2 qr}{2}\right).$$

Since $q \ge 3/4$ and $\delta \ge 1/3$,

$$\boxed{\Pr(\text{majority fails}) \le e^{-r/24}.}$$

Therefore failure probability at most ε requires only

$$\boxed{r \ge 24 \log \frac{1}{\varepsilon}.}$$

Exponential concentration converts constant success probability into very high confidence with only logarithmically many repetitions.

PART 7

Application: estimating a probability

7

Relative error becomes a Bernoulli-sum deviation

Goal: choose n so that

$$|\hat{p} - p| \leq \delta p$$

with probability at least $1 - \varepsilon$.
An event has unknown probability p. Draw n independent samples and form

$$X_i = \mathbf{1}_{\{\text{event occurs on sample } i\}}, \qquad S = \sum_{i=1}^{n} X_i, \qquad \hat{p} = \frac{S}{n}.$$

Relative error becomes a Bernoulli-sum deviation

Goal: choose n so that

$$|\hat{p} - p| \le \delta p$$

with probability at least $1 - \varepsilon$.
An event has unknown probability p. Draw n independent samples and form

$$X_i = \mathbf{1}_{\{\text{event occurs on sample } i\}}, \qquad S = \sum_{i=1}^{n} X_i, \qquad \hat{p} = \frac{S}{n}.$$

Since $\mu = \mathbb{E}S = np$,

$$\begin{aligned} |\hat{p} - p| \ge \delta p &\iff \left| \frac{S}{n} - p \right| \ge \delta p \\ &\iff |S - np| \ge \delta np \\ &\iff |S - \mu| \ge \delta \mu. \end{aligned}$$

Chernoff gives the required sample size

For $0 < \delta \leq 1$, the two-sided multiplicative bound gives

$$\boxed{\Pr\left(|\hat{p} - p| \geq \delta p\right) \leq 2e^{-\delta^2 pn/3}.}$$

Chernoff gives the required sample size

For $0 < \delta \leq 1$, the two-sided multiplicative bound gives

$$\boxed{\Pr\left(|\hat{p} - p| \geq \delta p\right) \leq 2e^{-\delta^2 pn/3}.}$$

To make this failure probability at most ε, it is enough that

$$2e^{-\delta^2 pn/3} \leq \varepsilon.$$

Chernoff gives the required sample size

For $0 < \delta \leq 1$, the two-sided multiplicative bound gives

$$\boxed{\Pr\left(|\hat{p} - p| \geq \delta p\right) \leq 2e^{-\delta^2 pn/3}.}$$

To make this failure probability at most ε, it is enough that

$$2e^{-\delta^2 pn/3} \leq \varepsilon.$$

Solving for n gives

$$\boxed{n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}.}$$

Rare events are expensive to estimate relatively

The sample size

$$n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}$$

grows like $1/p$: halving p doubles the work.

Rare events are expensive to estimate relatively

The sample size

$$n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}$$

grows like $1/p$: halving p doubles the work.

For $\delta = 0.1$ and $\varepsilon = 0.05$:

p	Chernoff samples
0.5	2.2×10^3
0.1	1.1×10^4
0.01	1.1×10^5
0.001	1.1×10^6

Rare events are expensive to estimate relatively

The sample size

$$n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}$$

grows like $1/p$: halving p doubles the work.

For $\delta = 0.1$ and $\varepsilon = 0.05$:

p	Chernoff samples
0.5	2.2×10^3
0.1	1.1×10^4
0.01	1.1×10^5
0.001	1.1×10^6

Hoeffding gives an additive guarantee independent of p:

$$\Pr(|\hat{p} - p| \geq t) \leq 2e^{-2nt^2}.$$

Rare events are expensive to estimate relatively

The sample size

$$n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}$$

grows like $1/p$: halving p doubles the work.

For $\delta = 0.1$ and $\varepsilon = 0.05$:

p	Chernoff samples
0.5	2.2×10^3
0.1	1.1×10^4
0.01	1.1×10^5
0.001	1.1×10^6

Hoeffding gives an additive guarantee independent of p:

$$\Pr(|\hat{p} - p| \geq t) \leq 2e^{-2nt^2}.$$

For $p = 0.01$, relative error $\delta = 0.1$ means $t = 10^{-3}$:

Chernoff: 1.1×10^5, Hoeffding: 1.8×10^6.

Rare events are expensive to estimate relatively

The sample size

$$n \geq \frac{3}{\delta^2 p} \log \frac{2}{\varepsilon}$$

grows like $1/p$: halving p doubles the work.

For $\delta = 0.1$ and $\varepsilon = 0.05$:

p	Chernoff samples
0.5	2.2×10^3
0.1	1.1×10^4
0.01	1.1×10^5
0.001	1.1×10^6

Hoeffding gives an additive guarantee independent of p:

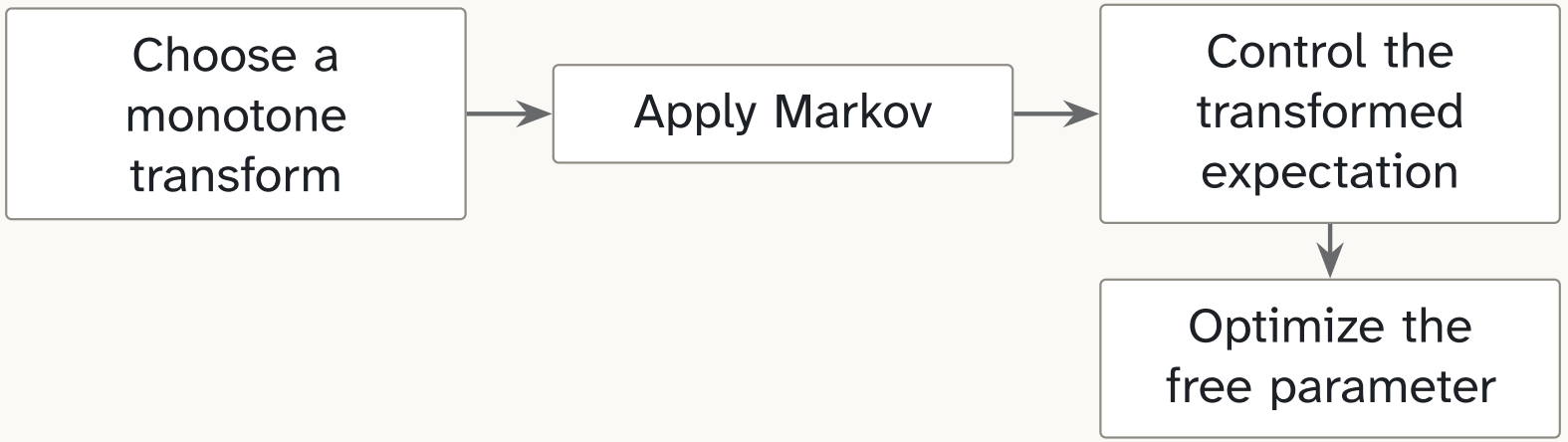
$$\Pr(|\hat{p} - p| \geq t) \leq 2e^{-2nt^2}.$$

For $p = 0.01$, relative error $\delta = 0.1$ means $t = 10^{-3}$:

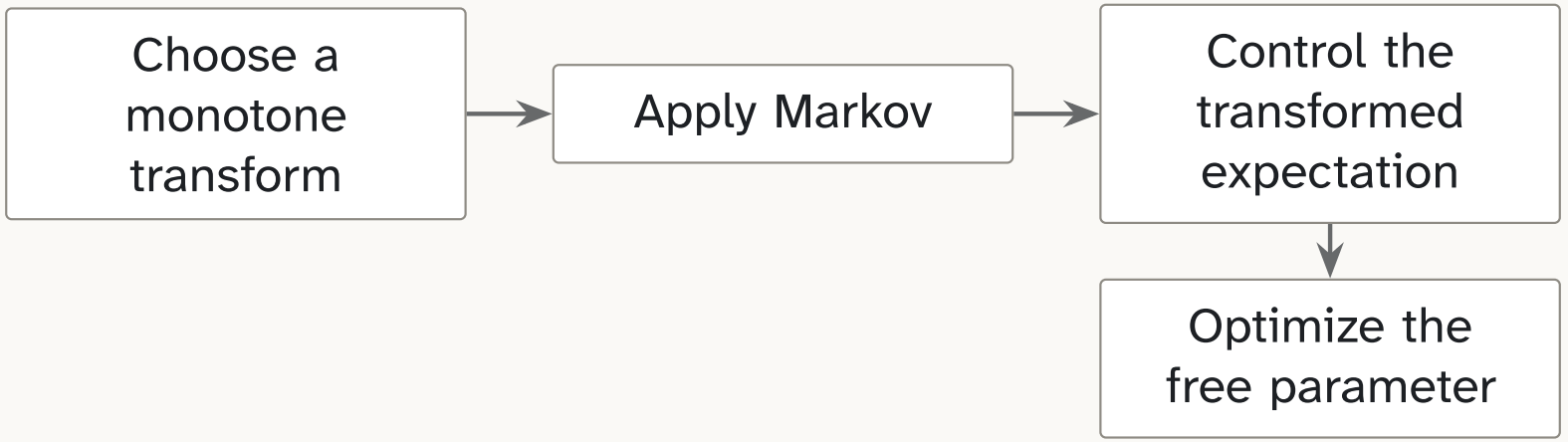
Chernoff: 1.1×10^5, Hoeffding: 1.8×10^6.

Relative accuracy is expensive for rare events: to estimate p relatively, the expected number of observed successes np must itself be large. Chernoff yields the natural $1/p$ dependence.

Markov + transform + optimization



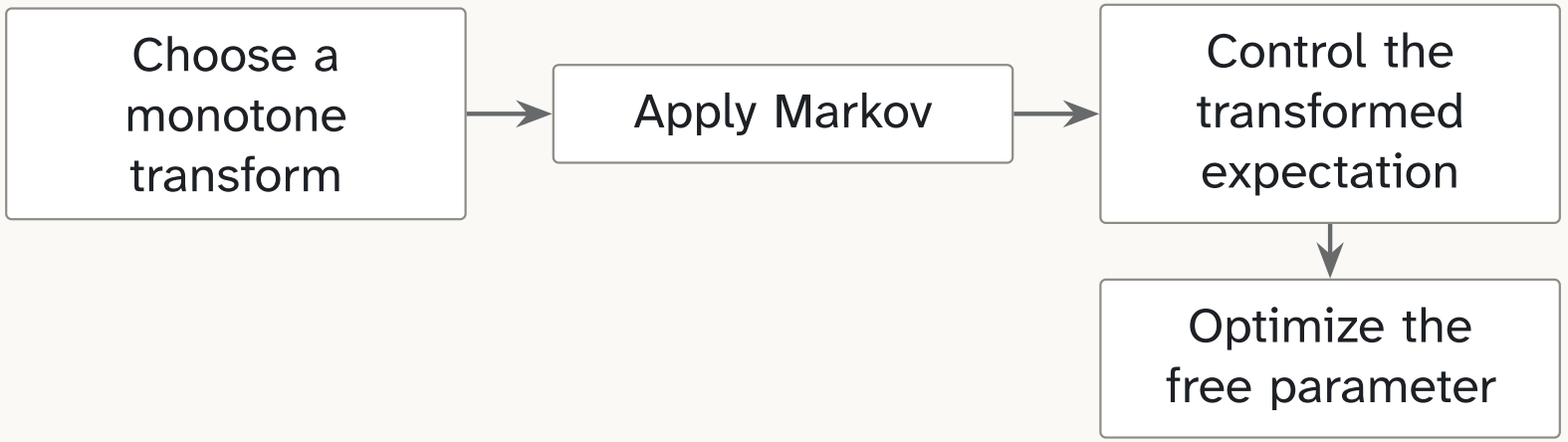
Markov + transform + optimization



Moment route:

$$|Y|^p \quad \to \quad \mathbb{E}|Y|^p \quad \to \quad \text{optimize } p.$$

Markov + transform + optimization



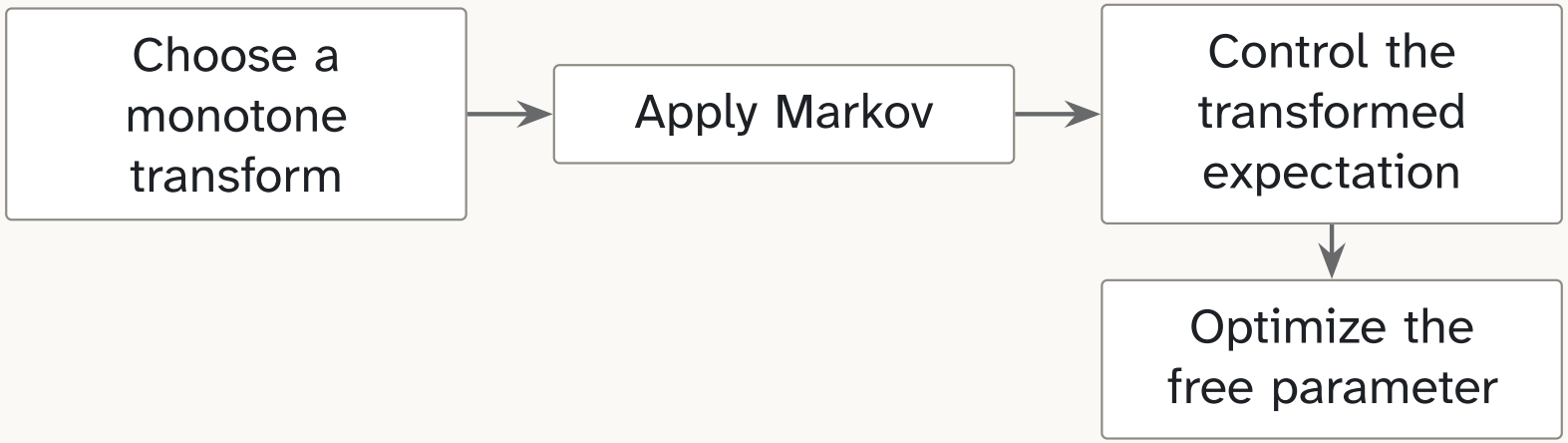
Moment route:

$$|Y|^p \quad \to \quad \mathbb{E}|Y|^p \quad \to \quad \text{optimize } p.$$

MGF route:

$$e^{\theta S} \quad \to \quad \mathbb{E}e^{\theta S} \quad \to \quad \text{optimize } \theta.$$

Markov + transform + optimization



Moment route:

$$|Y|^p \quad \to \quad \mathbb{E}|Y|^p \quad \to \quad \text{optimize } p.$$

MGF route:

$$e^{\theta S} \quad \to \quad \mathbb{E}e^{\theta S} \quad \to \quad \text{optimize } \theta.$$

When choosing a concentration inequality, ask what information its transformed expectation uses and what information it discards.

Questions?

PART 8

Appendix: supporting results and extensions

8

Appendix contents

A. Immediate technical detours

- integer moment rounding
- Gaussian moment growth and tail optimality
- CLT versus large deviations
- subgaussian equivalences
- exact Chernoff rate and reference table

B. Valiant routing

- random intermediates
- bit fixing and edge-load calculation
- Chernoff plus union bound

C. Beyond independence

- variance control without independence
- MGF control without factorization

D. Beyond bounded variables

- local MGFs and subexponential tails
- Bernstein and Bennett inequalities
- comparison of concentration tools

If only integer moments are available

Suppose

$$\|Y\|_p \le K\sqrt{p}$$

is known only for integer $p \ge 2$. The continuous calculation suggests

$$p_\star = \frac{x^2}{eK^2}.$$

If only integer moments are available

Suppose

$$\|Y\|_p \le K\sqrt{p}$$

is known only for integer $p \ge 2$. The continuous calculation suggests

$$p_\star = \frac{x^2}{eK^2}.$$

When $p_\star \ge 2$, choose the largest available integer below it:

$$p = \lfloor p_\star \rfloor.$$

If only integer moments are available

Suppose

$$\|Y\|_p \le K\sqrt{p}$$

is known only for integer $p \ge 2$. The continuous calculation suggests

$$p_\star = \frac{x^2}{eK^2}.$$

When $p_\star \ge 2$, choose the largest available integer below it:

$$p = \lfloor p_\star \rfloor.$$

Then

$$\boxed{\frac{p_\star}{2} \le p \le p_\star.}$$

So rounding down loses at most a factor of 2 in the moment order.

Integer rounding only changes constants

With $p = \lfloor p_\star \rceil$,

Integer rounding only changes constants

With $p = \lfloor p_\star \rfloor$,

$$\begin{aligned}\Pr(|Y| \geq x) &\leq \left(\frac{K\sqrt{p}}{x}\right)^p \\ &\leq \left(\frac{K\sqrt{p_\star}}{x}\right)^p = e^{-p/2}.\end{aligned}$$

Integer rounding only changes constants

With $p = \lfloor p_\star \rfloor$,

$$\begin{aligned}\Pr(|Y| \geq x) &\leq \left(\frac{K\sqrt{p}}{x}\right)^p \\ &\leq \left(\frac{K\sqrt{p_\star}}{x}\right)^p = e^{-p/2}.\end{aligned}$$

Since $p \geq p_\star/2$,

$$\Pr(|Y| \geq x) \leq e^{-p_\star/4} = \boxed{\exp\left(-\frac{x^2}{4eK^2}\right)}.$$

Integer rounding only changes constants

With $p = \lfloor p_\star \rfloor$,

$$\Pr(|Y| \geq x) \leq \left(\frac{K\sqrt{p}}{x}\right)^p \leq \left(\frac{K\sqrt{p_\star}}{x}\right)^p = e^{-p/2}.$$

Since $p \geq p_\star/2$,

$$\Pr(|Y| \geq x) \leq e^{-p_\star/4} = \boxed{\exp\left(-\frac{x^2}{4eK^2}\right)}.$$

Thus integer-valued moment information still gives

$$\boxed{\Pr(|Y| \geq x) \leq e^{-cx^2/K^2}}$$

for a universal constant $c > 0$: rounding changes the constant, not the tail shape.

Gaussian moments grow like $\sqrt{p}$

For $Z \sim N(0,1)$ and integer $k \geq 1$,

$$\mathbb{E}Z^{2k} = (2k-1)!! = \frac{(2k)!}{2^k k!}.$$

Gaussian moments grow like $\sqrt{p}$

For $Z \sim N(0,1)$ and integer $k \geq 1$,

$$\mathbb{E}Z^{2k} = (2k-1)!! = \frac{(2k)!}{2^k k!}.$$

Stirling's estimate $m! \asymp \sqrt{m}\,(m/e)^m$ gives

$$\frac{(2k)!}{2^k k!} \asymp \sqrt{2}\left(\frac{2k}{e}\right)^k.$$

Gaussian moments grow like $\sqrt{p}$

For $Z \sim N(0,1)$ and integer $k \geq 1$,

$$\mathbb{E}Z^{2k} = (2k-1)!! = \frac{(2k)!}{2^k k!}.$$

Stirling's estimate $m! \asymp \sqrt{m}\,(m/e)^m$ gives

$$\frac{(2k)!}{2^k k!} \asymp \sqrt{2}\left(\frac{2k}{e}\right)^k.$$

Taking the $2k$th root,

$$\|Z\|_{2k} \asymp \sqrt{\frac{2k}{e}} \lesssim \sqrt{2k}.$$

Gaussian moments grow like $\sqrt{p}$

For $Z \sim N(0,1)$ and integer $k \geq 1$,

$$\mathbb{E}Z^{2k} = (2k-1)!! = \frac{(2k)!}{2^k k!}.$$

Stirling's estimate $m! \asymp \sqrt{m}\,(m/e)^m$ gives

$$\frac{(2k)!}{2^k k!} \asymp \sqrt{2}\left(\frac{2k}{e}\right)^k.$$

Taking the $2k$th root,

$$\|Z\|_{2k} \asymp \sqrt{\frac{2k}{e}} \lesssim \sqrt{2k}.$$

Monotonicity of L_p norms extends the estimate to all real $p \geq 1$:

$$\boxed{\|Z\|_p \lesssim \sqrt{p}.}$$

The CLT does not prove large deviations

For i.i.d. X_i with mean μ and variance σ^2,

$$Z_n = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

The CLT does not prove large deviations

For i.i.d. X_i with mean μ and variance σ^2,

$$Z_n = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

The CLT controls probabilities at a *fixed* standardized threshold z.

The CLT does not prove large deviations

For i.i.d. X_i with mean μ and variance σ^2,

$$Z_n = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

The CLT controls probabilities at a *fixed* standardized threshold z.
A fixed fractional deviation of the sum corresponds to

$$z_n \asymp \sqrt{n},$$

which grows with n.

The CLT does not prove large deviations

For i.i.d. X_i with mean μ and variance σ^2,

$$Z_n = \frac{\sum_i X_i - n\mu}{\sigma\sqrt{n}} \Rightarrow N(0,1).$$

The CLT controls probabilities at a *fixed* standardized threshold z.
A fixed fractional deviation of the sum corresponds to

$$z_n \asymp \sqrt{n},$$

which grows with n.
Therefore the CLT suggests Gaussian-looking concentration, but does not by itself justify an e^{-cn} tail for that event. Hoeffding and Chernoff provide the needed finite-n bounds.

Gaussian tails make the exponent unavoidable

For $G \sim N(0,1)$ and $x \geq 1$, the Gaussian Mills bounds give

$$\sqrt{\frac{2}{\pi}} \frac{x}{1+x^2} e^{-x^2/2} \leq \Pr(|G| \geq x) \leq 2e^{-x^2/2}.$$

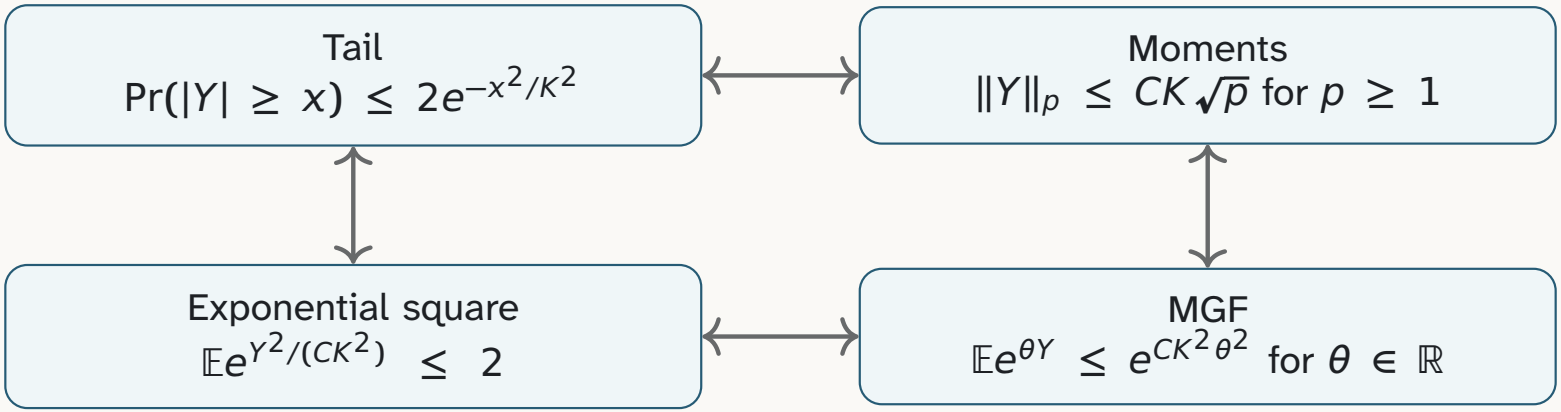
The factor in front of $e^{-x^2/2}$ is only polynomial in x. Thus a uniformly faster exponential rate cannot follow from Gaussian moment growth alone.

What "optimal up to constants" means

Some subgaussian variables have lighter tails—a bounded variable eventually has no tail at all. The Gaussian shows that no stronger exponent is valid for the entire class.

Subgaussianity has several equivalent forms

For centered Y, each condition below implies the other three after changing the numerical constant multiplying K.



The main lecture used the moments $\Longleftrightarrow$ tails characterization. In fact, subgaussianity has several equivalent formulations.

The exact Bernoulli rate has two regimes

The exact Chernoff exponent is

$$h(\delta) = (1+\delta)\log(1+\delta) - \delta.$$

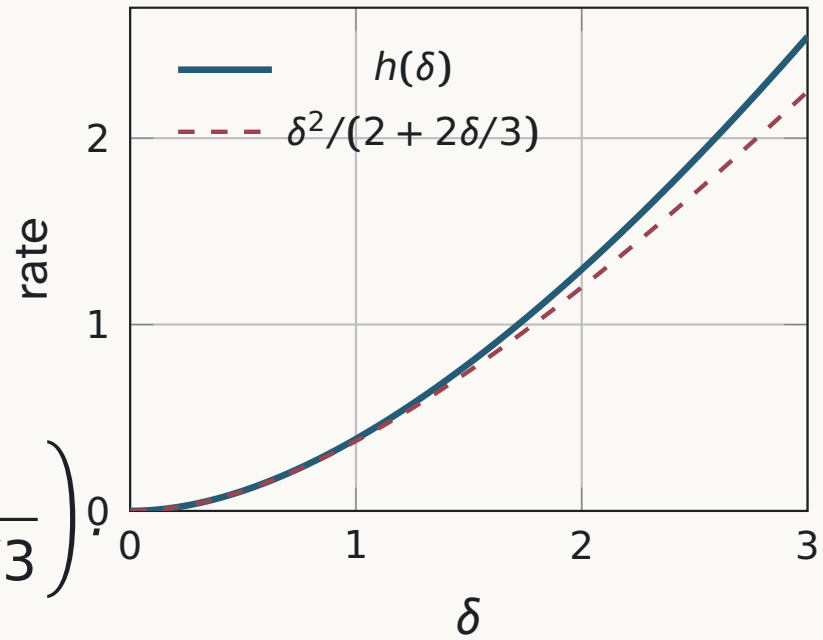
For every $\delta \geq 0$,

$$h(\delta) \geq \frac{\delta^2}{2 + 2\delta/3}.$$

Consequently,

$$\Pr(S \geq (1+\delta)\mu) \leq \exp\left(-\frac{\mu\delta^2}{2+2\delta/3}\right).$$

For $0 \leq \delta \leq 1$, the simpler lower bound $h(\delta) \geq \delta^2/3$ is often enough.



Reference table for the main bounds

Information	Method	Typical tail
$\mathbb{E}\lvert Y\rvert$	moment Markov	x^{-1}
$\mathrm{Var}(Y)$	Chebyshev	x^{-2}
$\lVert Y\rVert_p \le K\sqrt{p}$	optimize p	$e^{-x^2/(2eK^2)}$
independent bounded summands	Hoeffding	$e^{-2x^2/\sum_i (b_i-a_i)^2}$
Bernoulli sum with mean μ	multiplicative Chernoff	$e^{-\mu h(\delta)}$
local quadratic MGF	local-MGF Chernoff	$e^{-c\min\{x^2/\nu^2, x/b\}}$

Random routing breaks adversarial alignment

In a d-dimensional hypercube, the $N = 2^d$ nodes are bit strings, and an edge changes one bit.
The **congestion** is the largest number of packet routes using any one directed edge. An adversarial traffic pattern can force many direct routes to align.

Random routing breaks adversarial alignment

In a d-dimensional hypercube, the $N = 2^d$ nodes are bit strings, and an edge changes one bit.
The **congestion** is the largest number of packet routes using any one directed edge. An adversarial traffic pattern can force many direct routes to align.
Valiant's scheme gives every packet an independent uniformly random intermediate node:

$$\text{source} \longrightarrow \text{random intermediate} \longrightarrow \text{destination}.$$

Random routing breaks adversarial alignment

In a d-dimensional hypercube, the $N = 2^d$ nodes are bit strings, and an edge changes one bit.
The **congestion** is the largest number of packet routes using any one directed edge. An adversarial traffic pattern can force many direct routes to align.
Valiant's scheme gives every packet an independent uniformly random intermediate node:

$$\text{source} \longrightarrow \text{random intermediate} \longrightarrow \text{destination}.$$

Each leg uses **bit fixing**: change the disagreeing coordinates in a fixed order.

Random routing breaks adversarial alignment

In a d-dimensional hypercube, the $N = 2^d$ nodes are bit strings, and an edge changes one bit.
The **congestion** is the largest number of packet routes using any one directed edge. An adversarial traffic pattern can force many direct routes to align.
Valiant's scheme gives every packet an independent uniformly random intermediate node:

$$\text{source} \longrightarrow \text{random intermediate} \longrightarrow \text{destination}.$$

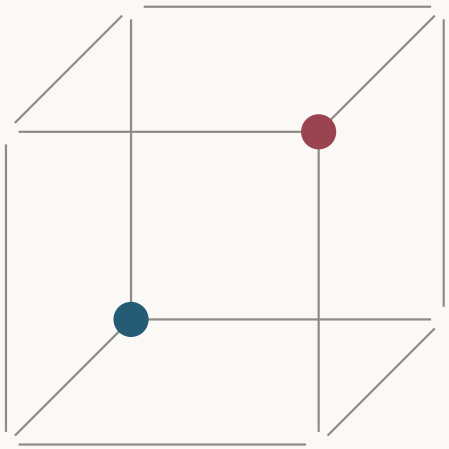
Each leg uses **bit fixing**: change the disagreeing coordinates in a fixed order.

> The random intermediate breaks the adversary's ability to align many source–destination routes along the same edges.

Bit fixing changes one coordinate at a time

In a 3-cube, route from 001 to 110 by fixing coordinates from left to right:

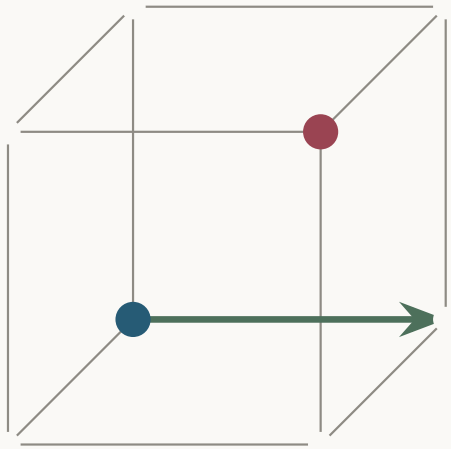
001 .



Bit fixing changes one coordinate at a time

In a 3-cube, route from 001 to 110 by fixing coordinates from left to right:

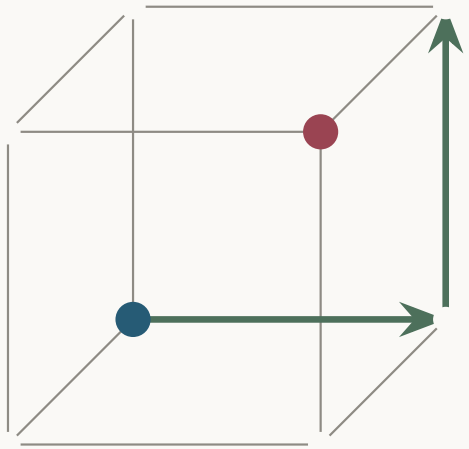
$$001 \longrightarrow 101 \qquad .$$



Bit fixing changes one coordinate at a time

In a 3-cube, route from 001 to 110 by fixing coordinates from left to right:

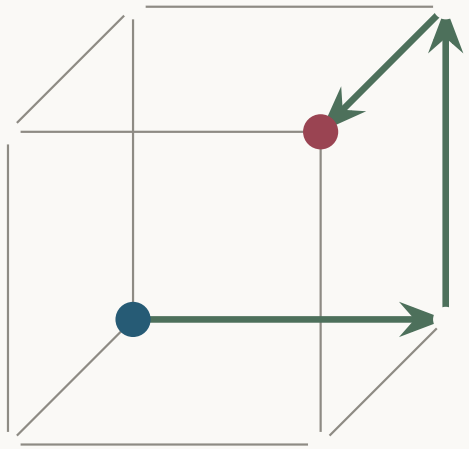
$$001 \longrightarrow 101 \longrightarrow 111 \quad .$$



Bit fixing changes one coordinate at a time

In a 3-cube, route from 001 to 110 by fixing coordinates from left to right:

$$001 \longrightarrow 101 \longrightarrow 111 \longrightarrow 110.$$



At each step, the route fixes one more bit to match the destination.

One edge calculation can be checked by hand

Fix the directed edge

$$e : 000 \longrightarrow 010,$$

which changes coordinate $i = 2$ under left-to-right bit fixing.

One edge calculation can be checked by hand

Fix the directed edge

$$e : 000 \longrightarrow 010,$$

which changes coordinate $i = 2$ under left-to-right bit fixing.
To reach the tail 000 immediately before fixing coordinate 2:

- the source has the form $*00$, giving 000 or 100;
- the intermediate has the form $01*$, giving 010 or 011.

One edge calculation can be checked by hand

Fix the directed edge

$$e : 000 \longrightarrow 010,$$

which changes coordinate $i = 2$ under left-to-right bit fixing.
To reach the tail 000 immediately before fixing coordinate 2:

- the source has the form $*00$, giving 000 or 100;
- the intermediate has the form $01*$, giving 010 or 011.

A uniform intermediate has prefix 01 with probability

$$\frac{2}{8} = \frac{1}{4}.$$

One edge calculation can be checked by hand

Fix the directed edge

$$e : 000 \longrightarrow 010,$$

which changes coordinate $i = 2$ under left-to-right bit fixing.
To reach the tail 000 immediately before fixing coordinate 2:

- the source has the form $*00$, giving 000 or 100;
- the intermediate has the form $01*$, giving 010 or 011.

A uniform intermediate has prefix 01 with probability

$$\frac{2}{8} = \frac{1}{4}.$$

Therefore the expected first-leg load of e is

$$2 \text{ possible sources} \times \frac{1}{4} = \frac{1}{2}.$$

The same counting works in dimension d

Fix a directed edge e in coordinate i.

The same counting works in dimension d

Fix a directed edge e in coordinate i.
A source can use e only if its suffix agrees with the tail of e. The first $i-1$ source bits are unrestricted, so there are

$$2^{i-1}$$

possible sources.

The same counting works in dimension d

Fix a directed edge e in coordinate i.
A source can use e only if its suffix agrees with the tail of e. The first $i-1$ source bits are unrestricted, so there are

$$2^{i-1}$$

possible sources.
The random intermediate must have a prescribed prefix of length i, which occurs with probability

$$2^{-i}.$$

The same counting works in dimension d

Fix a directed edge e in coordinate i.
A source can use e only if its suffix agrees with the tail of e. The first $i-1$ source bits are unrestricted, so there are

$$2^{i-1}$$

possible sources.
The random intermediate must have a prescribed prefix of length i, which occurs with probability

$$2^{-i}.$$

Hence the expected first-leg load is

$$2^{i-1}2^{-i} = \frac{1}{2}.$$

The second leg contributes another $1/2$, so

$$\boxed{\mathbb{E}L_e \le 1.}$$

The same counting works in dimension d

Fix a directed edge e in coordinate i.
A source can use e only if its suffix agrees with the tail of e. The first $i-1$ source bits are unrestricted, so there are

$$2^{i-1}$$

possible sources.
The random intermediate must have a prescribed prefix of length i, which occurs with probability

$$2^{-i}.$$

Hence the expected first-leg load is

$$2^{i-1}2^{-i} = \frac{1}{2}.$$

The second leg contributes another $1/2$, so

$$\boxed{\mathbb{E}L_e \le 1.}$$

The same argument applies to the second leg after exchanging the roles of source and destination.

Chernoff controls routing congestion

For each packet s, let

$$Y_s = \mathbf{1}_{\{\text{packet } s \text{ traverses } e\}}, \qquad L_e = \sum_s Y_s.$$

The random intermediates are independent, so the Y_s are independent and $\mathbb{E}L_e \le 1$.
For $k \ge 1$, Chernoff gives

$$\Pr(L_e \ge k) \le \left(\frac{e}{k}\right)^k.$$

There are $N \log_2 N$ directed edges. A union bound gives the explicit estimate

$$\boxed{\Pr\left(\max_e L_e \ge k\right) \le N \log_2 N \left(\frac{e}{k}\right)^k.}$$

Choose k so that this displayed expression is below the desired failure probability. In particular, a sufficiently large multiple of $\log N$ makes it very small.

Variance control does not require full independenc

For Chebyshev, what we need is control of

$$\mathrm{Var}\left(\sum_i X_i\right).$$

Variance control does not require full independenc

For Chebyshev, what we need is control of

$$\mathrm{Var}\left(\sum_i X_i\right).$$

Independence is one way to get

$$\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i).$$

Variance control does not require full independenc

For Chebyshev, what we need is control of

$$\operatorname{Var}\left(\sum_i X_i\right).$$

Independence is one way to get

$$\operatorname{Var}\left(\sum_i X_i\right) = \sum_i \operatorname{Var}(X_i).$$

But pairwise nonpositive covariance is already enough for

$$\operatorname{Var}\left(\sum_i X_i\right) \leq \sum_i \operatorname{Var}(X_i).$$

Variance control does not require full independenc

For Chebyshev, what we need is control of

$$\operatorname{Var}\left(\sum_i X_i\right).$$

Independence is one way to get

$$\operatorname{Var}\left(\sum_i X_i\right) = \sum_i \operatorname{Var}(X_i).$$

But pairwise nonpositive covariance is already enough for

$$\operatorname{Var}\left(\sum_i X_i\right) \leq \sum_i \operatorname{Var}(X_i).$$

So the proof tells us what structure matters: a usable variance estimate for the sum.

Chernoff needs MGF control, not independence

For the Chernoff method, what we need is a useful bound on

$$\mathbb{E}e^{\theta \sum_i X_i}.$$

Chernoff needs MGF control, not independence

For the Chernoff method, what we need is a useful bound on

$$\mathbb{E}e^{\theta \sum_i X_i}.$$

Independence gives the exact factorization

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

Chernoff needs MGF control, not independence

For the Chernoff method, what we need is a useful bound on

$$\mathbb{E}e^{\theta \sum_i X_i}.$$

Independence gives the exact factorization

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

But the same proof works whenever we can establish, for the relevant θ,

$$\boxed{\mathbb{E}e^{\theta \sum_i X_i} \leq \prod_i \mathbb{E}e^{\theta X_i}.}$$

Chernoff needs MGF control, not independence

For the Chernoff method, what we need is a useful bound on

$$\mathbb{E}e^{\theta \sum_i X_i}.$$

Independence gives the exact factorization

$$\mathbb{E}e^{\theta \sum_i X_i} = \prod_i \mathbb{E}e^{\theta X_i}.$$

But the same proof works whenever we can establish, for the relevant θ,

$$\boxed{\mathbb{E}e^{\theta \sum_i X_i} \leq \prod_i \mathbb{E}e^{\theta X_i}.}$$

Certain forms of negative dependence imply inequalities of this type. Pairwise negative covariance alone does *not* generally suffice.

Unbounded data can still concentrate

Boundedness is sufficient for Hoeffding, but it is not necessary for exponential concentration.
For example, service times and job lengths may be unbounded but have exponentially decreasing tails.
Let

$$E \sim \mathrm{Exp}(1), \qquad Y = E - 1.$$

Then $\mathbb{E}Y = 0$, but Y is not bounded above.

What replaces boundedness?

We will directly control the MGF, although only for tilts θ in a bounded interval around zero.

A centered exponential has a local MGF bound

For $0 < \theta < 1$,

$$\begin{aligned}\mathbb{E}e^{\theta Y} &= \mathbb{E}e^{\theta(E-1)} \\ &= \frac{e^{-\theta}}{1-\theta} \\ &= \exp\bigl(-\theta - \log(1-\theta)\bigr).\end{aligned}$$

For $0 < \theta \le 1/2$,

$$-\theta - \log(1-\theta) = \sum_{j \ge 2} \frac{\theta^j}{j} \le \frac{\theta^2}{2(1-\theta)} \le \theta^2,$$

so

$$\boxed{\mathbb{E}e^{\theta Y} \le e^{\theta^2}.}$$

The quadratic bound resembles subgaussian behavior, but it holds only for $\theta \le 1/2$.

The optimizer hits the MGF boundary

For independent copies $Y_1, \ldots, Y_n$ and $S = \sum_i Y_i$,

$$\Pr(S \geq x) \leq \exp(n\theta^2 - \theta x), \qquad 0 < \theta \leq \frac{1}{2}.$$

Differentiating gives the unconstrained optimizer

$$\theta^\star = \frac{x}{2n}.$$

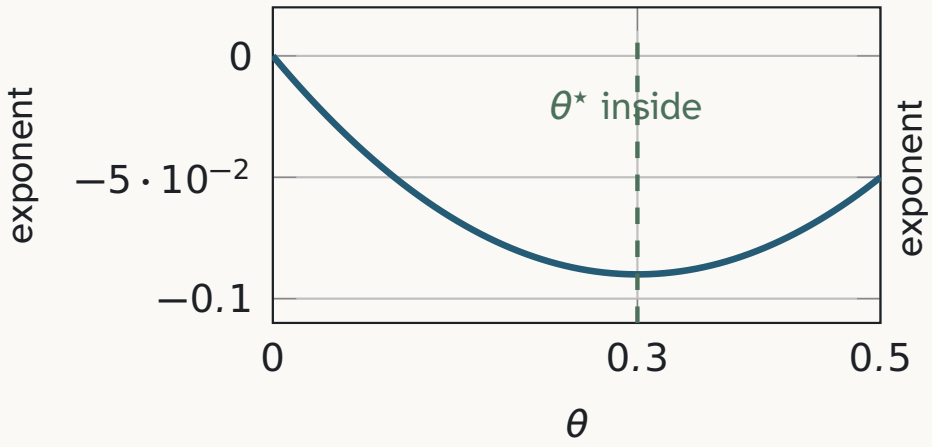
The optimizer hits the MGF boundary

For independent copies $Y_1, \ldots, Y_n$ and $S = \sum_i Y_i$,

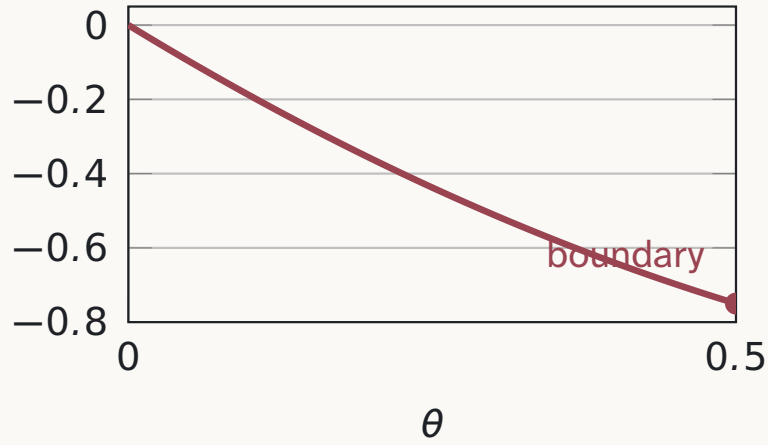
$$\Pr(S \geq x) \leq \exp(n\theta^2 - \theta x), \qquad 0 < \theta \leq \frac{1}{2}.$$

Differentiating gives the unconstrained optimizer

$$\theta^\star = \frac{x}{2n}.$$



$x = .6n$: use $\theta^\star$



$x = 2n$: stop at $\theta = 1/2$

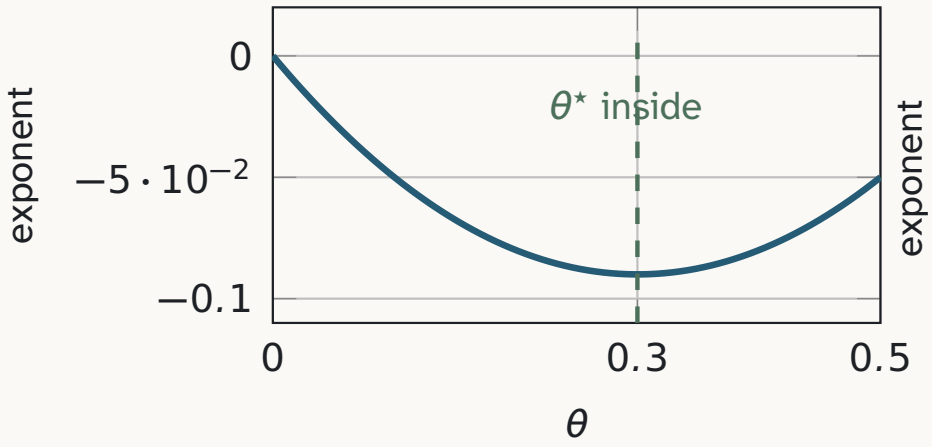
The optimizer hits the MGF boundary

For independent copies $Y_1, \ldots, Y_n$ and $S = \sum_i Y_i$,

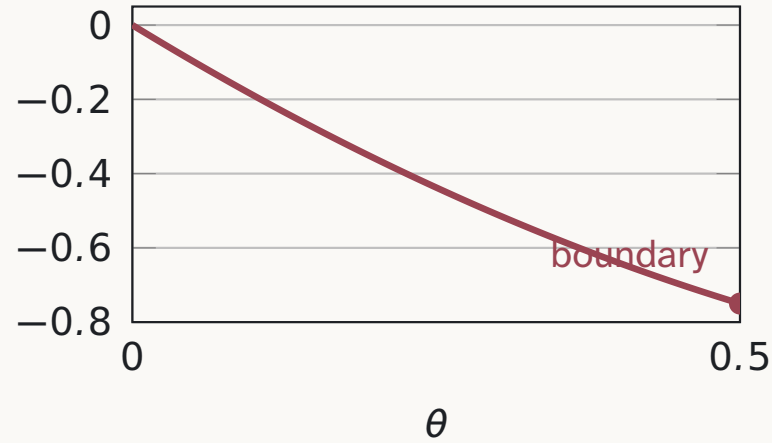
$$\Pr(S \geq x) \leq \exp(n\theta^2 - \theta x), \qquad 0 < \theta \leq \frac{1}{2}.$$

Differentiating gives the unconstrained optimizer

$$\theta^\star = \frac{x}{2n}.$$



$x = .6n$: use $\theta^\star$



$x = 2n$: stop at $\theta = 1/2$

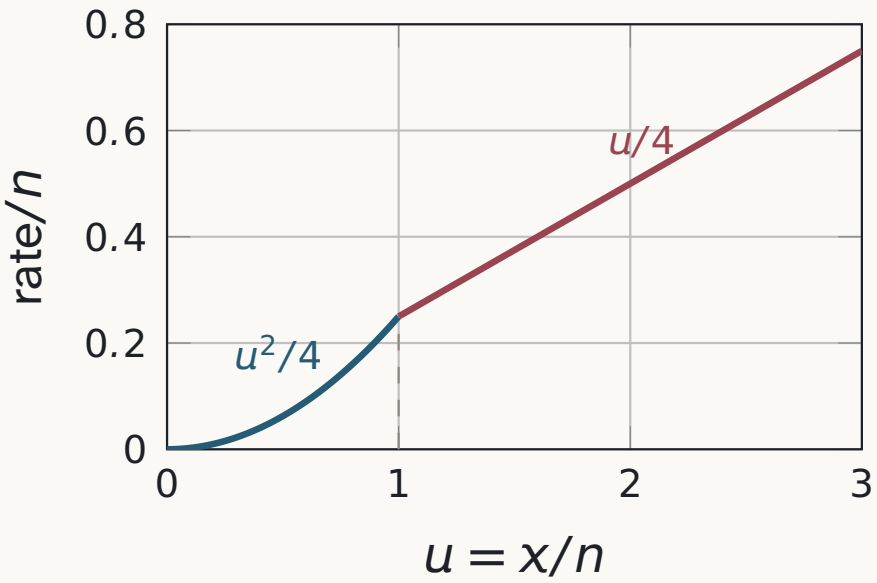
Thus $x \leq n$ gives exponent $-x^2/(4n)$, while $x > n$ gives exponent at most $-x/4$.

Subexponential sums have two regimes

Combining the two cases gives

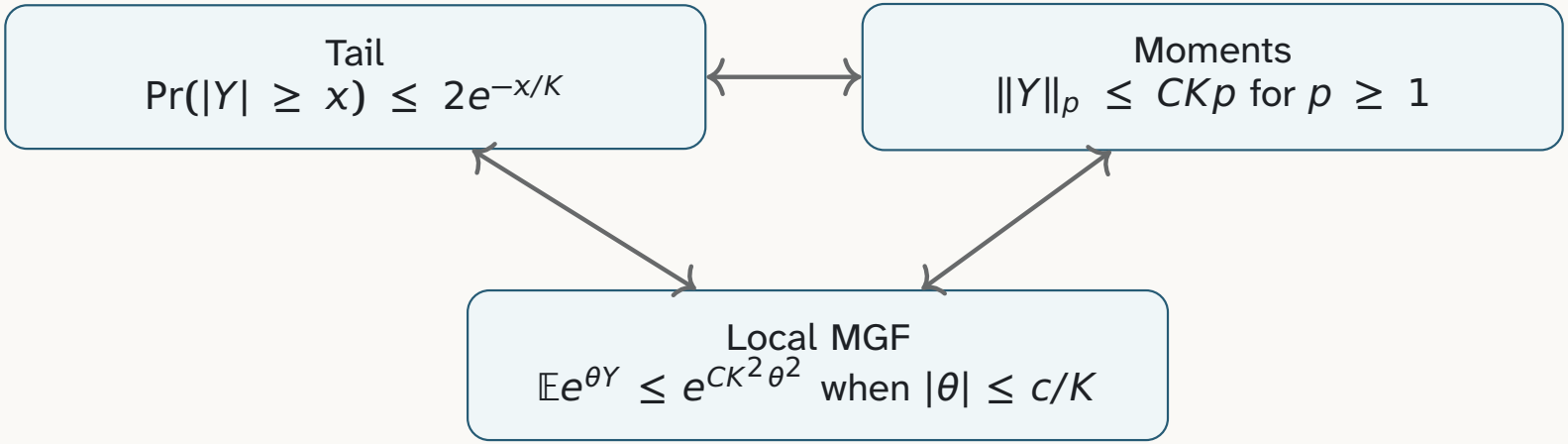
$$\boxed{\Pr(S \geq x) \leq \exp\left[-\frac{1}{4}\min\left\{\frac{x^2}{n}, x\right\}\right].}$$

- $x \leq n$: Gaussian moderate deviations, $e^{-x^2/(4n)}$;
- $x > n$: exponential large deviations, at most $e^{-x/4}$.



Subexponentiality also has equivalent forms

For centered Y, the following are equivalent after changing numerical constants:



The contrast with subgaussian variables is the MGF domain: all tilts for subgaussian variables, only a neighborhood of zero for subexponential variables.

Moment growth determines the tail shape

Suppose that for some $\alpha > 0$,

$$\|Y\|_p \le K p^{1/\alpha}, \qquad p \ge 1.$$

Optimizing

$$\Pr(|Y| \ge x) \le \inf_{p \ge 1} \left(\frac{K p^{1/\alpha}}{x} \right)^p$$

uses

$$p_\star = e^{-1} \left(\frac{x}{K} \right)^\alpha .$$

When $p_\star \ge 1$, substitution gives the explicit bound

$$\Pr(|Y| \ge x) \le \boxed{\exp\left[-\frac{1}{\alpha e} \left(\frac{x}{K} \right)^\alpha \right].}$$

$\alpha = 2$	$\alpha = 1$	$0 < \alpha < 1$
subgaussian	subexponential	stretched exponential

A Bernstein moment condition bounds the MGF

Let $X_1, \dots, X_n$ be independent and centered. Suppose that for every integer $k \geq 2$,

$$\mathbb{E}|X_i|^k \leq \frac{k!}{2}\sigma_i^2 b^{k-2}.$$

A Bernstein moment condition bounds the MGF

Let $X_1, \dots, X_n$ be independent and centered. Suppose that for every integer $k \geq 2$,

$$\mathbb{E}|X_i|^k \leq \frac{k!}{2}\sigma_i^2 b^{k-2}.$$

For $0 < \theta < 1/b$, the moment expansion gives

$$\begin{aligned}
\mathbb{E}e^{\theta X_i} &\leq 1 + \frac{\sigma_i^2\theta^2}{2}\sum_{k\geq 2}(b\theta)^{k-2} \\
&= 1 + \frac{\sigma_i^2\theta^2}{2(1-b\theta)} \\
&\leq \exp\left(\frac{\sigma_i^2\theta^2}{2(1-b\theta)}\right).
\end{aligned}$$

Independence combines the local MGF bounds

Writing $\nu^2 = \sum_i \sigma_i^2$ and using independence gives

$$\begin{aligned}
\mathbb{E}e^{\theta \sum_i X_i} &= \prod_i \mathbb{E}e^{\theta X_i} \\
&\leq \prod_i \exp\left(\frac{\sigma_i^2 \theta^2}{2(1 - b\theta)} \right) \\
&= \exp\left(\frac{\nu^2 \theta^2}{2(1 - b\theta)} \right).
\end{aligned}$$

Thus, for $0 < \theta < 1/b$,

$$\boxed{\mathbb{E}e^{\theta \sum_i X_i} \leq \exp\left(\frac{\nu^2 \theta^2}{2(1 - b\theta)} \right).}$$

A convenient tilt gives Bernstein's bound

The Laplace–Chernoff method gives, for $0 < \theta < 1/b$,

$$\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left(-\theta x + \frac{\nu^2 \theta^2}{2(1 - b\theta)}\right).$$

A convenient tilt gives Bernstein's bound

The Laplace–Chernoff method gives, for $0 < \theta < 1/b$,

$$\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left(-\theta x + \frac{\nu^2 \theta^2}{2(1-b\theta)}\right).$$

Choose

$$\theta = \frac{x}{\nu^2 + bx}, \qquad 1 - b\theta = \frac{\nu^2}{\nu^2 + bx}.$$

A convenient tilt gives Bernstein's bound

The Laplace–Chernoff method gives, for $0 < \theta < 1/b$,

$$\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left(-\theta x + \frac{\nu^2 \theta^2}{2(1 - b\theta)}\right).$$

Choose

$$\theta = \frac{x}{\nu^2 + bx}, \qquad 1 - b\theta = \frac{\nu^2}{\nu^2 + bx}.$$

This choice is always in $(0, 1/b)$, so it is permitted by the local MGF bound.

Substitution gives Bernstein's inequality

Use

$$\theta = \frac{x}{\nu^2 + bx}, \qquad 1 - b\theta = \frac{\nu^2}{\nu^2 + bx}.$$

Substitution yields

$$\boxed{\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left(-\frac{x^2}{2(\nu^2 + bx)}\right).}$$

Substitution gives Bernstein's inequality

Use

$$\theta = \frac{x}{\nu^2 + bx}, \qquad 1 - b\theta = \frac{\nu^2}{\nu^2 + bx}.$$

Substitution yields

$$\boxed{\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left(-\frac{x^2}{2(\nu^2 + bx)}\right).}$$

In particular,

$$\Pr\left(\sum_i X_i \geq x\right) \leq \exp\left[-\frac{1}{4}\min\left\{\frac{x^2}{\nu^2}, \frac{x}{b}\right\}\right].$$

Bennett retains a sharper rate function

Suppose instead that the independent centered variables satisfy $X_i \le b$ almost surely, and let

$$\nu^2 = \sum_i \operatorname{Var}(X_i).$$

Bennett's inequality states

$$\boxed{\Pr\left(\sum_i X_i \ge x\right) \le \exp\left[-\frac{\nu^2}{b^2} h\left(\frac{bx}{\nu^2}\right)\right],}$$

where $h(u) = (1+u)\log(1+u) - u$.

> Bernstein replaces this exact curved rate by a simpler quadratic-over-linear expression; Bennett can be sharper for large deviations.

Hoeffding, Bernstein, and Bennett differ

Bound	Information	Advantage	Cost
Hoeffding	ranges $[a_i, b_i]$	simple and distribution-free	does not use variances
Bernstein	variances plus a range or moment bound	uses small variance	more assumptions
Bennett	variance plus an upper bound	strong large-deviation rate	less convenient

Selection rule

Use the strongest structural information that is genuinely available. A sharper formula is not useful unless its assumptions have been verified.