Concentration Inequalities: Quick Reference

CSE 493S: Advanced Topics in Machine Learning

Autumn 2025

1 Overview

Concentration inequalities bound the probability that a random variable deviates from its expectation. They are fundamental tools in learning theory for proving generalization bounds, analyzing algorithms, and understanding sample complexity.

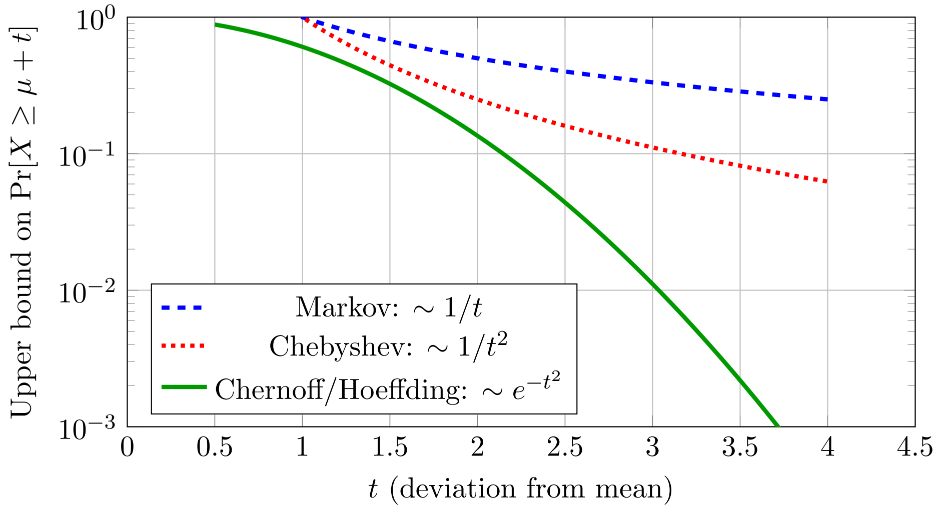


Figure 1: Tail bound comparison - exponential bounds (Chernoff/Hoeffding) decay much faster than polynomial bounds (Markov 1/t, Chebyshev 1/t^2)

2 The Three Inequalities

2.1 Markov's Inequality

Markov's Inequality

Theorem 1 (Markov)**.** *Let X be a non-negative random variable. Then for any $a > 0$:*

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Requirements:

- $X \geq 0$ (non-negative, i.e., lower bounded at 0)

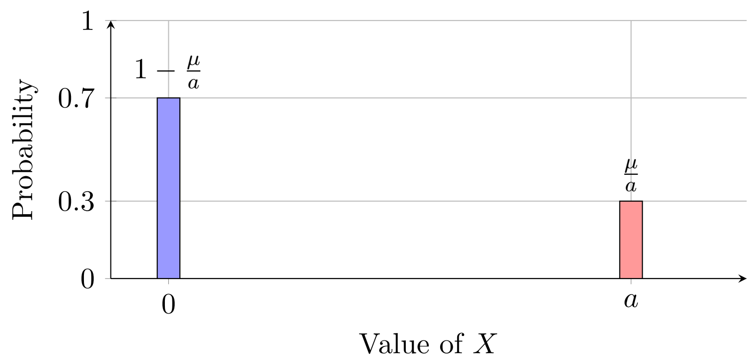
- Only need to know $\mathbb{E}[X]$
- No independence assumptions
- No upper boundedness assumptions needed

When to use:

- Simplest bound, works for any non-negative random variable
- When you only know the mean (no variance information)
- Quick rough bounds (often loose)

Example: If the average test score is 70, at most 70% of students can score ≥ 100.

Why Markov is tight: The bound cannot be improved in general. Consider the distribution that puts all mass at 0 and a:



Example: $X = 0$ with prob $1 - \mu/a$ and $X = a$ with prob μ/a achieves Markov's bound with equality

This distribution has $\mathbb{E}[X] = \mu$ and $\Pr[X \geq a] = \mu/a$, exactly matching Markov's bound. Since some distribution achieves the bound, you cannot do better without more information.

2.2 Chebyshev's Inequality

Chebyshev's Inequality

Theorem 2 (Chebyshev)**.** *Let X be a random variable with mean μ and variance σ^2. Then for any $t > 0$:*

$$\Pr[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}$$

Requirements:

- Need to know $\mathbb{E}[X]$ and $\text{Var}(X)$ to *apply* the bound
- No distribution assumptions
- No independence needed
- No boundedness assumptions

When to use:

- You have variance information (otherwise the bound σ^2/t^2 is meaningless)
- Better than Markov (quadratic vs linear decay)
- Distribution-free bounds

Key insight: Derived by applying Markov to $(X - \mu)^2 \geq t^2$.

Note: Technically, Chebyshev's inequality holds for any random variable with finite variance, but you need to know σ^2 to use the bound numerically. Without knowing the variance, the inequality tells you nothing quantitative.

2.3 Chernoff Bound

Chernoff Bound (Multiplicative Form)

Theorem 3 (Chernoff - Multiplicative). *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [0, 1]$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then for any $\epsilon > 0$:*

$$\Pr[X \geq (1+\epsilon)\mu] \leq e^{-\frac{\epsilon^2 \mu}{3}} \quad \textit{(upper tail)}$$

$$\Pr[X \leq (1-\epsilon)\mu] \leq e^{-\frac{\epsilon^2 \mu}{2}} \quad \textit{(lower tail)}$$

Multiplicative form: Bounds the *relative* error - useful when you care about deviation as a fraction of the mean.

Chernoff Bound (Additive Form)

Theorem 4 (Chernoff - Additive). *Let $X_1, \ldots, X_n$ be independent random variables with $X_i \in [0, 1]$. Let $X = \sum_{i=1}^{n} X_i$ and $\mu = \mathbb{E}[X]$. Then for any $\epsilon > 0$:*

$$\Pr[X \geq \mu + \epsilon] \leq e^{-\frac{2\epsilon^2}{n}}$$

$$\Pr[X \leq \mu - \epsilon] \leq e^{-\frac{2\epsilon^2}{n}}$$

$$\Pr[|X - \mu| \geq \epsilon] \leq 2e^{-\frac{2\epsilon^2}{n}}$$

Additive form: Bounds the *absolute* error - useful when you care about fixed deviation regardless of the mean. Note: This is identical to Hoeffding's bound for $[0, 1]$ variables!

Requirements:

- **Independence** of X_i
- Bounded random variables (typically $[0, 1]$)
- Sum of independent variables

When to use:

- Sums of independent bounded random variables

- **Exponential decay** - much tighter than Markov/Chebyshev
- **Multiplicative form:** When relative error matters (e.g., "within 10% of mean")
- **Additive form:** When absolute error matters (e.g., "within ±0.1")
- Applications: coin flips, sampling, randomized algorithms

2.4 Hoeffding's Inequality

Hoeffding's Inequality

Theorem 5 (Hoeffding)**.** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [a_i, b_i]$. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for any $t > 0$:*

$$\Pr[|\overline{X} - \mathbb{E}[\overline{X}]| \geq t] \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

For $X_i \in [a, b]$ (same bounds), this simplifies to:

$$\Pr[|\overline{X} - \mathbb{E}[\overline{X}]| \geq t] \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right)$$

Requirements:

- **Independence** of X_i
- Bounded random variables $X_i \in [a_i, b_i]$
- Works for **empirical average**

When to use:

- Empirical averages (sample means)
- **Learning theory:** bounding training error vs test error
- **Exponential decay** with explicit constants
- Works even when variables have different ranges

2.5 Bernstein's Inequality

Bernstein's Inequality

Theorem 6 (Bernstein)**.** *Let $X_1, \dots, X_n$ be independent random variables with $X_i \in [a, b]$. Let $X = \sum_{i=1}^{n} X_i$, $\mu = \mathbb{E}[X]$, and $\sigma^2 = Var(X)$. Then for any $t > 0$:*

$$\Pr[X - \mu \geq t] \leq \exp\left(-\frac{t^2/2}{\sigma^2 + (b-a)t/3}\right)$$

$$\Pr[|X - \mu| \geq t] \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + (b-a)t/3}\right)$$

Requirements:

- **Independence** of X_i
- Bounded random variables $X_i \in [a, b]$
- Need to know (or bound) the **variance** σ^2

When Bernstein beats Hoeffding:

- When variance $\sigma^2 \ll n(b-a)^2$ (much smaller than the worst case)
- For small deviations t: Bernstein gives $\sim e^{-t^2/\sigma^2}$ vs Hoeffding's $\sim e^{-t^2/(b-a)^2}$
- For large deviations t: Both give similar exponential decay
- **Example:** Sum of n Bernoulli(p) with $p \ll 1$: variance $\sigma^2 = np(1-p) \ll n$

Trade-off:

- **Bernstein:** Tighter when you know variance is small, but requires variance information
- **Hoeffding:** Always works with just boundedness, but can be loose if variance is small

3 Visual Comparison: When to Use What

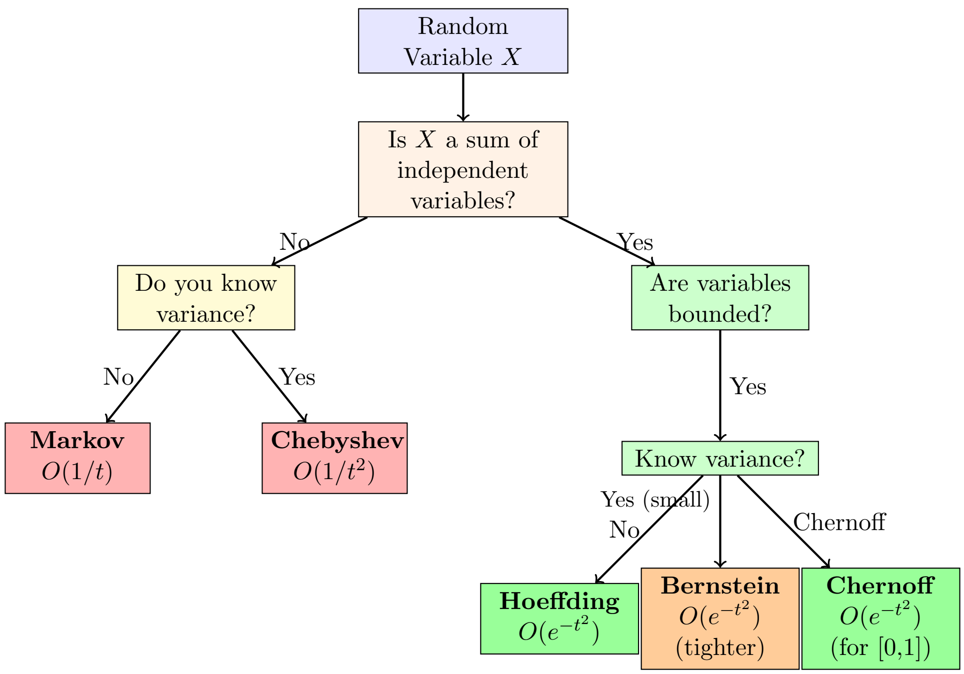


Figure 2: Decision tree for choosing the right concentration inequality

Property	Markov	Chebyshev	Bernstein	Chernoff/Hoeff.
Independence	Not needed	Not needed	**Required**	**Required**
Bounded vars	Not needed	Not needed	**Required**	**Required**
Variance info	Not needed	**Required**	**Required**	Not needed
Tail decay	$O(1/t)$	$O(1/t^2)$	$O(e^{-t^2})$	$O(e^{-t^2})$
Info needed	$\mathbb{E}[X]$	$\mathbb{E}[X], \sigma^2$	$\mathbb{E}[X], \sigma^2$, bounds	Bounds on X_i
Tightness	Loose	Medium	**Very tight**	**Tight**
Best when	Simple bound	No indep.	$\sigma^2 \ll n(b-a)^2$	General sums

Table 1: Comparison of concentration inequalities

4 Key Differences

4.1 Why Independence Matters

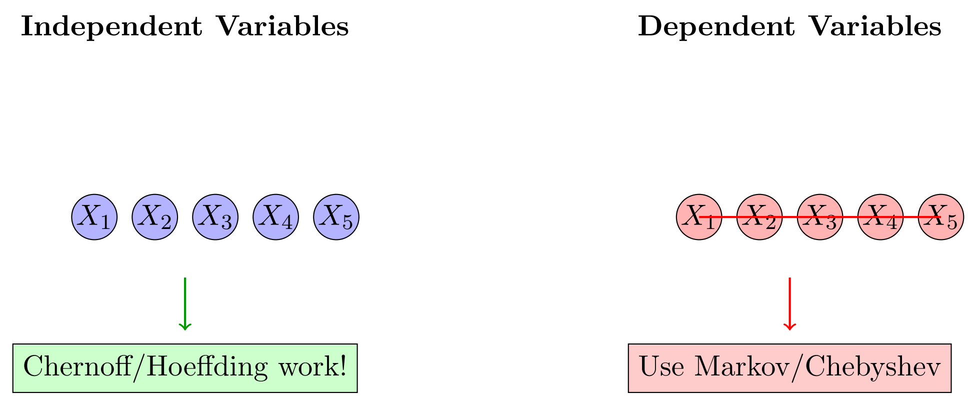


Figure 3: Independence is crucial for exponential concentration

Why? When variables are independent, deviations in different directions tend to cancel out. With dependence, they can all deviate together, giving worse concentration.

5 Typical Applications in Learning Theory

5.1 Uniform Convergence (Hoeffding)

Bounding the probability that empirical risk deviates from true risk:

Let $\ell(h, z_i) \in [0, 1]$ be the loss on sample z_i. By Hoeffding:

$$\Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i) - \mathbb{E}[\ell(h, Z)]\right| \geq \epsilon\right] \leq 2e^{-2n\epsilon^2}$$

5.2 PAC Learning (Chernoff/Hoeffding)

For n samples and error tolerance ϵ, confidence δ:

$$n \geq \frac{1}{2\epsilon^2} \log \frac{2}{\delta}$$

This comes from setting $2e^{-2n\epsilon^2} \leq \delta$ and solving for n.

5.3 Sample Complexity Visualization

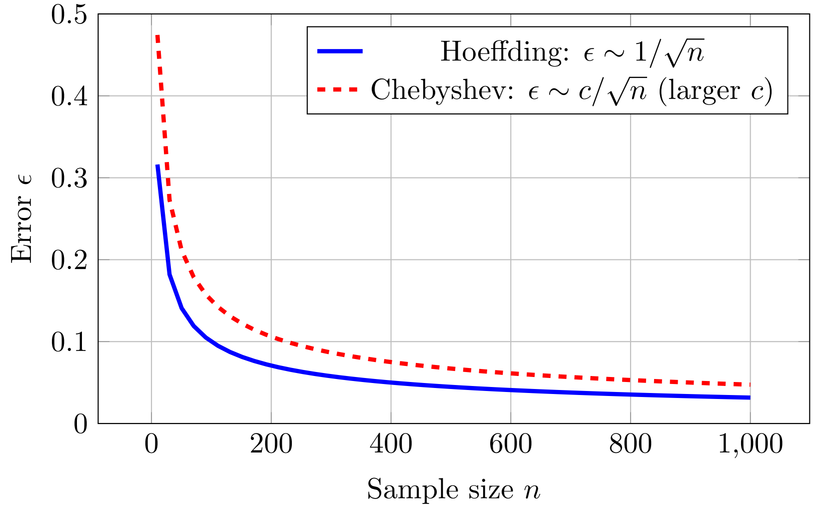


Figure 4: Error decreases as $O(1/\sqrt{n})$ - need 4x samples to halve error

6 Quick Reference: Which Bound to Use

Decision Guide

Use Markov when:

- You only know the mean
- Need a quick, rough bound
- Variables can be dependent or unbounded

Use Chebyshev when:

- You know the variance
- Variables might be dependent
- Want a distribution-free bound better than Markov

Use Chernoff when:

- Sum of **independent** Bernoulli or bounded $[0, 1]$ variables
- Want multiplicative error bounds (relative error)
- Analyzing randomized algorithms

Use Hoeffding when:

- Empirical averages of **independent** bounded variables
- Learning theory (training vs test error)
- Want additive error bounds with explicit constants
- Variables have different (but known) ranges

7 Common Pitfalls

1. **Forgetting independence:** Chernoff/Hoeffding require independent samples. If your data is correlated, these bounds don't apply.
2. **Wrong normalization:** Chernoff bounds $\sum X_i$, Hoeffding bounds $\frac{1}{n}\sum X_i$. Don't mix them up!
3. **Ignoring constants:** In practice, the constants matter. Hoeffding often gives better constants than Chernoff for [0,1] variables.
4. **One-sided vs two-sided:** Chernoff gives separate upper/lower tail bounds. Hoeffding is typically stated as a two-sided bound.