CSE 840: Computational Foundations of Artificial Intelligence April 9, 2025

Law of Large Numbers: Central Limit Theorem, Concentration Inequalities

Instructor: Vishnu Boddeti *Scribe: Jihoon Choi, Dae Young Kim*

1 Limit Theorems: LLN and CLT

1.1 Strong Law of Large Numbers (SLLN)

Let $X_n : \{\Omega, \mathcal{A}, P\} \to \mathbb{R}$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean $\mu := E(X_n) < \infty$ and finite variance $\sigma^2 := \text{Var}(X_n) < \infty$. The Strong Law of Large Numbers states:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = \mu \quad \text{almost surely (a.s.) and in } L^2.$$

Explanation: The sample average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to the true mean μ with probability 1 (almost surely), meaning the set of outcomes where convergence fails has measure zero. Additionally, it converges in the mean-square sense (L^2), where $E\left[\left(\frac{1}{n}\sum X_i - \mu\right)^2\right] \to 0$. This dual convergence makes the SLLN a powerful tool, guaranteeing that with enough i.i.d. samples, the empirical average reliably estimates the population mean.

Assumptions and Variations: The SLLN relies on:

- *Independence:* Each X_i is independent of all others.
- *Identical Distribution:* All X_i share the same probability distribution (same probability mass or density).
- *Finite Moments:* Both the mean μ and variance σ^2 must exist and be finite.

Note: These assumptions are standard for the SLLN, but variations exist that relax them—e.g., the variance need not be finite, or independence can be weakened to pairwise independence—broadening the theorem's applicability. In machine learning, the i.i.d. assumption often holds (e.g., random data sampling), but may fail in cases like time-series data, necessitating such relaxations.

Examples:

- **Train/Test Error in Machine Learning:** The average training or test error, computed over many i.i.d. samples, converges to the true expected error as the dataset size increases, supporting error estimation in models. This underpins error estimation in models like neural networks, where large datasets ensure the empirical loss approximates the population loss. For example, in a neural network, the average loss over n i.i.d. training samples approximates the expected loss over the entire data distribution as $n \to \infty$.

- **Statistical Testing:** In statistics, we compare the means of two distributions (e.g., μ_1 vs. μ_2) to determine if they differ. The SLLN ensures that sample averages from each distribution stabilize to their true means, enabling reliable hypothesis testing.
- **Generative Modeling:** When training generative models (e.g., GANs), comparing sample means from generated vs. real data distributions leverages the SLLN to assess model quality – i.e., does the sample mean of the generated data converge to the true mean of the real distribution?

AI Relevance: The SLLN is foundational in machine learning because it justifies why empirical averages (e.g., loss over a training set) approximate theoretical expectations as data grows. However, when the i.i.d. assumption breaks—say, in sequential data like video frames—alternative formulations or techniques (e.g., batch normalization) are needed.

1.2 Weak Law of Large Numbers (WLLN)

The Weak Law states that the sample average $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to the mean μ *in probability*:

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Remarks:

- **Variations of the Theorem:** Several versions of the Law of Large Numbers (LLN) exist, relaxing the strict i.i.d. (independent and identically distributed) assumption:
  - *Pairwise Independence:* Full independence between samples isn't always necessary; pairwise independence (i.e., $P(X_i, X_j) = P(X_i)P(X_j)$ for $i \neq j$) can suffice.
  - *Non-Identical Distributions:* The random variables X_i need not follow the same distribution. Convergence to the mean can still hold if variances are bounded, though the limit may be a weighted average of individual means.
  - *Unbounded Variances:* Some versions allow unbounded variances, yet still guarantee convergence under weaker conditions (e.g., Lyapunov's condition).

  Explanation: These relaxations broaden the LLN's applicability, e.g., in real-world datasets where samples may come from slightly different sources.
- **Strong vs. Weak Convergence:**
  - The Strong Law of Large Numbers (SLLN) guarantees convergence almost surely (a.s.), a stronger condition where the set of outcomes failing convergence has probability 0.
  - The Weak Law of Large Numbers (WLLN) ensures convergence in probability, a weaker condition where $P\left(\left|\frac{1}{n}\sum X_i - \mu\right| > \epsilon\right) \to 0$ for any $\epsilon > 0$.

  Note: SLLN implies WLLN, but the converse isn't true. For example, a sequence may oscillate yet satisfy WLLN without achieving the pointwise stability of SLLN.
- **Limitations and Failure Cases:** The LLN doesn't always hold, and caution is required:

- **Heavy-Tailed Distributions:** For distributions like the Cauchy (where the mean doesn't exist due to infinite tails), the sample average $\frac{1}{n}\sum X_i$ doesn't converge.
  - *Example:* If $X_i \sim \text{Cauchy}(0,1)$, the average of n samples remains Cauchy-distributed, not stabilizing to a single value.
  - *Relevance to AI:* Real-world data (e.g., financial returns, object sizes) often exhibit heavy tails, challenging LLN assumptions in model training.
- **Selection Bias:** If samples X_i embed systematic bias (e.g., from human behavior), the LLN won't eliminate it, even with large n.
  - *Example:* In job resume datasets, historical gender bias (e.g., favoring males for certain roles) persists in the sample mean, regardless of sample size.
  - *AI Context:* Modern AI models (e.g., hiring algorithms, facial recognition) inherit biases from training data. Collecting more biased data amplifies rather than mitigates these issues.
  - *Explanation:* The LLN assumes samples represent the true population. If the sampling process is skewed (e.g., by human economic or rational behavior), the "true mean" reflects the biased subset, not the intended population.

Caveat: Simply increasing sample size doesn't magically fix bias. In AI development, addressing selection bias requires careful data curation, not just volume.

- **Practical Implications for AI:** Understanding these limitations is critical:
  - Biases in gender, ethnicity, or age in datasets (e.g., skewed performance in predictive models) persist unless sampling corrects the underlying skew.
  - Heavy-tailed phenomena (e.g., rare but extreme events) require alternative tools like robust statistics or truncation methods.

Example: Consider X_i as the outcome of a fair coin toss, where $X_i = 1$ for heads and $X_i = 0$ for tails, with $E(X_i) = 0.5$. The sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ converges to 0.5 in probability as $n \to \infty$ (WLLN). However, for finite n, outliers such as all heads ($\frac{1}{n}\sum X_i = 1$) or all tails ($\frac{1}{n}\sum X_i = 0$) can occur, though their probability diminishes as n increases. This illustrates the WLLN's probabilistic nature versus the SLLN's almost sure stability.

1.3 Central Limit Theorem (CLT)

Let $(X_i)_{i\in\mathbb{N}}$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Define the sum $S_n := \sum_{i=1}^{n} X_i$ and normalize it as:

$$Y_n := \frac{S_n - n\mu}{\sqrt{n}\sigma}.$$

Then, Y_n converges in distribution to a standard normal random variable:

$$Y_n \xrightarrow{d} Y, \quad Y \sim N(0,1).$$

Explanation: While the Law of Large Numbers (LLN) ensures that the sample average $\frac{S_n}{n}$ converges to μ, the CLT goes further by quantifying the deviations of S_n from $n\mu$ and characterizing their distribution. Specifically, regardless of the underlying distribution of X_i (e.g., binomial, uniform, or Bernoulli), as long as $\sigma^2 < \infty$, the normalized sum Y_n approaches a standard normal $N(0,1)$

as n increases. The scaling by $\sqrt{n}\sigma$ is critical: dividing by n (as in LLN) collapses the deviation to zero, while a smaller divisor (e.g., $\sqrt[3]{n}$) causes it to diverge—only $\sqrt{n}$ reveals the Gaussian limit. This "magical" property justifies Gaussian approximations in statistics and beyond.

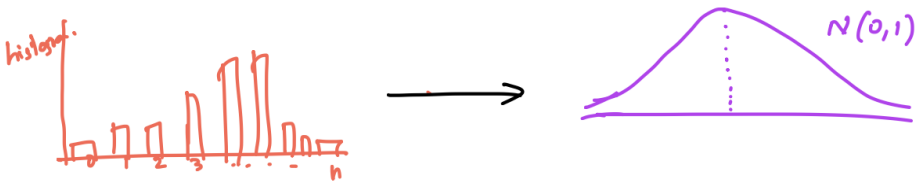

Figure 1: CLT Convergence and Distribution

Illustration: Consider X_i as a fair coin toss: heads $= 1$, tails $= 0$ ($\mu = 0.5$, $\sigma^2 = 0.25$). The sum $S_n = \sum_{i=1}^{n} X_i \in [0, n]$ represents the number of heads in n tosses. For small n, a histogram of S_n peaks around $0.5n$ and resembles a binomial distribution. As n grows, this histogram widens but, when normalized as:

$$Y_n = \frac{S_n - 0.5n}{\sqrt{n} \cdot 0.5} \xrightarrow{d} N(0, 1),$$

it tightens into a bell curve centered at 0 with unit variance.

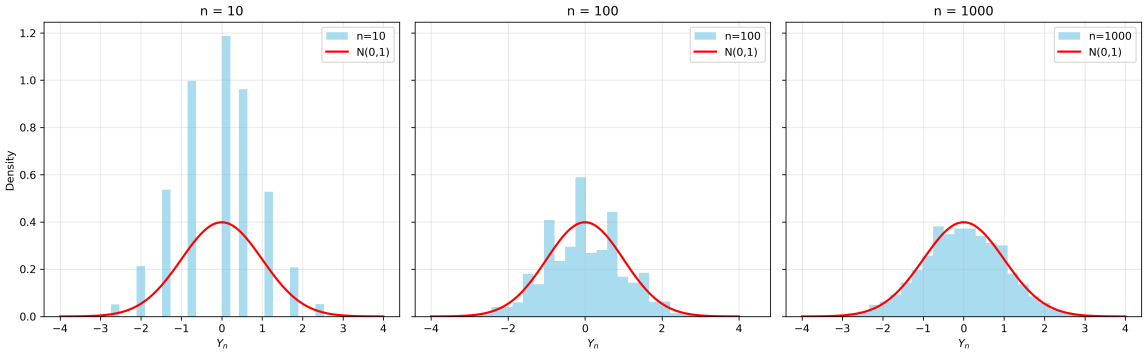

Figure 2: Distribution of Y_n for coin tosses with $n = 10, 50, 100$. As n increases, the histogram approaches the standard normal $N(0, 1)$ curve.

Insight: Figure 2 (simulated in Python) illustrates how Y_n's distribution converges to $N(0, 1)$ as n increases. For $n = 10$, the histogram is coarse; by $n = 100$, it closely matches the normal curve. This tightening reflects how larger sample sizes reduce variability relative to $\sqrt{n}$, a key insight for evaluating AI models where sample averages (e.g., loss) must approximate population behavior.

Practical Relevance: The CLT's power lies in its universality: it applies to sums of arbitrary i.i.d. variables, making it invaluable for real-world scenarios with unknown noise sources. In physics or observational studies, additive errors from many small, independent factors (e.g., measurement noise) often yield Gaussian deviations, even if their individual distributions are unknown. In AI, this underpins confidence intervals for model performance and explains why aggregated effects (e.g., weight updates in neural networks) often appear normal, enabling robust statistical analysis.

2 Concentration Inequalities

2.1 Motivation: Random Projections

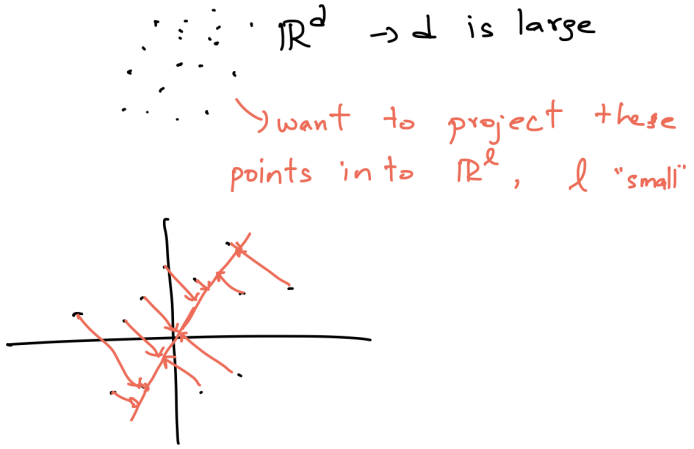

Figure 3: Random Projections

Concentration inequalities provide bounds on the probability that a random variable deviates significantly from its expected value, offering stronger tools than the Law of Large Numbers (LLN) and Central Limit Theorem (CLT). While LLN ensures convergence of the sample average to the true mean and CLT describes the distribution of deviations, concentration inequalities answer "how fast" this convergence occurs and "how many samples" are needed for the empirical average to be close to the mean. Unlike weaker bounds like Markov's or Chebyshev's inequalities (which require minimal assumptions), these stronger inequalities demand more conditions but yield tighter results. A key application is in random projections, used in machine learning for dimensionality reduction (e.g., compressing high-dimensional features into a manageable space).

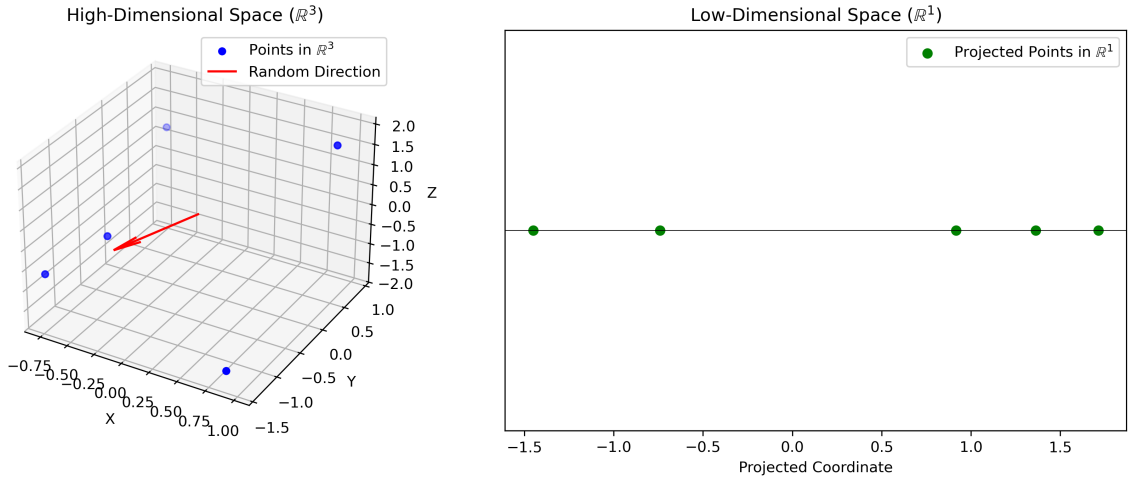

Figure 4: Random Projections from High to Low Dimensions

2.2 Johnson-Lindenstrauss Theorem

For vectors $x_i, x_j \in \mathbb{R}^d$ and a random projection $\pi : \mathbb{R}^d \to \mathbb{R}^l$ (where $l \ll d$), there exist parameters $\epsilon > 0$ and l such that:

$$(1-\epsilon)\|x_i - x_j\|_{\mathbb{R}^d} \leq \|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^l} \leq (1+\epsilon)\|x_i - x_j\|_{\mathbb{R}^d},$$

with high probability (e.g., $1-\delta$, where δ decreases as l increases).[1]

Construction Steps:

1. Assume $\|x_i - x_j\|_{\mathbb{R}^d} = 1$ to simplify analysis.
2. Compute the expected projected distance $E(\|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^l})$ over random projections.
3. Bound the deviation: $P\left(|\|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^l} - E(\|\pi(x_i) - \pi(x_j)\|_{\mathbb{R}^l})| > t\right)$, showing it's small for suitable l.

Explanation: This theorem ensures that distances between points in a high-dimensional space ($\mathbb{R}^d$) are approximately preserved when projected onto a lower-dimensional space ($\mathbb{R}^l$) using random directions. For instance, points in $\mathbb{R}^d$ can be projected onto l randomly chosen vectors (e.g., a single line for $l = 1$, or a subspace for $l > 1$), and their distances in $\mathbb{R}^l$ remain within a $(1-\epsilon, 1+\epsilon)$ factor of the originals. This property enables efficient computation in AI tasks like clustering or nearest neighbor search, where working directly in high dimensions is impractical.

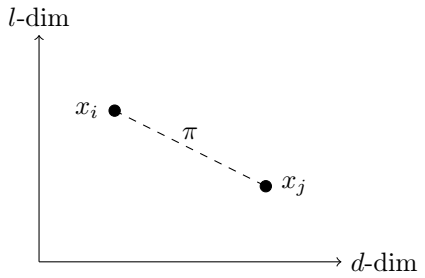

Figure 5: Random projection preserving distances from $\mathbb{R}^d$ to $\mathbb{R}^l$.

AI Example: In natural language processing, word embeddings (e.g., Word2Vec) in high dimensions can be projected to lower dimensions via random projections, retaining semantic similarity with minimal loss. Unlike principal component analysis (PCA), which seeks optimal projections, the JL theorem leverages random directions—uniformly chosen and computationally simpler—yet still preserves distances effectively, making it practical for large-scale ML tasks.

Additional Insight: The required dimension is $l \approx O(\epsilon^{-2} \log n)$, where n is the number of points. This logarithmic dependence on n and inverse quadratic on ϵ ensures efficiency, as l remains much smaller than d while controlling distortion.

[1] Here, $\|\cdot\|_{\mathbb{R}^d}$ and $\|\cdot\|_{\mathbb{R}^l}$ denote the Euclidean (L_2) norms in $\mathbb{R}^d$ and $\mathbb{R}^l$, respectively.

3 Hoeffding's Inequality

Theorem 1 (Hoeffding's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables, where each X_i takes values in the interval $[a_i, b_i]$ almost surely. Define*

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

Then for any $t > 0$, we have

$$\mathbb{P}\big(S_n > t\big) \;\le\; \exp\!\left(-\frac{2\,t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Remark 1 *Hoeffding's Inequality is a powerful exponential tail bound for sums of bounded independent random variables. It improves upon classical bounds such as Markov's or Chebyshev's inequality by giving exponentially decaying tail probabilities rather than polynomial decay.*

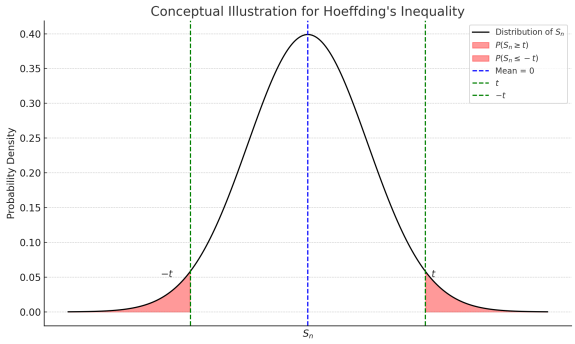

Figure 6: Conceptual Illustration for Hoeffding's Inequality

3.1 Application of Hoeffding's Inequality: Strong Law of Large Numbers (SLLN)

Setup

Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of i.i.d. random variables with common mean $\mathbb{E}[X_i] = \mu$ and each X_i taking values in $[a, b]$ almost surely (where $a < b$). We want to prove that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \;\xrightarrow{\text{a.s.}}\; \mu \quad \text{as } n \to \infty.$$

That is, the sample average converges to the true mean almost surely.

Proof Using Hoeffding's Inequality

Step 1: Apply Hoeffding's Inequality to the sum

Define

$$S_n \;=\; \sum_{i=1}^{n}(X_i - \mu).$$

Observe that each $X_i - \mu$ is bounded by $[\,a-\mu,\, b-\mu\,]$. Since all X_i are i.i.d. and bounded by $[a,b]$, the difference $(X_i - \mu)$ lies in an interval of length $(b-a)$. Thus we can set $(b_i - a_i) = (b-a)$ in Theorem 1.

Applying Hoeffding's Inequality, for any $t > 0$, we get

$$\mathbb{P}\Big(S_n > nt\Big) \;=\; \mathbb{P}\Big(\sum_{i=1}^{n}(X_i-\mu) > nt\Big) \;\le\; \exp\Big(-\tfrac{2(nt)^2}{n(b-a)^2}\Big) \;=\; \exp\Big(-\tfrac{2n\,t^2}{(b-a)^2}\Big).$$

Equivalently,

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu > t\Big) \;\le\; \exp\Big(-\tfrac{2n\,t^2}{(b-a)^2}\Big).$$

By symmetry, we also have

$$\mathbb{P}\Big(\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu < -t\Big) \;\le\; \exp\Big(-\tfrac{2n\,t^2}{(b-a)^2}\Big).$$

Combining these two bounds via the union bound gives

$$\mathbb{P}\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu\big| \ge t\Big) \;\le\; 2\,\exp\Big(-\tfrac{2n\,t^2}{(b-a)^2}\Big).$$

Step 2: Use the Borel–Cantelli Lemma for almost sure convergence

For a fixed $t > 0$, consider the infinite series

$$\sum_{n=1}^{\infty}\mathbb{P}\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu\big| \ge t\Big).$$

From the bound above,

$$\sum_{n=1}^{\infty}\mathbb{P}\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu\big| \ge t\Big) \;\le\; \sum_{n=1}^{\infty}2\,\exp\Big(-\tfrac{2n\,t^2}{(b-a)^2}\Big).$$

This is a sum of exponential terms in n, which converges (it is dominated by a geometric series). Since

$$\sum_{n=1}^{\infty}\mathbb{P}\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu\big| \ge t\Big) \;<\; \infty,$$

the Borel–Cantelli Lemma implies that

$$\mathbb{P}\Big(\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i - \mu\big| \ge t \text{ infinitely often}\Big) \;=\; 0.$$

Hence, for each $t > 0$, with probability 1, there exists some $N(\omega)$ (depending on the sample point ω) such that for all $n > N(\omega)$,

$$\big|\tfrac{1}{n}\sum_{i=1}^{n}X_i(\omega) - \mu\big| < t.$$

Since $t > 0$ was arbitrary, we conclude that

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu$$

Remark 2 *This approach shows that the probability of large deviations of the sample mean from its expected value decreases* exponentially *in n. Summing these probabilities over n yields a convergent series, which directly implies (by the Borel–Cantelli Lemma) that almost all sample paths can deviate from the mean by more than t only finitely many times. Therefore, the sample mean converges to the true mean almost surely.*

Remark 3 *For specific distributions (e.g. fair coin tosses, where $X_i \in \{0, 1\}$), Hoeffding's bound coincides with other well-known tail bounds (e.g. Chernoff bounds) and is often essentially tight for symmetric Bernoulli variables. For other distributions, more specialized bounds may give sharper results, but Hoeffding's inequality remains a very useful and general-purpose tool.*

4 Bernstein's Inequality

Theorem 2 (Bernstein's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables with zero mean, and assume $|X_i| \leq 1$ almost surely for each i. Define*

$$\sigma^2 \;=\; \sum_{i=1}^{n} \mathrm{Var}(X_i).$$

Then, for any $t > 0$, we have

$$\mathbb{P}\Big(\sum_{i=1}^{n} X_i \;\geq\; t\Big) \;\leq\; \exp\Big(-\frac{t^2}{2\big(\sigma^2 + \frac{t}{3}\big)}\Big).$$

By symmetry, the same bound holds for $\mathbb{P}\big(\sum_{i=1}^{n} X_i \leq -t\big)$.

Remark 4 *Bernstein's Inequality refines Hoeffding's Inequality by incorporating the variance term σ^2. When the variance is small, Bernstein's bound can be significantly tighter than Hoeffding's. It is often used when one knows both a bound on the individual random variables (i.e. $|X_i| \leq 1$) and their variance. This inequality plays an important role in scenarios where random variables have sub-exponential tails.*

5 Concentration Inequality for Functions with Bounded Differences

Many applications of concentration results involve not just sums of independent random variables but also functions of these variables. If the function does not change too much when a single input variable is changed (i.e. it has a "bounded difference"), then a Hoeffding-type tail bound still holds. The standard result is known as McDiarmid's Inequality.

5.1 Bounded Difference Property

Definition 1 (Bounded Difference Property) *Let $f : \mathcal{X}^n \to \mathbb{R}$ be a function of n variables, where $\mathcal{X}$ is any set. We say that f has the* bounded difference property *if there exist constants $c_1, c_2, \ldots, c_n \geq 0$ such that for all $i \in \{1, 2, \ldots, n\}$ and for any $(x_1, \ldots, x_n)$ and $(x_1, \ldots, x_i', \ldots, x_n)$ in $\mathcal{X}^n$, we have*

$$\big|\, f(x_1, \ldots, x_i, \ldots, x_n) \;-\; f(x_1, \ldots, x_i', \ldots, x_n)\big| \;\leq\; c_i.$$

Remark 5 *Intuitively, changing one coordinate of the input does not affect the function value by more than c_i. If each c_i is small, f is "stable" with respect to changes in its arguments.*

5.2 McDiarmid's Inequality

Theorem 3 (McDiarmid's Inequality) *Let $X_1, X_2, \ldots, X_n$ be independent random variables taking values in $\mathcal{X}$. Let $f : \mathcal{X}^n \to \mathbb{R}$ be a function that satisfies the bounded difference property with constants $c_1, \ldots, c_n$. Denote*

$$Y \;=\; f(X_1, X_2, \ldots, X_n) \quad \textit{and} \quad \mu \;=\; \mathbb{E}\big[Y\big].$$

Then, for any $t > 0$,

$$\mathbb{P}\big(Y - \mu \;\geq\; t\big) \;\leq\; \exp\Big(-\frac{2\,t^2}{\sum_{i=1}^{n} c_i^2}\Big).$$

Remark 6 *McDiarmid's Inequality can be viewed as a multidimensional analogue of Hoeffding's Inequality, where we measure how much each input coordinate can affect the function's value. This inequality is widely used in machine learning and combinatorial optimization to control the deviations of empirical processes, such as Rademacher averages, and to analyze algorithms.*

6 Additional Applications

The discussed concentration inequalities are fundamental in numerous areas. Below are several applications related to them.

- **Leave-One-Out Error Estimates**
  In machine learning, if removing a single data point leads to only a small change in error (i.e., the algorithm is stable), then McDiarmid's inequality guarantees that the leave-one-out error is tightly concentrated around its mean, reliably estimating the generalization error.
- **Stability in Machine Learning**
  Stability implies that slight changes in the training data cause only minor variations in the algorithm's output. Concentration inequalities like McDiarmid's ensure that the empirical error closely approximates the expected error with high probability.
- **Randomized Algorithms in Theoretical Computer Science**
  In randomized algorithms, concentration bounds (e.g., Hoeffding's and McDiarmid's inequalities) show that the performance is near its expected value with high probability. This is crucial for approximation algorithms in NP-hard problems such as the traveling salesman problem.

- **Largest Eigenvalue of Random Symmetric Matrices**
  Extensions of Bernstein's inequality to matrices (e.g., the Matrix Bernstein inequality) provide high-probability bounds for the largest eigenvalue of random symmetric matrices, with applications in spectral graph theory and high-dimensional statistics.