Lecture 4: Concentration Inequalities, Convexity, and Sampling

Undergraduate Mathematics Educator

April 21, 2025

Welcome back! Today, we'll delve into some powerful tools in probability theory: concentration inequalities, which tell us how likely random variables are to be close to their expected values, and the beautiful relationship between expectation and convex functions. We'll finish by discussing how we can actually generate random numbers following specific distributions, a crucial practical skill.

1 Concentration Inequalities (Section 1.4.1)

1.1 Motivation: How Concentrated is a Random Variable?

Often in probability and its applications, we're interested not just in the average value (expectation) of a random variable, but also in how likely it is to deviate *far* from that average. Think about the average height of students in a class versus the likelihood of finding someone exceptionally tall or short. Concentration inequalities provide mathematical bounds on the probability of such deviations.

In essence, they quantify how "concentrated" the distribution of a random variable is around a central value (like the mean or median). These inequalities are fundamental tools in:

- **Probability Theory:** Understanding the behavior of sums and averages of random variables.
- **Statistics:** Analyzing estimation errors and constructing confidence intervals.
- **Data Science & Machine Learning:** Analyzing probabilistic algorithms, bounding generalization errors in learning models.
- **Computer Science:** Analyzing randomized algorithms.

Some key examples of concentration inequalities include:

- **Markov's inequality:** A basic inequality for non-negative random variables, using only the expectation.
- **Chebyshev's inequality:** Uses both the expectation and variance to provide a tighter bound than Markov's in many cases.
- **Chernoff bounds:** Provide exponentially decreasing bounds for deviations of sums of independent (often Bernoulli) random variables from their mean. Typically much stronger than Markov or Chebyshev for sums.
- **Hoeffding's inequality:** Similar to Chernoff bounds but applies to sums of bounded independent random variables.
- **Azuma-Hoeffding inequality:** A powerful generalization of Hoeffding's inequality for martingales with bounded differences.

We'll start with the foundational ones: Markov's and Chebyshev's inequalities.

1.2 Markov's Inequality

This is often the first concentration inequality one encounters. It's remarkably simple, requiring only knowledge of the expectation and that the random variable is non-negative.

Theorem 1.1 (Markov's Inequality - Theorem 1.10)**.** *Let X be a random variable such that $X \geq 0$. Then for any constant $x > 0$,*

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

Proof. The proof relies on a clever use of indicator functions and the definition of expectation. Let X be a non-negative random variable and let $x > 0$. Consider the indicator function $I_{\{X \geq x\}}$, which is 1 if $X \geq x$ and 0 otherwise.

Since $X \geq 0$, we can write the expectation of X using the law of total expectation, partitioning on the event $\{X \geq x\}$ and its complement:

$$\mathbb{E}[X] = \mathbb{E}[X \cdot I_{\{X \geq x\}}] + \mathbb{E}[X \cdot I_{\{X < x\}}]$$

Because $X \geq 0$ and the indicator function $I_{\{X<x\}}$ is non-negative, the second term $\mathbb{E}[X \cdot I_{\{X<x\}}] \geq 0$. Therefore,

$$\mathbb{E}[X] \geq \mathbb{E}[X \cdot I_{\{X \geq x\}}]$$

Now, on the event $\{X \geq x\}$, the value of X is at least x. Thus, $X \cdot I_{\{X \geq x\}} \geq x \cdot I_{\{X \geq x\}}$. Taking expectations (which preserves inequalities):

$$\mathbb{E}[X \cdot I_{\{X \geq x\}}] \geq \mathbb{E}[x \cdot I_{\{X \geq x\}}]$$

Since x is a constant, we can pull it out of the expectation:

$$\mathbb{E}[x \cdot I_{\{X \geq x\}}] = x \cdot \mathbb{E}[I_{\{X \geq x\}}]$$

Recall that the expectation of an indicator function is the probability of the event it indicates: $\mathbb{E}[I_{\{X \geq x\}}] = \mathbb{P}(X \geq x)$.

Combining these steps, we have:

$$\mathbb{E}[X] \geq \mathbb{E}[X \cdot I_{\{X \geq x\}}] \geq \mathbb{E}[x \cdot I_{\{X \geq x\}}] = x \cdot \mathbb{P}(X \geq x)$$

So, $\mathbb{E}[X] \geq x \cdot \mathbb{P}(X \geq x)$. Since $x > 0$, we can rearrange to get the desired inequality:

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

□

Remark 1.2. Markov's inequality gives a bound on the "right tail" probability. It tells us that if the average value $\mathbb{E}[X]$ is small, then the probability of observing a very large value of X must also be small.

Corollary 1.3 (Polynomial Tail Decay from Moments)**.** *Suppose X is a random variable such that $\mathbb{E}[|X|^p] < \infty$ for some $p > 0$. Then the tail probability $\mathbb{P}(|X| \geq x)$ decays at least as fast as x^{-p} as $x \to \infty$.*

Proof. Let $Y = |X|^p$. Since $p > 0$, Y is a non-negative random variable. We can apply Markov's inequality to Y. For any $x > 0$: The event $\{|X| \geq x\}$ is equivalent to the event $\{|X|^p \geq x^p\}$ (since $z \mapsto z^p$ is increasing for $z \geq 0$ and $p > 0$). This is the event $\{Y \geq x^p\}$. Applying Markov's inequality to Y with the threshold $x^p > 0$:

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

Thus, $\mathbb{P}(|X| \geq x) \leq \frac{\mathbb{E}[|X|^p]}{x^p}$. Since $\mathbb{E}[|X|^p]$ is a finite constant, the probability decays at least like $1/x^p$. □

Remark 1.4. This shows that the existence of higher moments implies faster decay of the tail probabilities. However, Markov's inequality often provides quite loose bounds in practice. Furthermore, it provides no direct bound on the "left tail" probability $\mathbb{P}(X \leq x)$ (for $x < \mathbb{E}[X]$) using only $\mathbb{E}[X]$.

1.3 Chebyshev's Inequality

If we know not only the mean but also the variance of a random variable, we can obtain a generally tighter bound on deviations from the mean using Chebyshev's inequality. It bounds the probability of being far from the mean in *either* direction.

Theorem 1.5 (Chebyshev's Inequality)**.** *Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and finite variance $\sigma^2 = Var(X) = \mathbb{E}[(X-\mu)^2]$. Then for any constant $k > 0$,*

$$\mathbb{P}(|X-\mu| \geq k) \leq \frac{Var(X)}{k^2} = \frac{\sigma^2}{k^2}$$

Equivalently, letting $k = \epsilon\sigma$ for some $\epsilon > 0$ (assuming $\sigma > 0$),

$$\mathbb{P}(|X-\mu| \geq \epsilon\sigma) \leq \frac{1}{\epsilon^2}$$

Proof. The proof is a neat application of Markov's inequality. Let $Y = (X-\mu)^2$. Since Y is a squared value, it must be non-negative, i.e., $Y \geq 0$. The expectation of Y is, by definition of variance, $\mathbb{E}[Y] = \mathbb{E}[(X-\mu)^2] = \mathrm{Var}(X) = \sigma^2$.

Now, consider the event $|X-\mu| \geq k$. Since both sides are non-negative, this is equivalent to squaring both sides: $(X-\mu)^2 \geq k^2$. This is precisely the event $Y \geq k^2$. Since $Y \geq 0$ and $k^2 > 0$ (as $k > 0$), we can apply Markov's inequality (Theorem 1.10) to the random variable Y with the threshold value k^2:

$$\mathbb{P}(Y \geq k^2) \leq \frac{\mathbb{E}[Y]}{k^2}$$

Substituting back $Y = (X-\mu)^2$ and $\mathbb{E}[Y] = \sigma^2$, we get:

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

Since the event $(X-\mu)^2 \geq k^2$ is identical to the event $|X-\mu| \geq k$, we have arrived at the desired result:

$$\mathbb{P}(|X-\mu| \geq k) \leq \frac{\sigma^2}{k^2}$$

□

Remark 1.6. Like Markov's inequality, Chebyshev's inequality is powerful because it holds for *any* distribution with finite mean and variance. It doesn't require knowledge of the specific probability density or mass function, only the existence and values of the first two moments. The price for this generality is that the bound might still be loose compared to what one might get if the distribution were known (e.g., for a Normal distribution).

Corollary 1.7 (Concentration of the Sample Mean)**.** *Let $X_1, X_2, \ldots, X_n$ be random variables that are identically distributed (with mean μ and variance σ^2) and* ***uncorrelated*** *(i.e., $Cov(X_i, X_j) = 0$ for $i \neq j$). Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then for any $\epsilon > 0$,*

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$$

Proof. First, let's find the mean and variance of the sample mean $\bar{X}_n$. Using linearity of expectation:

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu$$

So, the sample mean is an unbiased estimator of the population mean. Now, let's find the variance. Since the variables are uncorrelated, the variance of the sum is the sum of the variances:

$$\begin{aligned}
\mathrm{Var}(\bar{X}_n) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \quad \text{(Property of variance)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \quad \text{(Uncorrelated variables)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n}
\end{aligned}$$

The variance of the sample mean decreases as n increases, which is intuitive: averaging more variables should lead to less spread.

Now we apply Chebyshev's inequality to the random variable $\bar{X}_n$. It has mean μ and variance σ^2/n. Using the threshold $\epsilon > 0$:

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{(\sigma^2/n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

This proves the result. □

Remark 1.8. This corollary provides a proof of the **Weak Law of Large Numbers** under the condition of uncorrelated variables with finite variance. It shows that as the sample size n increases, the probability that the sample mean $\bar{X}_n$ deviates from the true mean μ by more than any fixed amount ϵ converges to zero. The rate of convergence guaranteed by Chebyshev is at least $1/n$. The distribution of $\bar{X}_n$ becomes increasingly concentrated around μ.

2 Expectation and Convexity (Section 1.4.2)

We now shift gears to explore a fascinating interaction between the expectation operator and a geometric property of functions known as convexity. This relationship leads to a fundamental inequality with wide-ranging applications.

Definition 2.1 (Convex and Concave Functions - Definition 1.12)**.** Let $A \subseteq \mathbb{R}$ be an interval. A function $g : A \to \mathbb{R}$ is called **convex** on A if for all $x, y \in A$ and all $\lambda \in [0, 1]$,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$$

Geometrically, this means the line segment connecting any two points $(x, g(x))$ and $(y, g(y))$ on the graph of g lies on or above the graph itself.

A function g is called **concave** on A if the inequality is reversed:

$$g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y)$$

Equivalently, g is concave if and only if $-g$ is convex. Thus, results for convex functions can be readily adapted to concave functions by flipping the inequality sign.

Remark 2.2. Convexity is a fundamental concept appearing in optimization (convex functions have unique minima under certain conditions), inequalities, functional analysis, and many other areas. Examples of convex functions include $g(x) = x^2$, $g(x) = e^x$, $g(x) = |x|$. Examples of concave functions include $g(x) = \sqrt{x}$ (for $x \geq 0$) and $g(x) = \ln(x)$ (for $x > 0$). A linear function $g(x) = ax + b$ is both convex and concave.

The following lemma provides an alternative characterization of convexity, involving tangent or supporting lines, which is often useful and provides geometric insight.

Lemma 2.3 (Supporting Line Characterization - Lemma 1.13)**.** *A function $g : A \to \mathbb{R}$ defined on an interval A is convex if and only if for every $x_0 \in A$, there exists a value $v(x_0)$ (representing the slope of a supporting line at x_0) such that for all $y \in A$,*

$$g(y) \geq g(x_0) + v(x_0)(y - x_0)$$

Geometrically, this means that the graph of g lies entirely on or above its supporting (or tangent, if differentiable) line at any point x_0.

Proof. ($\Longleftarrow$) Suppose such a function $v(x)$ exists. Let $x_1, x_2 \in A$ and $\lambda \in [0, 1]$. Let $x_0 = \lambda x_1 + (1 - \lambda)x_2$. Since A is an interval, $x_0 \in A$. Applying the assumed inequality at the point x_0, for $y = x_1$ and $y = x_2$:

$$\begin{aligned} g(x_1) &\geq g(x_0) + v(x_0)(x_1 - x_0) \\ g(x_2) &\geq g(x_0) + v(x_0)(x_2 - x_0) \end{aligned}$$

Multiply the first inequality by λ (≥ 0) and the second by $(1 - \lambda)$ (≥ 0) and add them:

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \geq \lambda g(x_0) + (1 - \lambda)g(x_0) + v(x_0)[\lambda(x_1 - x_0) + (1 - \lambda)(x_2 - x_0)]$$

Simplifying the right side:

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \geq g(x_0) + v(x_0)[\lambda x_1 + (1 - \lambda)x_2 - (\lambda + 1 - \lambda)x_0]$$

Since $x_0 = \lambda x_1 + (1 - \lambda)x_2$, the term in the square brackets is $x_0 - x_0 = 0$. Thus,

$$\lambda g(x_1) + (1 - \lambda)g(x_2) \geq g(x_0) = g(\lambda x_1 + (1 - \lambda)x_2)$$

This is precisely the definition of convexity (with the inequality reversed from the definition, matching the required inequality).

($\Longrightarrow$) Suppose g is convex. For a fixed $x_0 \in A$, the slopes of the chords connecting $(x_0, g(x_0))$ to $(y, g(y))$ are non-decreasing as y moves away from x_0. If g is differentiable at x_0, the tangent line $L(y) = g(x_0) + g'(x_0)(y - x_0)$ serves as the supporting line, so we can choose $v(x_0) = g'(x_0)$. The inequality $g(y) \geq g(x_0) + g'(x_0)(y - x_0)$ is a standard property derived from the definition of convexity for differentiable functions. Even if g is not differentiable everywhere, a convex function is continuous on the interior of A and possesses left and right derivatives, $g'_-(x_0)$ and $g'_+(x_0)$, at any interior point x_0. Any value $v(x_0)$ such that $g'_-(x_0) \leq v(x_0) \leq g'_+(x_0)$ will satisfy the supporting line inequality. For instance, one can choose $v(x_0) = g'_+(x_0)$. The inequality can be derived more formally from the definition of convexity. $\square$

Example 2.4 (Checking Convexity via Second Derivative - Example 1.14)**.** If a function g is twice differentiable on an interval A, there's a very convenient way to check for convexity. Using Taylor's theorem, or as derived by integrating g'', we find:

$$g(y) = g(x) + g'(x)(y - x) + \int_x^y \int_x^u g''(v)dvdu$$

Rearranging gives:

$$g(y) - g(x) - g'(x)(y - x) = \int_x^y \left(\int_x^u g''(v)dv \right) du$$

From Lemma 1.13 (with $x_0 = x$), g is convex if and only if the left-hand side is non-negative for all $x, y \in A$. This means the double integral on the right must be non-negative. This condition holds if and only if the innermost integrand $g''(v)$ is non-negative for all v in the interval A. Therefore, for a twice-differentiable function g:

$$g \text{ is convex on } A \iff g''(x) \geq 0 \text{ for all } x \in A$$

Similarly, g is concave if and only if $g''(x) \leq 0$ for all $x \in A$. This provides a practical test.

Examples:

- $g(x) = e^x$. Then $g'(x) = e^x$ and $g''(x) = e^x$. Since $e^x > 0$ for all $x \in \mathbb{R}$, $g(x) = e^x$ is strictly convex on $\mathbb{R}$.
- $g(x) = x^p$ for $x > 0$. Then $g'(x) = px^{p-1}$ and $g''(x) = p(p-1)x^{p-2}$. For $g''(x) \geq 0$ (given $x > 0$), we need $p(p-1) \geq 0$. This occurs when $p \leq 0$ or $p \geq 1$. The function $g(x) = x^p$ is convex on $(0, \infty)$ for $p \in (-\infty, 0] \cup [1, \infty)$. (The specific case $p \geq 2$ mentioned in the source notes is indeed convex). It is concave for $p \in [0, 1]$.

Now we arrive at the central result connecting convexity and expectation.

Theorem 2.5 (Jensen's Inequality - Theorem 1.15)**.** *Let X be a random variable such that its expectation $\mu = \mathbb{E}[X]$ is finite and lies within an interval A. Let $g : A \to \mathbb{R}$ be a* ***convex*** *function. If $\mathbb{E}[g(X)]$ also exists and is finite, then*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

If g is ***concave*** *on A, the inequality is reversed:*

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

Proof. We use the supporting line characterization from Lemma 1.13. Since g is convex, let $x_0 = \mathbb{E}[X]$. As $x_0 \in A$, there exists a value $v(x_0)$ such that for any value y in the range of X (assuming the range is contained within A),

$$g(y) \geq g(x_0) + v(x_0)(y - x_0)$$

Let's substitute the random variable X for y and the constant $\mathbb{E}[X]$ for x_0. This inequality holds for each possible outcome of X:

$$g(X) \geq g(\mathbb{E}[X]) + v(\mathbb{E}[X])(X - \mathbb{E}[X])$$

Now, we take the expectation of both sides. Expectation is a linear operator and preserves inequalities (given that the expectations exist):

$$\mathbb{E}[g(X)] \geq \mathbb{E}[g(\mathbb{E}[X]) + v(\mathbb{E}[X])(X - \mathbb{E}[X])]$$

Using linearity of expectation on the right side:

$$\mathbb{E}[g(X)] \geq \mathbb{E}[g(\mathbb{E}[X])] + \mathbb{E}[v(\mathbb{E}[X])(X - \mathbb{E}[X])]$$

Since $g(\mathbb{E}[X])$ and $v(\mathbb{E}[X])$ are constants (because $\mathbb{E}[X]$ is a constant value), we can pull them out of the expectation:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) + v(\mathbb{E}[X])\, \mathbb{E}[X - \mathbb{E}[X]]$$

The last expectation term is $\mathbb{E}[X - \mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X] - \mathbb{E}[X] = 0$. So, the inequality simplifies to:

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X]) + v(\mathbb{E}[X]) \cdot 0$$

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[X])$$

This proves Jensen's inequality for convex functions. The proof for concave functions follows by applying the result to the convex function $-g$, which reverses the inequality. □

Remark 2.6. Jensen's inequality essentially states that for a convex function, the expectation of the function's value is greater than or equal to the function applied to the expectation. For a concave function, it's the other way around. This has important interpretations, for instance, in utility theory (risk aversion corresponds to concave utility functions) and information theory.

Jensen's inequality is a powerful tool for proving other inequalities. One important consequence is Lyapunov's inequality, which relates the different moments of a random variable.

Theorem 2.7 (Lyapunov's Inequality - Theorem 1.16)**.** *Let X be a random variable. For any real numbers $p \geq q > 0$, if the p-th absolute moment $\mathbb{E}[|X|^p]$ is finite, then the q-th absolute moment $\mathbb{E}[|X|^q]$ is also finite, and*

$$(\mathbb{E}[|X|^q])^{1/q} \leq (\mathbb{E}[|X|^p])^{1/p}$$

This means the function $f(r) = (\mathbb{E}[|X|^r])^{1/r}$, which represents the L_r-norm of X, is a non-decreasing function of r for $r > 0$.

Proof. Let $p \geq q > 0$. Define $r = p/q$. Since $p \geq q$, we have $r \geq 1$. Consider the function $g(z) = z^r = z^{p/q}$ for $z \geq 0$. Let's check its second derivative: $g'(z) = rz^{r-1}$, $g''(z) = r(r-1)z^{r-2}$. Since $r \geq 1$, we have $r - 1 \geq 0$. Also, $r > 0$ and $z^{r-2} \geq 0$ for $z \geq 0$ (interpret $0^0 = 1$ if $r = 1$). Thus, $g''(z) \geq 0$, which means $g(z) = z^r$ is a **convex** function for $r \geq 1$.

Let $Y = |X|^q$. Note that Y is a non-negative random variable. Apply Jensen's inequality (Theorem 1.15) to the random variable Y and the convex function $g(z) = z^r$:

$$\mathbb{E}[g(Y)] \geq g(\mathbb{E}[Y])$$

provided the expectations exist. Substituting $Y = |X|^q$ and $g(z) = z^r$:

$$\mathbb{E}[(|X|^q)^r] \geq (\mathbb{E}[|X|^q])^r$$

Simplifying the left side: $(|X|^q)^r = (|X|^q)^{p/q} = |X|^p$. So,

$$\mathbb{E}[|X|^p] \geq (\mathbb{E}[|X|^q])^{p/q}$$

Since $\mathbb{E}[|X|^p]$ is finite by assumption, the term $(\mathbb{E}[|X|^q])^{p/q}$ must also be finite. Since $p/q \geq 1$, this implies $\mathbb{E}[|X|^q]$ must be finite (and non-negative). Now, since both sides are non-negative and $p > 0$, we can take the positive $1/p$ power of both sides. The function $x \mapsto x^{1/p}$ is increasing for $x \geq 0$, so the inequality direction is preserved:

$$(\mathbb{E}[|X|^p])^{1/p} \geq ((\mathbb{E}[|X|^q])^{p/q})^{1/p}$$

Simplifying the right side exponent: $(p/q) \times (1/p) = 1/q$.

$$(\mathbb{E}[|X|^p])^{1/p} \geq (\mathbb{E}[|X|^q])^{1/q}$$

This is Lyapunov's inequality. □

Corollary 2.8 (Finiteness of Lower Moments - Corollary 1.17)**.** *Let X be a random variable. If $\mathbb{E}[|X|^p] < \infty$ for some $p > 0$, then $\mathbb{E}[|X|^q] < \infty$ for all $0 < q < p$.*

Proof. This follows directly from the proof of Lyapunov's inequality (Theorem 1.16). In the proof, we showed that if $\mathbb{E}[|X|^p] < \infty$, then $\mathbb{E}[|X|^q]$ must also be finite for $0 < q \leq p$. The corollary statement just restricts q to be strictly less than p. Alternatively, applying the main result of Lyapunov's inequality: $(\mathbb{E}[|X|^q])^{1/q} \leq (\mathbb{E}[|X|^p])^{1/p}$. Since the right side is a finite number (as $\mathbb{E}[|X|^p] < \infty$), the left side must also be finite. Since $q > 0$, this implies $\mathbb{E}[|X|^q]$ must be finite. □

Administrative Notes:

- **Applications of Jensen's Inequality:** Due to time constraints in lecture, we couldn't explore the many practical applications of Jensen's inequality. It plays a significant role in various fields including:
  - Economics (e.g., modelling risk aversion using concave utility functions).
  - Information Theory (e.g., proving properties of entropy and relative entropy/KL divergence).
  - Optimization (e.g., in proofs related to convex optimization problems).
  - Statistical Physics.
- **Further Exploration (Perplexity AI Link):** For those interested in seeing more examples and finding references, I previously used the AI model Perplexity to gather some information. You can explore its findings and follow related queries starting from this link: Uses of Jensen Inequality (Perplexity Search) The linked search provides an initial summary and can be a useful starting point for further investigation.

Exercise (Not covered in lecture)

Prove that for a positive random variable X (i.e., $X > 0$ with probability 1), if $\mathbb{E}[X]$ is finite and positive, then

$$\mathbb{E}[\ln(X)] \leq \ln(\mathbb{E}[X])$$

(This inequality relates the expected logarithm to the logarithm of the expectation, sometimes connected to comparing geometric and arithmetic means, or ideas of logarithmic utility in economics).

Solution: The function $g(x) = \ln(x)$ is defined for $x > 0$. We need to determine its convexity/concavity. Let's examine its second derivative: $g'(x) = 1/x$ $g''(x) = -1/x^2$ Since X is positive, its possible values x are greater than 0. For $x > 0$, $g''(x) = -1/x^2 < 0$. Therefore, $g(x) = \ln(x)$ is a strictly **concave** function on the interval $(0, \infty)$.

We are given that $X > 0$ and $\mathbb{E}[X]$ exists, is finite, and positive. Thus, $\mathbb{E}[X]$ lies in the domain $(0, \infty)$ where g is defined and concave. We also need $\mathbb{E}[\ln(X)]$ to exist (it could potentially be $-\infty$). Assuming $\mathbb{E}[\ln(X)]$ exists, we can apply Jensen's inequality for concave functions (Theorem 1.15, with the inequality reversed):

$$\mathbb{E}[g(X)] \leq g(\mathbb{E}[X])$$

Substituting $g(x) = \ln(x)$:

$$\mathbb{E}[\ln(X)] \leq \ln(\mathbb{E}[X])$$

This confirms the desired inequality. ———

3 Sampling from a Random Variable (Section 1.5)

So far, we've discussed the abstract properties of random variables and their distributions. A crucial practical question arises: how can we actually *generate* numerical values (realizations or samples) that follow a specific probability distribution? This process is fundamental for:

- **Simulation:** Modeling random phenomena in science, engineering, finance, etc.
- **Monte Carlo Methods:** Approximating integrals or expectations numerically.
- **Statistical Inference:** Generating data for bootstrapping or permutation tests.
- **Algorithm Testing:** Providing random inputs to evaluate algorithms.

Modern scientific computing environments like R, Python (with libraries like NumPy/SciPy), MATLAB, etc., provide convenient built-in functions to sample from many common distributions.

Example 3.1 (Sampling Functions in R)**.** Here are examples of R commands for generating n random samples from various distributions, corresponding to those listed in the original notes:

- Standard Uniform distribution $U(0,1)$: 'runif(n, min = 0, max = 1)' (or simply 'runif(n)')
- Standard Normal distribution $N(0,1)$: 'rnorm(n, mean = 0, sd = 1)' (or simply 'rnorm(n)')
- Exponential distribution with rate $\lambda = 1$: 'rexp(n, rate = 1)' (or simply 'rexp(n)')
- Geometric distribution with success probability p: 'rgeom(n, prob = p)' (Note: R's 'rgeom' counts the number of *failures* before the first success, starting from 0).

Many other functions like 'rpois', 'rbinom', 'rgamma', 'rbeta', etc., are available.

But what if we need to sample from a distribution that isn't built-in, or if we want to understand the fundamental principle behind how these generators might work? A widely applicable and elegant technique is the **Inverse Transform Method**. It leverages our ability to generate samples from the standard uniform distribution $U(0,1)$ – the foundation upon which many other random number generations are built.

The core idea is that if we can calculate the inverse of the Cumulative Distribution Function (CDF), F^{-1}, then we can transform a uniform random variable $V \sim U(0,1)$ into a random variable $X = F^{-1}(V)$ that follows the distribution F.

3.1 Inverse Transform Method: Discrete Case

Let's first see how this works for sampling from a discrete distribution.

Lemma 3.2 (Inverse Transform Sampling (Discrete) - Based on Lemma 1.8)**.** *Let X be a discrete random variable with possible values $x_1 < x_2 < x_3 < \dots$ and CDF $F(x) = \mathbb{P}(X \leq x)$. Let $p_j = \mathbb{P}(X = x_j) = F(x_j) - F(x_{j-1})$ (where $F(x_0) = 0$). Generate a random number V from the $U(0,1)$ distribution. Define a new random variable Y as follows:*

$$Y = x_j \quad \text{if} \quad F(x_{j-1}) < V \leq F(x_j)$$

This is equivalent to finding the smallest j such that $F(x_j) \geq V$. That is, $Y = \min\{x_j : F(x_j) \geq V\}$. Then the random variable Y has the same distribution as X.

Proof. We need to show that $\mathbb{P}(Y = x_j)$ is equal to the probability mass $p_j = \mathbb{P}(X = x_j)$ for every possible value x_j. By the definition of Y, the event $\{Y = x_j\}$ occurs if and only if the generated uniform random number V falls into the interval $(F(x_{j-1}), F(x_j)]$. Since $V \sim U(0,1)$, the probability of V falling into any interval $(a, b] \subseteq [0,1]$ is simply the length of the interval, $b - a$. Therefore,

$$\mathbb{P}(Y = x_j) = \mathbb{P}(F(x_{j-1}) < V \leq F(x_j))$$

The length of this interval is $F(x_j) - F(x_{j-1})$. By the definition of a discrete CDF, this difference is exactly the probability mass $p_j = \mathbb{P}(X = x_j)$. Since $\mathbb{P}(Y = x_j) = \mathbb{P}(X = x_j)$ for all possible values x_j, the random variable Y generated by this procedure has the same distribution as X. □

Remark 3.3. To implement this, one typically generates $V \sim U(0,1)$ and then checks inequalities sequentially: Is $V \leq F(x_1)$? If yes, $Y = x_1$. If no, is $V \leq F(x_2)$? If yes, $Y = x_2$, and so on.

3.2 Inverse Transform Method: Continuous Case

The same principle applies beautifully to continuous distributions, especially those with strictly increasing CDFs where the inverse F^{-1} is uniquely defined.

Lemma 3.4 (Inverse Transform Sampling (Continuous) - Lemma 1.9)**.** *Let F be a continuous and strictly increasing CDF defined on some interval (the support of the random variable). Let $F^{-1}(v)$ denote the inverse function of F, which maps $(0,1)$ back to the support interval.*

1. *(**Sampling**) If $V \sim U(0,1)$, then the random variable $X = F^{-1}(V)$ has CDF F.*
2. *(**Probability Integral Transform**) Conversely, if X is a random variable with continuous and strictly increasing CDF G, then the random variable $V = G(X)$ follows the $U(0,1)$ distribution.*

Proof.

1. We want to find the CDF of $X = F^{-1}(V)$. Let $F_X(x) = \mathbb{P}(X \leq x)$.

$$F_X(x) = \mathbb{P}(F^{-1}(V) \leq x)$$

Since F is strictly increasing, applying F to both sides of the inequality $F^{-1}(V) \leq x$ preserves the inequality direction:

$$\mathbb{P}(F^{-1}(V) \leq x) = \mathbb{P}(F(F^{-1}(V)) \leq F(x)) = \mathbb{P}(V \leq F(x))$$

Now, since $V \sim U(0,1)$, its CDF is $F_V(v) = v$ for $v \in [0,1]$. Since $F(x)$ maps the support of X into the interval $(0,1)$ (because F is continuous and strictly increasing), the value $F(x)$ lies in the range where the CDF of V is defined. Therefore,

$$\mathbb{P}(V \leq F(x)) = F_V(F(x)) = F(x)$$

Thus, $F_X(x) = F(x)$. This shows that the random variable $X = F^{-1}(V)$ has the desired CDF F.

2. We want to find the CDF of $V = G(X)$. Let $F_V(v) = \mathbb{P}(V \leq v)$ for $v \in (0,1)$. (The range of $G(X)$ will be $(0,1)$ since G is a continuous strictly increasing CDF).

$$F_V(v) = \mathbb{P}(G(X) \leq v)$$

Since G is strictly increasing, it has a unique inverse G^{-1}. Applying G^{-1} to both sides preserves the inequality:

$$\mathbb{P}(G(X) \leq v) = \mathbb{P}(G^{-1}(G(X)) \leq G^{-1}(v)) = \mathbb{P}(X \leq G^{-1}(v))$$

By definition, $\mathbb{P}(X \leq x)$ is given by the CDF of X, which is $G(x)$. So, applying this with $x = G^{-1}(v)$:

$$\mathbb{P}(X \leq G^{-1}(v)) = G(G^{-1}(v)) = v$$

Therefore, $F_V(v) = v$ for $v \in (0, 1)$. This is exactly the CDF of the $U(0, 1)$ distribution.

□

Remark 3.5. Part 2, the Probability Integral Transform, is a fascinating result stating that applying its own CDF transformation to *any* continuous random variable yields a standard uniform variable. This has important applications in statistics, for instance, in constructing goodness-of-fit tests or transforming data to uniformity. The original notes mentioned that a version of this lemma was related to a previous homework problem (Exercise 1, Question 3).

Example 3.6 (Sampling from an Exponential Distribution - Example 1.20)**.** Let's apply the inverse transform method (Lemma 1.19, part 1) to generate a sample from an Exponential distribution with rate parameter $\lambda > 0$. The specific case mentioned in the original lecture was $\lambda = 3$. The CDF of the $Exp(\lambda)$ distribution is:

$$F(x) = 1 - e^{-\lambda x}, \quad \text{for } x \geq 0$$

This CDF is continuous and strictly increasing on $[0, \infty)$. Its range is $[0, 1)$.

To apply the method, we need to find the inverse function $F^{-1}(v)$. We set $v = F(x)$ for $v \in (0, 1)$ and solve for x:

$$\begin{aligned} v &= 1 - e^{-\lambda x} \\ e^{-\lambda x} &= 1 - v \\ -\lambda x &= \ln(1 - v) \quad \text{(taking natural log of both sides)} \\ x &= -\frac{1}{\lambda} \ln(1 - v) \end{aligned}$$

So, the inverse CDF is $F^{-1}(v) = -\frac{1}{\lambda} \ln(1 - v)$.

According to the lemma, if we generate a standard uniform random variable $V \sim U(0, 1)$, then the transformed random variable

$$X = F^{-1}(V) = -\frac{1}{\lambda} \ln(1 - V)$$

will have an $Exp(\lambda)$ distribution.

For the specific case $\lambda = 3$ from the lecture:

$$X = -\frac{1}{3} \ln(1 - V)$$

where $V \sim U(0, 1)$.

Computational Note: If $V \sim U(0, 1)$, then the random variable $V' = 1 - V$ also follows a $U(0, 1)$ distribution (its CDF is $\mathbb{P}(1 - V \leq v) = \mathbb{P}(V \geq 1 - v) = 1 - (1 - v) = v$). Therefore, we can equivalently use the computationally slightly simpler formula:

$$X = -\frac{1}{\lambda} \ln(V')$$

where V' is a *different* $U(0, 1)$ sample. This form avoids the subtraction $1 - V$ and is often the one implemented in software libraries.

The inverse transform method is a powerful and general technique, providing a constructive way to sample from any distribution for which we can compute the inverse CDF, F^{-1}. This might be done analytically (as in the exponential example) or numerically if an analytical inverse is unavailable.