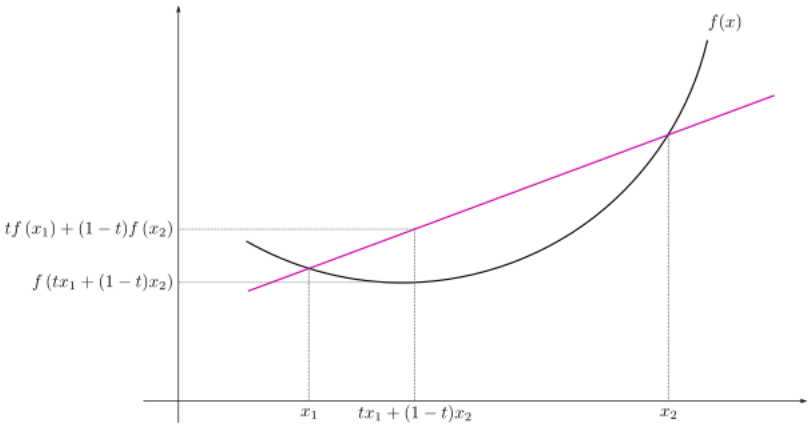
Jensen's inequality

In mathematics, **Jensen's inequality**, named after the Danish mathematician Johan Jensen, relates the value of a convex function of an integral to the integral of the convex function. It was proven by Jensen in 1906.[1] Given its generality, the inequality appears in many forms depending on the context, some of which are presented below. In its simplest form the inequality states that the convex transformation of a mean is less than or equal to the mean applied after convex transformation; it is a simple corollary that the opposite is true of concave transformations.



Jensen's inequality generalizes the statement that a secant line of a convex function lies above the graph.

Jensen's inequality generalizes the statement that the secant line of a convex function lies *above* the graph of the function, which is Jensen's inequality for two points: the secant line consists of weighted means of the convex function (for $t \in [0,1]$),

$$tf(x_1) + (1-t)f(x_2),$$

while the graph of the function is the convex function of the weighted means,

$$f\left(tx_1 + (1-t)x_2\right).$$

Thus, Jensen's inequality is

$$f\left(tx_1 + (1-t)x_2\right) \leq tf(x_1) + (1-t)f(x_2).$$

In the context of probability theory, it is generally stated in the following form: if X is a random variable and φ is a convex function, then

$$\varphi\left(\mathbf{E}[X]\right) \leq \mathbf{E}[\varphi(X)].$$

Contents

Statements

The classical form of Jensen's inequality involves several numbers and weights. The inequality can be stated quite generally using either the language of measure theory or (equivalently) probability. In the probabilistic setting, the inequality can be further generalized to its *full strength*.

Finite form

For a real convex function φ, numbers $x_1, x_2, \ldots, x_n$ in its domain, and positive weights a_i, Jensen's inequality can be stated as:

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \leq \frac{\sum a_i \varphi(x_i)}{\sum a_i} \qquad (1)$$

and the inequality is reversed if φ is concave, which is

$$\varphi\left(\frac{\sum a_i x_i}{\sum a_i}\right) \geq \frac{\sum a_i \varphi(x_i)}{\sum a_i}. \qquad (2)$$

Equality holds if and only if $x_1 = x_2 = \cdots = x_n$ or φ is linear.

As a particular case, if the weights a_i are all equal, then (1) and (2) become

$$\varphi\left(\frac{\sum x_i}{n}\right) \leq \frac{\sum \varphi(x_i)}{n} \qquad (3)$$

$$\varphi\left(\frac{\sum x_i}{n}\right) \geq \frac{\sum \varphi(x_i)}{n} \qquad (4)$$

For instance, the function $\log(x)$ is *concave*, so substituting $\varphi(x) = \log(x)$ in the previous formula (4) establishes the (logarithm of the) familiar arithmetic mean-geometric mean inequality:

$$\log\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) \geq \frac{\sum_{i=1}^{n} \log(x_i)}{n} \quad \text{or} \quad \frac{x_1 + x_2 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \cdot x_2 \cdots x_n}$$

A common application has x as a function of another variable (or set of variables) t, that is, $x_i = g(t_i)$. All of this carries directly over to the general continuous case: the weights a_i are replaced by a non-negative integrable function $f(x)$, such as a probability distribution, and the summations are replaced by integrals.

Measure-theoretic and probabilistic form

Let (Ω, A, μ) be a probability space, such that $\mu(\Omega) = 1$. If g is a real-valued function that is μ-integrable, and if φ is a convex function on the real line, then:

$$\varphi\left(\int_\Omega g\,d\mu\right) \leq \int_\Omega \varphi\circ g\,d\mu.$$

In real analysis, we may require an estimate on

$$\varphi\left(\int_a^b f(x)\,dx\right),$$

where $a, b \in \mathbb{R}$, and $f : [a, b] \to \mathbb{R}$ is a non-negative Lebesgue-integrable function. In this case, the Lebesgue measure of $[a, b]$ need not be unity. However, by integration by substitution, the interval can be rescaled so that it has measure unity. Then Jensen's inequality can be applied to get[2]

$$\varphi\left(\frac{1}{b-a}\int_a^b f(x)\,dx\right) \leq \frac{1}{b-a}\int_a^b \varphi(f(x))\,dx.$$

The same result can be equivalently stated in a probability theory setting, by a simple change of notation. Let $(\Omega, \mathfrak{F}, \mathbf{P})$ be a probability space, X an integrable real-valued random variable and φ a convex function. Then:

$$\varphi\left(\mathbf{E}[X]\right) \leq \mathbf{E}[\varphi(X)].$$

In this probability setting, the measure μ is intended as a probability $\mathbf{P}$, the integral with respect to μ as an expected value $\mathbf{E}$, and the function g as a random variable X.

Notice that the equality holds if X is constant (degenerate random variable) or if φ is linear, and even if there is $A \subset \mathbb{R}$ (a Borel set, in fact) such that

$$\mathbf{Pr}(X \in A) = 1$$

and φ is a linear function over A (that is, there are $a, b \in \mathbb{R}$ such that $\varphi(x) = ax + b, \forall x \in A$).

General inequality in a probabilistic setting

More generally, let T be a real topological vector space, and X a T-valued integrable random variable. In this general setting, *integrable* means that there exists an element $\mathbf{E}[X]$ in T, such that for any element z in the dual space of T: $\mathbf{E}\left|\langle z, X\rangle\right| < \infty$, and $\langle z, \mathbf{E}[X]\rangle = \mathbf{E}[\langle z, X\rangle]$. Then, for any measurable convex function φ and any sub-σ-algebra $\mathfrak{G}$ of $\mathfrak{F}$:

$$\varphi\left(\mathbf{E}[X \mid \mathfrak{G}]\right) \leq \mathbf{E}[\varphi(X) \mid \mathfrak{G}].$$

Here $\mathbf{E}[\cdot \mid \mathfrak{G}]$ stands for the expectation conditioned to the σ-algebra $\mathfrak{G}$. This general statement reduces to the previous ones when the topological vector space T is the real axis, and $\mathfrak{G}$ is the trivial σ-algebra $\{\varnothing, \Omega\}$.[3]

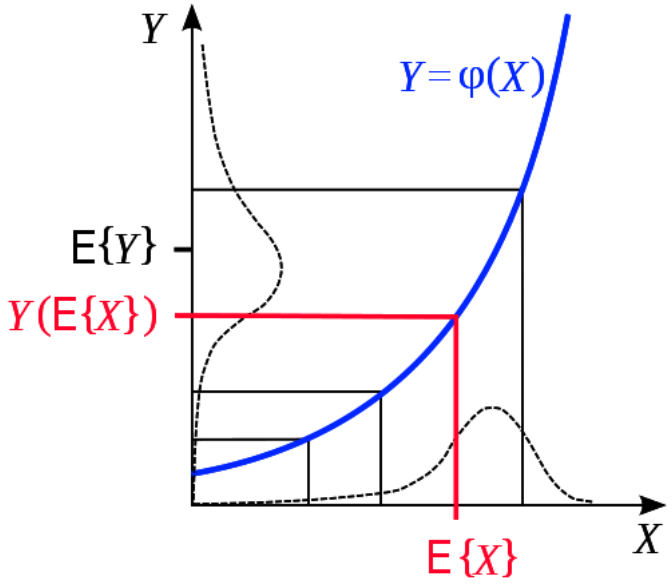
Proofs

Jensen's inequality can be proved in several ways, and three different proofs corresponding to the different statements above will be offered. Before embarking on these mathematical derivations, however, it is worth analyzing an intuitive graphical argument based on the probabilistic case where X is a real number (see figure). Assuming a hypothetical distribution of X values, one can immediately identify the position of $\mathbf{E}[X]$ and its image $\varphi(\mathbf{E}[X])$ in the graph. Noticing that for convex mappings $Y = \varphi(X)$ the corresponding distribution of Y values is increasingly "stretched out" for increasing values of X, it is easy to see that the distribution of Y is broader in the interval corresponding to $X > X_0$ and narrower in $X < X_0$ for any X_0; in particular, this is also true for $X_0 = \mathbf{E}[X]$. Consequently, in this picture the expectation of Y will always shift upwards with respect to the position of $\varphi(\mathbf{E}[X])$. A similar reasoning holds if the distribution of X covers a decreasing portion of the convex function, or both a decreasing and an increasing portion of it. This "proves" the inequality, i.e.

$$\varphi(\mathbf{E}[X]) \leq \mathbf{E}[\varphi(X)] = \mathbf{E}[Y],$$

with equality when $\varphi(X)$ is not strictly convex, e.g. when it is a straight line, or when X follows a degenerate distribution (i.e. is a constant).

The proofs below formalize this intuitive notion.



A graphical "proof" of Jensen's inequality for the probabilistic case. The dashed curve along the X axis is the hypothetical distribution of X, while the dashed curve along the Y axis is the corresponding distribution of Y values. Note that the convex mapping $Y(X)$ increasingly "stretches" the distribution for increasing values of X.

Proof 1 (finite form)

If λ_1 and λ_2 are two arbitrary nonnegative real numbers such that $\lambda_1 + \lambda_2 = 1$ then convexity of φ implies

$$\forall x_1, x_2 : \qquad \varphi\left(\lambda_1 x_1 + \lambda_2 x_2\right) \leq \lambda_1 \varphi(x_1) + \lambda_2 \varphi(x_2).$$

This can be easily generalized: if λ_1, ..., λ_n are nonnegative real numbers such that $\lambda_1 + \ldots + \lambda_n = 1$, then

$$\varphi(\lambda_1 x_1 + \lambda_2 x_2 + \cdots + \lambda_n x_n) \leq \lambda_1 \varphi(x_1) + \lambda_2 \varphi(x_2) + \cdots + \lambda_n \varphi(x_n),$$

for any x_1, ..., x_n. This *finite form* of the Jensen's inequality can be proved by induction: by convexity hypotheses, the statement is true for $n = 2$. Suppose it is true also for some n, one needs to prove it for $n + 1$. At least one of the λ_i is strictly positive, say λ_1; therefore by convexity inequality:

$$\begin{aligned}\varphi\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= \varphi\left(\lambda_1 x_1 + (1-\lambda_1)\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} x_i\right) \\ &\leq \lambda_1 \varphi(x_1) + (1-\lambda_1)\varphi\left(\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} x_i\right).\end{aligned}$$

Since

$$\sum_{i=2}^{n+1} \frac{\lambda_i}{1-\lambda_1} = 1,$$

one can apply the induction hypotheses to the last term in the previous formula to obtain the result, namely the finite form of the Jensen's inequality.

In order to obtain the general inequality from this finite form, one needs to use a density argument. The finite form can be rewritten as:

$$\varphi\left(\int x \, d\mu_n(x)\right) \leq \int \varphi(x) \, d\mu_n(x),$$

where μ_n is a measure given by an arbitrary convex combination of Dirac deltas:

$$\mu_n = \sum_{i=1}^{n} \lambda_i \delta_{x_i}.$$

Since convex functions are continuous, and since convex combinations of Dirac deltas are weakly dense in the set of probability measures (as could be easily verified), the general statement is obtained simply by a limiting procedure.

Proof 2 (measure-theoretic form)

Let g be a real-valued μ-integrable function on a probability space Ω, and let φ be a convex function on the real numbers. Since φ is convex, at each real number x we have a nonempty set of subderivatives, which may be thought of as lines touching the graph of φ at x, but which are at or below the graph of φ at all points (support lines of the graph).

Now, if we define

$$x_0 := \int_\Omega g \, d\mu,$$

because of the existence of subderivatives for convex functions, we may choose a and b such that

$$ax + b \leq \varphi(x),$$

for all real x and

$$ax_0 + b = \varphi(x_0).$$

But then we have that

$$\varphi \circ g(x) \geq ag(x) + b$$

for all x. Since we have a probability measure, the integral is monotone with $\mu(\Omega) = 1$ so that

$$\int_\Omega \varphi \circ g \, d\mu \geq \int_\Omega (ag + b) \, d\mu = a \int_\Omega g \, d\mu + b \int_\Omega d\mu = ax_0 + b = \varphi(x_0) = \varphi\left(\int_\Omega g \, d\mu \right),$$

as desired.

Proof 3 (general inequality in a probabilistic setting)

Let X be an integrable random variable that takes values in a real topological vector space T. Since $\varphi : T \to \mathbb{R}$ is convex, for any $x, y \in T$, the quantity

$$\frac{\varphi(x + \theta y) - \varphi(x)}{\theta},$$

is decreasing as θ approaches 0^+. In particular, the *subdifferential* of φ evaluated at x in the direction y is well-defined by

$$(D\varphi)(x) \cdot y := \lim_{\theta \downarrow 0} \frac{\varphi(x + \theta y) - \varphi(x)}{\theta} = \inf_{\theta \neq 0} \frac{\varphi(x + \theta y) - \varphi(x)}{\theta}.$$

It is easily seen that the subdifferential is linear in y (that is false and the assertion requires Hahn-Banach theorem to be proved) and, since the infimum taken in the right-hand side of the previous formula is smaller than the value of the same term for $\theta = 1$, one gets

$$\varphi(x) \leq \varphi(x+y) - (D\varphi)(x) \cdot y.$$

In particular, for an arbitrary sub-σ-algebra $\mathfrak{G}$ we can evaluate the last inequality when $x = \mathbf{E}[X \mid \mathfrak{G}]$, $y = X - \mathbf{E}[X \mid \mathfrak{G}]$ to obtain

$$\varphi(\mathbf{E}[X \mid \mathfrak{G}]) \leq \varphi(X) - (D\varphi)(\mathbf{E}[X \mid \mathfrak{G}]) \cdot (X - \mathbf{E}[X \mid \mathfrak{G}]).$$

Now, if we take the expectation conditioned to $\mathfrak{G}$ on both sides of the previous expression, we get the result since:

$$\mathbf{E}[[(D\varphi)(\mathbf{E}[X \mid \mathfrak{G}]) \cdot (X - \mathbf{E}[X \mid \mathfrak{G}])] \mid \mathfrak{G}] = (D\varphi)(\mathbf{E}[X \mid \mathfrak{G}]) \cdot \mathbf{E}[(X - \mathbf{E}[X \mid \mathfrak{G}]) \mid \mathfrak{G}] = 0,$$

by the linearity of the subdifferential in the y variable, and the following well-known property of the conditional expectation:

$$\mathbf{E}[(\mathbf{E}[X \mid \mathfrak{G}]) \mid \mathfrak{G}] = \mathbf{E}[X \mid \mathfrak{G}].$$

Applications and special cases

Form involving a probability density function

Suppose Ω is a measurable subset of the real line and $f(x)$ is a non-negative function such that

$$\int_{-\infty}^{\infty} f(x)\,dx = 1.$$

In probabilistic language, f is a probability density function.

Then Jensen's inequality becomes the following statement about convex integrals:

If g is any real-valued measurable function and φ is convex over the range of g, then

$$\varphi\left(\int_{-\infty}^{\infty} g(x) f(x)\,dx\right) \leq \int_{-\infty}^{\infty} \varphi(g(x)) f(x)\,dx.$$

If $g(x) = x$, then this form of the inequality reduces to a commonly used special case:

$$\varphi\left(\int_{-\infty}^{\infty} x\, f(x)\,dx\right) \leq \int_{-\infty}^{\infty} \varphi(x)\, f(x)\,dx.$$

Example: even moments of a random variable

If $g(x) = x^{2n}$, and X is a random variable, then g is convex as

$$\frac{d^2 g}{dx^2}(x) = 2n(2n-1)x^{2n-2} \geq 0 \quad \forall\, x \in \mathbb{R}$$

and so

$$g(\mathbf{E}[X]) = (\mathbf{E}[X])^{2n} \leq \mathbf{E}[X^{2n}].$$

In particular, if some even moment $2n$ of X is finite, X has a finite mean. An extension of this argument shows X has finite moments of every order $l \in \mathbb{N}$ dividing n.

Alternative finite form

Let $\Omega = \{x_1, \ldots x_n\}$, and take μ to be the counting measure on Ω, then the general form reduces to a statement about sums:

$$\varphi\left(\sum_{i=1}^{n} g(x_i)\lambda_i\right) \leq \sum_{i=1}^{n} \varphi(g(x_i))\lambda_i,$$

provided that $\lambda_i \geq 0$ and

$$\lambda_1 + \cdots + \lambda_n = 1.$$

There is also an infinite discrete form.

Statistical physics

Jensen's inequality is of particular importance in statistical physics when the convex function is an exponential, giving:

$$e^{\mathbf{E}[X]} \leq \mathbf{E}\left[e^{X}\right],$$

where the expected values are with respect to some probability distribution in the random variable X.

The proof in this case is very simple (cf. Chandler, Sec. 5.5). The desired inequality follows directly, by writing

$$\mathbf{E}\left[e^{X}\right] = e^{\mathbf{E}[X]}\,\mathbf{E}\left[e^{X-\mathbf{E}[X]}\right]$$

and then applying the inequality $e^X \geq 1 + X$ to the final exponential.

Information theory

If $p(x)$ is the true probability distribution for x, and $q(x)$ is another distribution, then applying Jensen's inequality for the random variable $Y(x) = q(x)/p(x)$ and the function $\varphi(y) = -\log(y)$ gives

$$\mathbf{E}[\varphi(Y)] \geq \varphi(\mathbf{E}[Y])$$

Therefore:

$$-D(p(x)\|q(x)) = \int p(x)\log\left(\frac{q(x)}{p(x)}\right)dx \leq \log\left(\int p(x)\frac{q(x)}{p(x)}\,dx\right) = \log\left(\int q(x)\,dx\right) = 0$$

a result called Gibbs' inequality.

It shows that the average message length is minimised when codes are assigned on the basis of the true probabilities p rather than any other distribution q. The quantity that is non-negative is called the Kullback–Leibler divergence of q from p.

Since $-\log(x)$ is a strictly convex function for $x > 0$, it follows that equality holds when $p(x)$ equals $q(x)$ almost everywhere.

Rao–Blackwell theorem

If L is a convex function and $\mathfrak{G}$ a sub-sigma-algebra, then, from the conditional version of Jensen's inequality, we get

$$L(\mathbf{E}[\delta(X) \mid \mathfrak{G}]) \leq \mathbf{E}[L(\delta(X)) \mid \mathfrak{G}] \quad \Longrightarrow \quad \mathbf{E}[L(\mathbf{E}[\delta(X) \mid \mathfrak{G}])] \leq \mathbf{E}[L(\delta(X))].$$

So if δ(*X*) is some estimator of an unobserved parameter θ given a vector of observables *X*; and if *T*(*X*) is a sufficient statistic for θ; then an improved estimator, in the sense of having a smaller expected loss *L*, can be obtained by calculating

$$\delta_1(X) = \mathbf{E}_\theta[\delta(X') \mid T(X') = T(X)],$$

the expected value of δ with respect to θ, taken over all possible vectors of observations *X* compatible with the same value of *T*(*X*) as that observed.

This result is known as the Rao–Blackwell theorem.

See also

- Karamata's inequality for a more general inequality
- Popoviciu's inequality
- Law of averages
- A proof without words of Jensen's inequality

Notes

1. Jensen, J. L. W. V. (1906). "Sur les fonctions convexes et les inégalités entre les valeurs moyennes". *Acta Mathematica*. **30** (1): 175–193. doi:10.1007/BF02418571 (https://doi.org/10.1007%2FBF02418571).
2. Niculescu, Constantin P. "Integral inequalities" (http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.625.7106&rep=rep1&type=pdf), P. 12.
3. Attention: In this generality additional assumptions on the convex function and/ or the topological vector space are needed, see Example (1.3) on p. 53 in Perlman, Michael D. (1974). "Jensen's Inequality for a Convex Vector-Valued Function on an Infinite-Dimensional Space". *Journal of Multivariate Analysis*. **4** (1): 52–65. doi:10.1016/0047-259X(74)90005-0 (https://doi.org/10.1016%2F0047-259X%2874%2990005-0).

References

- David Chandler (1987). *Introduction to Modern Statistical Mechanics*. Oxford. ISBN 0-19-504277-8.
- Tristan Needham (1993) "A Visual Explanation of Jensen's Inequality", American Mathematical Monthly 100(8):768–71.
- Nicola Fusco, Paolo Marcellini, Carlo Sbordone (1996). *Analisi Matematica Due*. Liguori. ISBN 978-88-207-2675-1.
- Walter Rudin (1987). *Real and Complex Analysis*. McGraw-Hill. ISBN 0-07-054234-1.

External links

- Jensen's Operator Inequality (https://arxiv.org/abs/math/0204049) of Hansen and Pedersen.
- Hazewinkel, Michiel, ed. (2001) [1994], "Jensen inequality" (https://www.encyclopediaofmath.org/index.php?title=p/j054220), *Encyclopedia of Mathematics*, Springer Science+Business Media B.V. / Kluwer Academic Publishers, ISBN 978-1-55608-010-4
- Weisstein, Eric W. "Jensen's inequality" (http://mathworld.wolfram.com/JensensInequality.html). *MathWorld*.
- Arthur Lohwater (1982). "Introduction to Inequalities" (http://www.mediafire.com/?1mw1tkgozzu). Online e-book in PDF format.