Inequalities, Theorems, & Functions

Contents

1 Useful Inequalities

Inequality 1.1 (Jensen's Inequality)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, X integrable, and $\varphi : \mathbb{R} \to \mathbb{R}$ some convex function with $\mathbb{E}[\,|\varphi(X)|\,] < \infty$. Then*

$$\varphi(\mathbb{E}[X]) \;\leq\; \mathbb{E}[\varphi(X)].$$

Equality *holds if X is degenerate (a.s. constant), or if φ is affine (linear plus constant) on the support of X.*

Remark 1.2 (Convexity)*.* A function $\varphi : \mathbb{R} \to \mathbb{R}$ is *convex* if for all $x, y \in \mathbb{R}$ and $\lambda \in [0, 1]$,

$$\varphi(\lambda x + (1-\lambda)y) \;\leq\; \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

Equivalently:

- If φ is differentiable, convexity is equivalent to φ' being nondecreasing.
- If φ is twice differentiable, convexity is equivalent to $\varphi''(x) \geq 0$ for all x.

Inequality 1.3 (Cauchy–Schwarz Inequality)**.** *For x, y in an inner-product space,*

$$|\langle x, y\rangle| \leq \|x\|\,\|y\|.$$

Integral form: if $f, g \in L^2(\mu)$,

$$\Big|\int fg\,d\mu\Big| \leq \Big(\int |f|^2\,d\mu\Big)^{1/2}\Big(\int |g|^2\,d\mu\Big)^{1/2}.$$

Equality *iff x, y are linearly dependent (a.e. for functions).*

Inequality 1.4 (Markov's Inequality)**.** *If $X \geq 0$ and $a > 0$,*

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

More generally, for $p > 0$, $\mathbb{P}(X \geq a) \leq \mathbb{E}[X^p]/a^p$.

Inequality 1.5 (Chebyshev's Inequality)**.** *(A generalization of Markov's inequality.)*

Let X be a random variable with finite mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathrm{Var}(X) < \infty$. Then for any $k > 0$,

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

Equivalently, for any $\epsilon > 0$,

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

Interpretation: *The probability that X deviates from its mean by more than k standard deviations is at most $1/k^2$.*

Remark 1.6 (Big-O Interpretation). Chebyshev's inequality provides an explicit bound on tail probabilities:

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

In asymptotic notation this means the probability of a k-standard-deviation deviation is bounded above on the order of $O(1/k^2)$. It is stronger than a generic Big-O statement; however, it should not be interpreted as an $o(1/k^2)$ bound, since the decay of the tail probability may in some cases be exactly of order $1/k^2$, not strictly faster.

Inequality 1.7 (Hölder's Inequality)**.** *Let $p, q \in (1, \infty)$ be* conjugate exponents*, meaning*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

If $f \in L^p(\mu)$ and $g \in L^q(\mu)$, then

$$\int |fg| \, d\mu \;\leq\; \|f\|_{L^p} \, \|g\|_{L^q}.$$

Where:

- *$L^p(\mu)$ is the space of measurable functions with finite p-norm:*

$$L^p(\mu) = \Big\{ f : \int |f|^p \, d\mu < \infty \Big\}.$$

- *The p-norm of f is*

$$\|f\|_{L^p} = \Big(\int |f|^p \, d\mu \Big)^{1/p}.$$

- *The exponents p and q are linked: e.g. if $p = 2$, then $q = 2$; if $p = 3$, then $q = 3/2$.*

Equality *holds if and only if $|f|^p$ and $|g|^q$ are proportional almost everywhere.*

Inequality 1.8 (Minkowski's Inequality (Triangle in L^p)). *For $p \in [1, \infty]$ and $f, g \in L^p(\mu)$,*

$$\|f + g\|_{L^p} \ \le \ \|f\|_{L^p} + \|g\|_{L^p}.$$

This applies strictly to sums like $f + g$. For differences, one typically uses the triangle inequality in the form

$$\|f - g\|_{L^p} = \|f + (-g)\|_{L^p} \ \le \ \|f\|_{L^p} + \|g\|_{L^p},$$

which follows by applying Minkowski with $-g$ in place of g.

Inequality 1.9 (Triangle Inequality). *For any normed vector space $(V, \|\cdot\|)$ and any $x, y \in V$,*

$$\|x + y\| \ \le \ \|x\| + \|y\|.$$

In words: the length of one side of a triangle is at most the sum of the lengths of the other two sides.

- *In $\mathbb{R}^n$ with the Euclidean norm, this corresponds to the geometric triangle inequality.*
- *In L^p spaces, this is exactly Minkowski's inequality.*

Inequality 1.10 (Young's Inequality (for products)). *If $a, b \ge 0$ and $p, q > 1$ with $1/p + 1/q = 1$, then*

$$ab \le \frac{a^p}{p} + \frac{b^q}{q}.$$

Equality *iff $a^p = b^q$ (equivalently, $a^{p-1} = b^{q-1}$).*

Inequality 1.11 (Cramér–Rao (CR) Inequality). *Let $X_1, \ldots, X_n$ have density f_θ satisfying standard regularity conditions. If $T = T(X_1, \ldots, X_n)$ is unbiased for $g(\theta)$, $\mathbb{E}_\theta[T] = g(\theta)$, then*

$$\operatorname{Var}_\theta(T) \ \ge \ \frac{\big(g'(\theta)\big)^2}{\mathcal{I}_n(\theta)}, \qquad \mathcal{I}_n(\theta) = n\,\mathcal{I}(\theta)$$

Equality *holds if the estimator is efficient, i.e. it achieves the bound. In many cases, such an estimator is also UMVU.*

Inequality 1.12 (Bernstein's Inequality (bounded/sub-exponential)). *Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}[X_i] = 0$, $|X_i| \le M$ a.s., and $\sum_{i=1}^n \operatorname{Var}(X_i) = \sigma^2$. For all $t > 0$,*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \ge t\right) \le \exp\left(-\frac{t^2}{2\sigma^2 + \frac{2}{3}Mt}\right).$$

Equivalently, for $\bar{X} = \frac{1}{n}\sum X_i$,

$$\mathbb{P}\Big(\bar{X} \ge \epsilon\Big) \le \exp\left(-\frac{n\epsilon^2}{2\operatorname{Var}(X_1) + \frac{2}{3}M\epsilon}\right).$$

Note: Stronger moment or tail assumptions can yield sharper bounds with faster rates of convergence, but for this course, we typically use Bernstein's inequality in the bounded/sub-exponential form given above (thus far).

Inequality 1.13 (Hoeffding's Inequality (bounded differences))**.** *Let $X_1, \ldots, X_n$ be independent with $a_i \le X_i \le b_i$ a.s. Set $S_n = \sum_{i=1}^n X_i$ and $\mu = \mathbb{E}[S_n]$. Then for all $t > 0$,*

$$\mathbb{P}(|S_n - \mu| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

In particular, if $a \le X_i \le b$,

$$\mathbb{P}\Big(|\bar{X} - \mathbb{E}[X_1]| \ge \epsilon\Big) \le 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

Note: same note as for Bernstein's Inequality.

Lemma 1.14. *For any two random variables X and Y with finite variances,*

$$\mathrm{Var}[X \pm Y] \;\le\; 2\,\mathrm{Var}X \;+\; 2\,\mathrm{Var}Y.$$

Proof. *From the variance of the sum of two random variables we have*

$$0 \;\le\; \mathrm{Var}[X \pm Y] \;=\; \mathrm{Var}X \;+\; \mathrm{Var}Y \;\pm\; 2\,\mathrm{Cov}[X,Y],$$

and

$$|2\,\mathrm{Cov}[X,Y]| \;\le\; \mathrm{Var}X \;+\; \mathrm{Var}Y.$$

Substituting the latter equation into the first gives the result.

Lemma 1.15. *Let X and Y be any two random variables with $\mathbb{E}[X] < \infty$, $\mathrm{Var}X < \infty$, and $\mathrm{Var}Y < \infty$. Further, assume there exists a $B \ge 0$ such that $\mathbb{P}(|Y| \le B) = 1$. Then*

$$\mathrm{Var}[XY] \;\le\; 2\,\|Y\|_\infty^2\,\mathrm{Var}X \;+\; 2\,(\mathbb{E}[X])^2\,\mathrm{Var}Y,$$

where

$$\|Y\|_\infty = \inf\{B \ge 0 : \mathbb{P}(|Y| \le B) = 1\}.$$

2 Useful Theorems

Theorem 2.1 (Fubini's Theorem)**.** *Let f be Lebesgue integrable on the rectangle $X \times Y \subset \mathbb{R}^2$. Then the double integral of f can be computed as an iterated integral:*

$$\iint_{X\times Y} f(x,y)\, d(x,y) = \int_X \left(\int_Y f(x,y)\, dy \right) dx = \int_Y \left(\int_X f(x,y)\, dx \right) dy.$$

Remark 2.2 (Tonelli (Nonnegative Fubini)). If $f \geq 0$ is measurable on $X \times Y$, then the same equalities hold (with value possibly $+\infty$) without assuming $f \in L^1$.

Theorem 2.3 (Volterra's Theorem: Differentiation under the Integral Sign)**.** *Let $f : [a,b] \times (\alpha, \beta) \to \mathbb{R}$ be continuous and $\partial f/\partial\theta$ be continuous on $[a,b] \times (\alpha,\beta)$. Then for $\theta \in (\alpha,\beta)$,*

$$\frac{d}{d\theta} \int_a^b f(x,\theta)\, dx = \int_a^b \frac{\partial}{\partial\theta} f(x,\theta)\, dx.$$

A common generalization is that continuity of $\frac{\partial}{\partial\theta} f(x,\theta)$ can be replaced by the existence of an integrable dominating function $g(x)$ *such that*

$$\left| \frac{\partial}{\partial\theta} f(x,\theta) \right| \leq g(x) \quad \textit{for all } \theta \textit{ in the parameter range.}$$

Then, by the Dominated Convergence Theorem, the derivative can be moved inside the integral:

$$\frac{d}{d\theta} \int f(x,\theta)\, dx \;=\; \int \frac{\partial}{\partial\theta} f(x,\theta)\, dx.$$

Lemma 2.4 (Bochner (1955))**.** *Suppose that the kernel K satisfies the following properties:*

$$\textit{(A1)} \qquad \int |K(u)|\, du < \infty,$$

$$\textit{(A2)} \qquad \lim_{|u|\to\infty} |uK(u)| = 0.$$

Let a function g satisfy $\int |g(u)|\, du < \infty$ and let $\{h_n\}$ be a sequence of positive constants such that $\lim_{n\to\infty} h_n = 0$. Define

$$g_n(x) \;=\; \frac{1}{h_n} \int K\left(\frac{u}{h_n} \right) g(x-u)\, du,$$

then at every point of continuity of g we have

$$\lim_{n\to\infty} g_n(x) \;=\; g(x) \int K(u)\, du.$$

Theorem 2.5 (Squeeze (Sandwich) Theorem).
(Sequences). *If a_n, x_n, b_n satisfy*

$$a_n \;\le\; x_n \;\le\; b_n \quad \textit{for all } n,$$

and

$$\lim_{n\to\infty} a_n \;=\; \lim_{n\to\infty} b_n \;=\; L,$$

then

$$\lim_{n\to\infty} x_n \;=\; L.$$

(Functions). *If $g(x) \le f(x) \le h(x)$ near x_0 and*

$$\lim_{x\to x_0} g(x) \;=\; \lim_{x\to x_0} h(x) \;=\; L,$$

then

$$\lim_{x\to x_0} f(x) \;=\; L.$$

3 Useful Functions

Definition 3.1 (Convolution)**.** Let $f, g : \mathbb{R} \to \mathbb{R}$ (or $\mathbb{C}$) be integrable functions. Their *convolution* is defined by

$$(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau, \qquad t \in \mathbb{R}.$$

Properties:

- Commutativity: $f * g = g * f$.
- Associativity: $(f * g) * h = f * (g * h)$.
- Distributivity: $f * (g + h) = f * g + f * h$.
- Convolution theorem: $\widehat{f * g}(\xi) = \widehat{f}(\xi)\, \widehat{g}(\xi)$ (Fourier transform turns convolution into multiplication).

In probability, if X and Y are independent random variables with densities f_X and f_Y, then the density of $X + Y$ is the convolution $f_X * f_Y$.

Definition 3.2 (Characteristic Function)**.** For a real-valued random variable X with distribution function F_X, the characteristic function is

$$\varphi_X(t) \;=\; \mathbb{E}\Big[e^{itX}\Big] = \int_{-\infty}^{\infty} e^{itx}\, dF_X(x), \qquad t \in \mathbb{R}.$$

If X has a probability density function f_X, then

$$\varphi_X(t) \;=\; \int_{-\infty}^{\infty} e^{itx}\, f_X(x)\, dx.$$

The characteristic function uniquely determines the distribution of X, satisfies $\varphi_X(0) = 1$, and has some other properties not noted here.

$$\varphi_{aX+b}(t) \;=\; e^{itb}\, \varphi_X(at), \qquad a, b \in \mathbb{R}.$$

Definition 3.3 (Empirical Characteristic Function)**.** Given a random sample $X_1, X_2, \ldots, X_n$ from a distribution, the empirical characteristic function (ECF) is defined by

$$\hat{\varphi}_n(t) \;=\; \frac{1}{n} \sum_{j=1}^{n} e^{itX_j}, \qquad t \in \mathbb{R}.$$

Properties:

- $\hat{\varphi}_n(0) = 1$ always.
- Each term e^{itX_j} has modulus 1, hence $\hat{\varphi}_n(t)$ always exists for any finite sample.
- $\hat{\varphi}_n(t)$ is an unbiased estimator of the true characteristic function:

$$\mathbb{E}[\hat{\varphi}_n(t)] = \varphi_X(t).$$

- As $n \to \infty$, $\hat{\varphi}_n(t) \to \varphi_X(t)$ almost surely for each fixed t (law of large numbers).

Remark 3.4 (Why Fourier Transforms?). The Fourier transform is often used to simplify problems involving integrals and convolutions by moving from the time or spatial domain into the *frequency (spectral) domain.* In this domain, operations such as convolution become easier to evaluate, and dependence on variables like x can often be removed or simplified.

Definition 3.5 (Fourier Transform)**.** For $f \in L^1(\mathbb{R})$, define

$$\widehat{f}(\xi) \;=\; \int_{-\infty}^{\infty} f(x)\, e^{-i\xi x}\, dx, \qquad \xi \in \mathbb{R}.$$

This extends to $L^2(\mathbb{R})$ (up to sets of measure zero) with Plancherel's theorem (not detailed here).

Definition 3.6 (Inverse Fourier Transform)**.** If $\widehat{f} \in L^1(\mathbb{R})$, then

$$f(x) \;=\; \frac{1}{2\pi} \int_{-\infty}^{\infty} \widehat{f}(\xi)\, e^{i\xi x}\, d\xi, \qquad x \in \mathbb{R}.$$

Under mild conditions (e.g., both f and $\widehat{f}$ in L^1), this inversion holds pointwise at continuity points of f (conditions not exhaustively detailed, but seem fairly tame compared to typical kde-based assumptions).