Proof. Again, we may prove the result just for $\theta = 0$. The proof follows mostly from Lemma A.37, with some additional observations.

1. (i) $\implies$ (ii): Assume u is convex. Since convexity is defined along lines, we see that $g(t) = u(x + tv)$ is convex for all $x, v \in \mathbb{R}^d$, and by Lemma A.37 $g''(t) \geq 0$ for all t. By (A.10) we have

$$g''(t) = \frac{d^2}{dt^2} u(x + tv) = \sum_{i=1}^{d} \sum_{j=1}^{d} u_{x_i x_j}(x) v_i v_j = v \cdot \nabla^2 u(x) v, \tag{A.25}$$

and so $\nabla^2 u(x) \geq 0$ for all $x \in \mathbb{R}^d$.

2. (ii) $\implies$ (iii): Assume (ii) holds and let $g(t) = u(x + tv)$ for $x, v \in \mathbb{R}^d$. Let $y \in \mathbb{R}^d$. Then by (A.25) we have $g''(t) \geq 0$ for all t, and so by Lemma A.37

$$g(t) \geq g(s) + g'(s)(t - s)$$

for all s, t. Set $v = y - x$, $t = 1$ and $s = 0$ to obtain

$$u(y) \geq u(x) + \nabla u(x) \cdot (y - x),$$

where we used the fact that

$$g'(0) = \frac{d}{dt}\Big|_{t=0} u(x + tv) = \nabla u \cdot v.$$

3. (iii) $\implies$ (iv): The proof is similar to Lemma A.37.

4. (iv) $\implies$ (i): Assume (iv) holds, and define $g(t) = u(x + tv)$ for $x, v \in \mathbb{R}^d$. Then we have

$$(g'(t) - g'(s))(t - s) = (\nabla u(x + tv) - \nabla u(x + sv)) \cdot v(t - s) \geq 0$$

for all t, s. By Lemma A.37 we have that g is convex for al $x, v \in \mathbb{R}^d$, from which it easily follows that u is convex. □

A.9 Probability

Here, we give a brief overview of basic probability. For more details we refer the reader to [25].

A.9.1 Basic definitions

A *probability space* is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, where $\mathcal{F}$ is a σ-algebra of measurable subsets of Ω and $\mathbb{P}$ is a nonegative measure on $\mathcal{F}$ with $\mathbb{P}(\Omega) = 1$ (i.e., a probability measure). Each $A \subset \Omega$ with $A \in \mathcal{F}$ is an event, with probability $\mathbb{P}(A)$. We think of each $\omega \in \Omega$ as a trial and if $\omega \in A$ then event A occured. For two events $A, B \in \mathcal{F}$

the union $A \cup B$ is the event that A or B occured, and the intersection $A \cap B$ is the event that both A and B occured. By subadditivity of measures we have

$$\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B),$$

which is called the *union bound*.

Example A.2. Consider rolling a 6-sided die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F}$ consists of all subsets of Ω, and $\mathbb{P}(A) = \#A/6$. If we roll the die twice, then $\Omega = \{1, 2, 3, 4, 5, 6\}^2$ and $\mathbb{P}(A) = \#A/36$. △

Example A.3. Consider drawing a number uniformly at random in the interval $[0, 1]$. Here, $\Omega = [0, 1]$, $\mathcal{F}$ is all Lebesgue measureable subsets of $[0, 1]$, and $\mathbb{P}(A)$ is the Lebesgue measure of $A \in \mathcal{F}$. △

We will from now on omit the σ-algebra $\mathcal{F}$ when referring to probability spaces.

Let $(\Omega, \mathbb{P})$ be a probability space. A *random variable* is a measurable function $X : \Omega \to \mathbb{R}^d$. That is, to each trial $\omega \in \Omega$ we associate the value $X(\omega)$.

Example A.4. In Example (A.2), suppose we win 10 times the number on the die in dollars. Then the random variable $X(\omega) = 10\omega$ describes our winnings. △

The image of Ω under X, denoted $\Omega^X = \{X(\omega) : \omega \in \Omega\} \subset \mathbb{R}^d$ is the *sample space* of X, and we often say X is a random variable on Ω^X. The random variable $X : \Omega \to \Omega^X$ defines a measure on Ω^X which we denote by $\mathbb{P}_X$. Indeed, for any $B \subset \Omega^X$, the probability that X lies in B, written $\mathbb{P}_X(X \in B)$ is

$$\mathbb{P}_X(X \in B) := \mathbb{P}(X^{-1}(B)).$$

With this new notation we can write

$$\mathbb{P}_X(X \in B) = \int_B d\,\mathbb{P}_X(x).$$

We say that X has a *density* if there exists a nonnegative Lebesgue measurable $\rho : \Omega^X \to \mathbb{R}$ such that

$$\mathbb{P}_X(X \in B) = \int_B \rho(x)\,dx.$$

Let $g : \Omega^X \to \mathbb{R}^m$. Then $Y = g(X)$ is a random variable. We define the *expectation* $\mathbb{E}_X[g(X)]$ to be

$$\mathbb{E}_X[g(X)] = \int_{\Omega^X} g(x)\,d\,\mathbb{P}_X(x) = \int_\Omega g(X(\omega))\,d\,\mathbb{P}(\omega).$$

In particular

$$\mathbb{E}_X[X] = \int_{\Omega^X} x\,d\,\mathbb{P}_X(x) = \int_\Omega X(\omega)\,d\,\mathbb{P}(\omega).$$

If X has a density then

$$\mathbb{E}_X[g(X)] = \int_{\Omega^X} g(x)\rho(x)\,dx.$$

We note that the expectation is clearly linear, so that

$$\mathbb{E}_X[f(X) + g(X)] = \mathbb{E}_X[f(X)] + \mathbb{E}_X[g(X)],$$

due to linearity of the integral.

A.9.2 Markov and Chebyshev inequalities

We introduce here basic estimates for bounding probabilities of random variables. An important result is Markov's inequality.

Proposition A.39 (Markov's inequality)**.** *Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to [0, \infty)$ be a nonnegative random variable. Then for any $t > 0$*

$$\mathbb{P}_X(X \geq t) \leq \frac{\mathbb{E}_X[X]}{t}. \tag{A.26}$$

Proof. By definition we have

$$\mathbb{P}_X(X \geq t) = \int_t^\infty d\,\mathbb{P}_X(x) \leq \int_t^\infty \frac{x}{t}\,d\,\mathbb{P}_X(x) = \frac{1}{t}\int_0^\infty x\,d\,\mathbb{P}_X(x) = \frac{\mathbb{E}_X[X]}{t}. \qquad \square$$

Markov's inequality can be improved if we have information about the variance of X. We define the *variance* of a random variable X as

$$\mathrm{Var}(X) = \mathbb{E}_X[(X - \mathbb{E}_X[X])^2]. \tag{A.27}$$

Proposition A.40 (Chebyshev's inequality)**.** *Let $(\Omega, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be a random variable with finite mean $\mathbb{E}_X[X]$ and variance $Var(X)$. Then for any $t > 0$*

$$\mathbb{P}_X(|X - \mathbb{E}_X[X]| \geq t) \leq \frac{Var\,(X)}{t^2}. \tag{A.28}$$

Proof. Let $Y = (X - \mathbb{E}_X[X])^2$. Then Y is a nonegative random variable and by Markov's inequality (A.26) we have

$$\mathbb{P}_X(|X - \mathbb{E}_X[X]| \geq t) = \mathbb{P}_X(Y \geq t^2) \leq \frac{\mathbb{E}_X[Y]}{t^2} = \frac{\mathrm{Var}\,(X)}{t^2}. \qquad \square$$

A.9.3 Sequences of indepedendent random variables

Let $(\Omega, \mathbb{P})$ be a probability space, and $X : \Omega \to \mathbb{R}^d$ be a random variable. We often want to construct other *independent* copies of the random variable X. For example, if we roll a die several times, then we have many instances of the same random variable. We clearly cannot use the same probability space for each roll of the die, otherwise all the rolls would always produce the same value (and would not be indepedent).

To construct an independent copy of X, we consider the product probability space $(\Omega \times \Omega, \mathbb{P} \times \mathbb{P})$ with the product probability measure $\mathbb{P} \times \mathbb{P}$. The product measure is the unique measure satisfying

$$(\mathbb{P} \times \mathbb{P})(A \times B) = \mathbb{P}(A)\mathbb{P}(B)$$

for all measurable $A, B \subset \Omega$. On the product probability space $\Omega^2 = \Omega \times \Omega$ the two independent copies of X are constructed via the random variable

$$(\omega_1, \omega_2) \mapsto (X(\omega_1), X(\omega_2)).$$

We normally give the random variables different names, so that $X_1(\omega_1, \omega_2) := X(\omega_1)$ and $X_2(\omega_1, \omega_2) := X(\omega_2)$. Then X_1 and X_2 are themselves random variables (now on Ω^2), and we say X_1 and X_2 are independent random variables with the same distribution as X, or *independent and identically distributed* random variables.

An important property concerns the expectation of products of independent random variables. If X_1 and X_2 are independent and identically distributed random variables with the same distribution as X (as above) then

$$\mathbb{E}_{(X_1,X_2)}[f(X_1)g(X_2)] = \mathbb{E}_X[f(X)]\mathbb{E}_X[g(X)]. \tag{A.29}$$

Indeed, we have

$$\begin{aligned}\mathbb{E}_{(X_1,X_2)}[f(X_1)g(X_2)] &= \int_\Omega \int_\Omega f(x)g(y)\, d\,\mathbb{P}_X(x)\, d\,\mathbb{P}_X(y) \\ &= \int_\Omega f(x)\, d\,\mathbb{P}_X(x) \int_\Omega g(y)\, d\,\mathbb{P}_X(y) \\ &= \mathbb{E}_X[f(X)]\mathbb{E}_X[g(X)].\end{aligned}$$

We also notice that

$$\mathbb{E}_X[f(X)] = \mathbb{E}_{(X_1,X_2)}[f(X_1)],$$

since

$$\mathbb{E}_{(X_1,X_2)}[f(X_1)] = \int_\Omega \int_\Omega f(x)\, d\,\mathbb{P}_X(x)\, d\,\mathbb{P}_X(y) = \int_\Omega f(x)\, d\,\mathbb{P}_X(x) = \mathbb{E}_X[f(x)].$$

We can continue constructing as many independent and identically distributed copies of X as we like. The construction is as follows. Let $n \geq 1$ and consider the product probability space $(\Omega^n, \mathbb{P}^n)$ with product measure

$$\mathbb{P}^n = \underbrace{\mathbb{P} \times \mathbb{P} \times \cdots \times \mathbb{P}}_{n \text{ times}}.$$

For $i = 1, \ldots, n$ we define the random variable $X_i : \Omega^n \to \mathbb{R}^d$ by

$$X_i(\omega_1, \omega_2, \ldots, \omega_n) = X(\omega_i).$$

We say that $X_1, X_2, \ldots, X_n$ is a sequence of n *independent and identically distributed* (*i.i.d.*) random variables. It is important to note how all X_i for $i = 1, \ldots, n$ are defined on the same probability space, which allows us to compute probabilities involving all the n random variables. As above, we have the product of expectations formula

$$\mathbb{E}_{(X_1,X_2,\ldots,X_n)}[f_1(X_1)f_2(X_2)\cdots f_n(X_n)] = \prod_{i=1}^{n} \mathbb{E}_X[f_i(X)]. \tag{A.30}$$

We leave it to the reader to verify (A.30). In applications of probability theory, we will not burden the notation and will write $\mathbb{P}$ in place of $\mathbb{P}^n$ and $\mathbb{E}$ in place of $\mathbb{E}_X$ and $\mathbb{E}_{(X_1,X_2,\ldots,X_n)}$. It will almost always be clear from context which probability measures and expectations are being used, and when it is not clear we will specifically denote the dependence. As above we have

$$\mathbb{E}_X[f(X)] = \mathbb{E}_{(X_1,X_2,\ldots,X_n)}[f(X_i)],$$

for any i, so the choice of which expectation to use is irrelevant. Since we do not wish to always specify the base random variable X on which the sequence is constructed, we often write X_1 or X_i in place of X.

A.9.4 Law of large numbers

To get some practice using probability, we give a proof of the weak law of large numbers, using only the tools from Sections A.9.2 and A.9.3.

Theorem A.41 (Weak law of large numbers)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent and identically distributed random variables with finite mean $\mu := \mathbb{E}[X_i]$ and variance $\sigma^2 := Var(X_i)$. Let $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\varepsilon > 0$ we have*

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mu| \geq \varepsilon) = 0. \tag{A.31}$$

Remark A.42. The limit in (A.31) shows that $S_n \to \mu$ in probability as $n \to \infty$, which is known as the *weak law of large numbers*. In fact, inspecting the proof below, we have proved the slightly stronger statement

$$\lim_{n\to\infty} \mathbb{P}(|S_n - \mu| \geq \varepsilon n^{-\alpha}) = 0,$$

for any $\alpha \in (0, \frac{1}{2})$.

Proof. Note that $\mathbb{E}[S_n] = \mu$ and compute

$$\begin{aligned}\operatorname{Var}(S_n) &= \mathbb{E}[(S_n - \mu)^2] \\ &= \frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^n X_i - \mu\right)^2\right] \\ &= \frac{1}{n^2}\mathbb{E}\left[\sum_{i,j=1}^n (X_i - \mu)(X_j - \mu)\right] \\ &= \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[(X_i - \mu)(X_j - \mu)].\end{aligned}$$

If $i \neq j$, then due to (A.30) we have $\mathbb{E}[(X_i - \mu)(X_j - \mu)] = 0$ and so

$$\operatorname{Var}(S_n) = \frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}[(X_i - \mu)^2] = \frac{\sigma^2}{n}.$$

By Chebyshev's inequality (Proposition A.40) we have

$$\mathbb{P}(|S_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

for all $\varepsilon > 0$, which completes the proof. □

A.10 Miscellaneous results

A.10.1 Vanishing lemma

Lemma A.43. *Let $U \subset \mathbb{R}^d$ be open and bounded and let $u \in C(U)$. If*

$$\int_U u(x)\varphi(x)\,dx = 0 \quad \text{for all } \varphi \in C_c^\infty(U)$$

then $u(x) = 0$ for all $x \in U$.

Proof. Let us sketch the proof. Assume to the contrary that $u(x_0) \neq 0$ at some $x_0 \in U$. We may assume, without loss of generality that $\varepsilon := u(x_0) > 0$. Since u is continuous, there exists $\delta > 0$ such that

$$u(x) \geq \frac{\varepsilon}{2} \quad \text{whenever } |x - x_0| < \delta.$$

Now let $\varphi \in C_c^\infty(U)$ be a test function satisfying $\varphi(x) > 0$ for $|x - x_0| < \delta$ and $\varphi(x) = 0$ for $|x - x_0| \geq \delta$. Then

$$0 = \int_U u(x)\varphi(x)\,dx = \int_{B(x_0,\delta)} u(x)\varphi(x)\,dx \geq \frac{\varepsilon}{2}\int_{B(x_0,\delta)} \varphi(x)\,dx > 0,$$

which is a contradiction. □