Essentials of the probability part of MASD

Torben Krüger

This is a summary of some important facts from the probability part of MASD. These essentials are not exhaustive and reading them is not a substitution for reading the lecture notes or exercise solutions.

1 Axioms of Probability Theory

Definition 1.1 (Probability space). A probability space is a pair $(\Omega, \mathbb{P})$, where Ω is the sample space (set of all outcomes) and the probability distribution $\mathbb{P}$ assigns the probability $\mathbb{P}(A) \in [0,1]$ to any event $A \subset \Omega$ and satisfies

1. $\mathbb{P}(\Omega) = 1$
2. For mutually exclusive events $A_1, A_2, \dots \subset \Omega$, additivity holds, i.e. $\mathbb{P}\big(\cup_{i=1}^{\infty} A_i\big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Theorem 1.2 (Properties of probability). Probability distributions satisfy the following properties:

1. Complement probability: $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ for any event A.
2. Inclusion-exclusion principle: For any events $A_1, \dots, A_n$ we have

$$\mathbb{P}(\cup_{i=1}^{n} A_i) \;\; = \sum_{\emptyset \neq I \subset \{1,\dots,n\}} (-1)^{1+\#I}\, \mathbb{P}(\cap_{i\in I} A_i)\,.$$

Definition 1.3 (Binomial coefficient and Factorial). The binomial coefficient 'n choose k' is defined as

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}\,,$$

with the factorial $0! = 1! = 1$, $2! = 2$, $3! = 6$ and $n! = n(n-1)(n-2)\dots 1$.

Example 1.4 (Combinatorics). The number of k-tuples $(x_1, \dots, x_k) \in \{1, \dots, n\}^k$ is n^k. These model sampling k times out of n choices with replacement

The set of permutations (one-to-one maps) of $\{1, \dots, n\}$ is denoted S_n and has $n!$ elements. These model sampling n times out of n choices without replacement.

The number of subsets of $\{1, \dots, n\}$ with k elements is $\binom{n}{k}$.

Theorem 1.5 (Binomial theorem). For $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ we have

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

2 Random variables and their distributions

Definition 2.1 (Random variables). These are basic definitions concerning random variables.

- Random variables are functions $X : \Omega \to \Sigma$ on a probability space Ω.

- The probability distribution $\mathbb{P}_X$ on Σ defined by $\mathbb{P}_X(A) := \mathbb{P}(X \in A)$ is called the distribution of X. We write $X \sim \mathbb{P}_X$ for "X has distribution $\mathbb{P}_X$".
- If Σ is countable, then X (and also its distribution $\mathbb{P}_X$) is called discrete and

  $$p_X : \Sigma \to [0,1],\ x \mapsto \mathbb{P}(X = x)$$

  is called its probability mass function (pmf). This function determines the distribution of X via

  $$\mathbb{P}(X \in A) = \sum_{x \in A} p_X(x)\,.$$

- If $\Sigma = \mathbb{R}^n$, then X (and also its distribution $\mathbb{P}_X$) is called continuous whenever there exists a probability density function (pdf) $\rho_X : \mathbb{R}^n \to \mathbb{R}$ such that

  $$\mathbb{P}(X \in A) = \int_A \rho_X(x)\mathrm{d}x = \int_{\mathbb{R}^n} \rho_X(x)\mathbb{1}_A(x)\mathrm{d}x\,,$$

  for all n-dimensional rectangles $A = A_1 \times \cdots \times A_n \subset \mathbb{R}^n$. In particular, for real continuous random variables X we have

  $$\mathbb{P}(X \in [a,b]) = \int_a^b \rho_X(x)\mathrm{d}x\,.$$

  For a continuous random 2-dimensional vector $X = (X_1, X_2)$ we have

  $$\mathbb{P}(X \in [a,b] \times [c,d]) = \mathbb{P}(X_1 \in [a,b], X_2 \in [c,d]) = \int_a^b \int_c^d \rho_X(x,y)\,\mathrm{d}y\mathrm{d}x\,.$$

Definition 2.2 (Joint and Marginal Distributions). For random variables $X_1, \ldots, X_n$ we call the distribution $\mathbb{P}_X$ of the tuple $X = (X_1, \ldots, X_n)$ the joint distribution of $X_1, \ldots, X_n$. Given the joint distribution, the (marginal) distribution of X_1 is computed by integrating out or summing over the other variables. If X is discrete this means that the pmf of X_1 is

$$p_{X_1}(x_1) = \sum_{x_2,\ldots,x_n} p_X(x_1, \ldots, x_n)\,,$$

where the sum is over all values x_i that X_i takes with positive probability. If X is continuous, then the pdf of X_1 is

$$\rho_{X_1}(x_1) = \int \rho_X(x_1, \ldots, x_n)\mathrm{d}x_2 \ldots \mathrm{d}x_n\,.$$

For the marginal distribution of any other X_i the formulas are analogously obtained by integrating out or summing over all variables corresponding to X_j with $j \neq i$. In particular, if X, Y are real continuous random variables whose joint distribution has pdf $\rho_{(X,Y)} : \mathbb{R}^2 \to \mathbb{R}$, then

$$\rho_X(x) = \int_{-\infty}^{\infty} \rho_{(X,Y)}(x,y)\mathrm{d}y\,, \qquad \rho_Y(y) = \int_{-\infty}^{\infty} \rho_{(X,Y)}(x,y)\mathrm{d}x\,.$$

Theorem 2.3 (Properties of cdf). The cumulative distribution function (cdf) of a real random variable X is $F_X : \mathbb{R} \to [0,1], x \mapsto F_X(x) = \mathbb{P}(X \leq x)$. The function $F = F_X$ satisfies

1. Monotonicity: $F(x) \leq F(y)$ for $x \leq y$
2. Right continuity: $F(x) = \lim_{y \downarrow x} F(y)$
3. Limits at infinity: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

Every function $F : \mathbb{R} \to [0,1]$ with these properties is the cdf of some random variable. Furthermore, if X is continuous with pdf ρ_X, then $F_X'(x) = \rho_X(x)$. If X is discrete with values in $\Sigma \subset \mathbb{R}$ and pmf p_X, then for every $x \in \Sigma$ we have

$$p_X(x) = \mathbb{P}(X = x) = F_X(x) - \lim_{\varepsilon \downarrow 0} F_X(x - \varepsilon)\,.$$

3 Expectation and variance

DEFINITION 3.1 (Expectation). For a discrete real random variable X with pmf $p_X : \Sigma \to [0,1]$ we define its expectation through

$$\mathbb{E}X = \sum_{x\in\Sigma} x\, p_X(x)\,.$$

For a continuous real random variable X with pdf ρ_X we define its expectation through

$$\mathbb{E}X = \int_{\mathbb{R}} x\, \rho_X(x)\mathrm{d}x\,.$$

PROPOSITION 3.2 (Rules for computing expectations). We summarise a few rules for computing expectations.

- For a discrete random variable $X : \Omega \to \Sigma$ and a real valued function $f : \Sigma \to \mathbb{R}$ we have

$$\mathbb{E}f(X) = \sum_{x\in\Sigma} f(x)\mathbb{P}(X = x) = \sum_{x\in\Sigma} f(x)p_X(x)\,.$$

- For a continuous random vector $X : \Omega \to \mathbb{R}^n$ with pdf ρ_X and $f : \mathbb{R}^n \to \mathbb{R}$ we have

$$\mathbb{E}f(X) = \int_{\mathbb{R}^n} f(x)\rho_X(x)\mathrm{d}x\,.$$

  In particular, for a real continuous random variable X with pdf ρ_X and $f : \mathbb{R} \to \mathbb{R}$ we have

$$\mathbb{E}f(X) = \int_{-\infty}^{\infty} f(x)\rho_X(x)\mathrm{d}x\,.$$

- Linearity of expectation:

$$\mathbb{E}(\alpha X + Y) = \alpha\mathbb{E}X + \mathbb{E}Y\,.$$

- Relationship between probability and expectation of indicator function:

$$\mathbb{E}\mathbb{1}(X \in A) = \mathbb{E}\mathbb{1}_A(X) = \mathbb{P}(X \in A)$$

DEFINITION 3.3 (Variance and Covariance). For real random variables X, Y we define covariance and variance through

$$\begin{aligned}\mathrm{Cov}(X,Y) &:= \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y) = \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y)\,,\\ \mathrm{Var}\,X &:= \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}(X^2) - (\mathbb{E}X)^2\,.\end{aligned}$$

PROPOSITION 3.4 (Properties of covariance and variance). Covariance and variance obey the following computational rules (Here X, Y, Z are random variables and $\alpha, \mu \in \mathbb{R}$).

1. Symmetry: $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$
2. Shift invariance: $\mathrm{Cov}(X+\mu, Y) = \mathrm{Cov}(X,Y)$
3. Bilinearity:

$$\begin{aligned}\mathrm{Cov}(\alpha X + Z, Y) &= \alpha\mathrm{Cov}(X,Y) + \mathrm{Cov}(Z,Y)\\ \mathrm{Cov}(Y, \alpha X + Z) &= \alpha\mathrm{Cov}(Y,X) + \mathrm{Cov}(Y,Z)\end{aligned}$$

4. Relation between variance and covariance:

$$\mathrm{Cov}(X,X) = \mathrm{Var}\,X\,.$$

5. Affine transformation:

$$\mathrm{Var}(\alpha X + \mu) = \alpha^2\,\mathrm{Var}\,X\,.$$

4 Conditional probability and independence

DEFINITION 4.1 (Conditional probability). The conditional probability of A, given B with $\mathbb{P}(B) > 0$ is

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}\,.$$

The assignment $A \mapsto \mathbb{P}(A|B)$ is a probability distribution.

THEOREM 4.2 (Law of total probability). Let $B_1, \ldots, B_n$ be a partition of the sample space with $\mathbb{P}(B_i) > 0$. Then

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A|B_i)\mathbb{P}(B_i)\,, \qquad \text{for any event } A\,.$$

THEOREM 4.3 (Bayes' rule). Let $B_1, \ldots, B_n$ be a partition of the sample space with $\mathbb{P}(B_i) > 0$. Then

$$\mathbb{P}(B_i|A) = \frac{\mathbb{P}(A|B_i)\mathbb{P}(B_i)}{\sum_{j=1}^{n} \mathbb{P}(A|B_j)\mathbb{P}(B_j)}\,, \qquad \text{for any event } A \text{ with } \mathbb{P}(A > 0)\,.$$

DEFINITION 4.4 (Independence of two events). Two events A, B are called independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)\,.$$

DEFINITION 4.5 (Independence many random variables). Random variables $X_1, \ldots, X_n$ are called independent if

$$\mathbb{P}(X_1 \in A_1, \ldots, X_n \in A_n) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i) \tag{4.1}$$

holds for all possible choices of $A_1, \ldots, A_n$.

THEOREM 4.6 (Product rule). The random variables $X_1, \ldots, X_n$ are independent if and only if

$$\mathbb{E} \prod_{i=1}^{n} f_i(X_i) = \prod_{i=1}^{n} \mathbb{E} f_i(X_i)\,, \tag{4.2}$$

holds for all possible choices of real valued functions f_i. In particular, for discrete random variables $X_1, \ldots, X_n$, independence is equivalent to factorisation of the joint pmf

$$p_X(x_1, \ldots, x_n) = p_{X_1}(x_1) \ldots p_{X_n}(x_n)$$

and for continuous random variables $X_1, \ldots, X_n$, independence is equivalent to factorisation of the joint pdf

$$\rho_X(x_1, \ldots, x_n) = \rho_{X_1}(x_1) \ldots \rho_{X_n}(x_n)\,,$$

where $X = (X_1, \ldots, X_n)$.

DEFINITION 4.7 (i.i.d.). We use the abbreviation i.i.d. for independent and identically distributed random variables, i.e. we say that $X_1, \ldots, X_n$ are i.i.d. if they are independent and have all the same distribution $X_i \sim \mathbb{P}_{X_1}$ for all i.

LEMMA 4.8 (Convolution rule). Let X and Y be independent continuous random variables with pdf ρ_X and ρ_Y, respectively. Then $X + Y$ has pdf

$$\rho_{X+Y}(z) = \int \rho_X(x)\rho_Y(z - x)\mathrm{d}x\,.$$

LEMMA 4.9 (Bienaymé formula). Let $X_1, \ldots, X_n$ be pairwise uncorrelated. Then

$$\operatorname{Var} \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} \operatorname{Var} X_i\,.$$

THEOREM 4.10 (Weak law of large numbers). Let $X_1, \ldots, X_n$ be i.i.d. random variables and $\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

$$\mathbb{P}(|\overline{X}_n - \mathbb{E}X_1| \geq \varepsilon) \leq \frac{\operatorname{Var} X_1}{\varepsilon^2 n} \to 0\,, \qquad n \to \infty\,.$$

Examples of Distributions

- Discrete uniform distribution: $\mathbb{P}(A) = \frac{\#A}{\#\Omega}$ on finite sample space Ω with pmf $\mathbb{P}(\{\omega\}) = \frac{1}{\#\Omega}$.
- Bernoulli distribution: $\mathbb{P} = \mathrm{Bern}(p)$ on $\{0,1\}$ with $\mathbb{P}(\{0\}) = 1-p$ and $\mathbb{P}(\{1\}) = p$ for $p \in (0,1)$.
- Binomial distribution: $\mathbb{P} = \mathrm{Bin}(n,p)$ for $p \in (0,1)$ on $\{0,\dots,n\}$ with pmf

$$\mathbb{P}(\{k\}) = \binom{n}{k} p^k (1-p)^{n-k}.$$

- Poisson distribution: $\mathbb{P} = \mathrm{Poi}(\lambda)$ for $\lambda > 0$ on non-negative integers $\{0,1,2,3,\dots\}$ with pmf

$$\mathbb{P}(\{k\}) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}.$$

- Continuous uniform distribution on $\Omega \subset \mathbb{R}^n$: $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$ with pdf $\rho(x) = \frac{1}{|\Omega|} \mathbb{1}_\Omega(x)$. In particular, we have the uniform distribution on the interval $[a,b] \subset \mathbb{R}$ with pdf

$$\rho(x) = \frac{1}{b-a} \mathbb{1}_{[a,b]}(x).$$

- Exponential distribution: $\mathbb{P} = \mathrm{Exp}(\alpha)$ for $\alpha > 0$ on $\mathbb{R}$ (or more precisely on $(0,\infty)$) and pdf

$$\rho(x) = \alpha \, \mathrm{e}^{-\alpha x} \mathbb{1}_{(0,\infty)}(x).$$

- Gaussian distribution: $\mathbb{P} = N(a,v)$ for $a \in \mathbb{R}$ and $v > 0$ on $\mathbb{R}$ with pdf

$$\rho(x) = \frac{1}{\sqrt{2\pi v}} \, \mathrm{e}^{-\frac{(x-a)^2}{2v}}.$$

  Here, a is the expectation and v the variance of a $N(a,v)$-distributed random variable. The distribution $N(0,1)$ is called standard normal distribution.

- 2-dimensional Gaussian distribution: $\mathbb{P} = N(m,A)$ for $m \in \mathbb{R}^2$ and $A \in \mathbb{R}^{2\times 2}$ positive definite on $\mathbb{R}^2$ with pdf

$$\rho_X(x) = \frac{\mathrm{e}^{-\frac{1}{2}(x-m)^T A^{-1}(x-m)}}{2\pi\sqrt{\det A}}.$$

2-dimensional Gaussian random variables satisfy the following computational rules

PROPOSITION 4.11 (2-dimensional Gaussian). Let X be a 2-dimensional Gaussian vector with mean $m \in \mathbb{R}^2$ and covariance matrix $A = (a_{ij})_{i,j=1}^2$, i.e. $X = (X_1, X_2) \sim N(m,A)$. Then

1. The expectation of X_i is $\mathbb{E}X_i = m_i$ for $i = 1,2$.
2. The covariances of the entries of X are $\mathrm{Cov}(X_i, X_j) = a_{ij}$ for all $i,j = 1,2$.
3. The marginals X_1 and X_2 are also Gaussian with $X_1 \sim N(m_1, a_{11})$ and $X_2 \sim N(m_2, a_{22})$.
4. X_1, X_2 are independent if and only if $\mathrm{Cov}(X_1, X_2) = 0$
5. For any $T \in \mathbb{R}^{2\times 2}$ and $b \in \mathbb{R}^2$ the affine transformation $TX + b$ is a Gaussian vector.

Tools from Calculus

For the probability part we introduced the following additional tools from calculus.

- Gaussian integral formula for $\alpha > 0$ and $\beta \in \mathbb{R}$:

$$\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{\alpha}{2}x^2+\beta} \mathrm{d}x = \sqrt{\frac{2\pi}{\alpha}}\, \mathrm{e}^{\frac{\beta^2}{2\alpha}} .$$

- Convolution of two functions $f, g : \mathbb{R} \to \mathbb{R}$ defined through

$$(f \star g)(x) := \int_{\mathbb{R}} f(y) g(x-y) \mathrm{d}y \, .$$

- Gamma function, defined through

$$\Gamma(s) := \int_0^{\infty} x^{s-1} \mathrm{e}^{-x} \, \mathrm{d}x$$

  for $s > 0$ with the relation to factorials

$$\Gamma(k+1) = (-1)^k \left(\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k} \int_0^{\infty} \mathrm{e}^{-\alpha x} \, \mathrm{d}x \right) \Bigg|_{\alpha=1} = (-1)^k \left(\frac{\mathrm{d}^k}{\mathrm{d}\alpha^k} \frac{1}{\alpha} \right) \Bigg|_{\alpha=1} = k! \, .$$