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, the online course or exercise solutions.

1 Axioms of Probability Theory

DEFINITION 1.1 (Probability space). A probability space is a pair $(S, \mathbb{P})$, where S 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 $A \subset S$ and satisfies

1. $\mathbb{P}(S) = 1$
2. For mutually exclusive events $A_1, A_2, \dots \subset S$, 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 (e.g. shuffling of a card deck).

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 $X : S \to \mathbb{R}$ on a probability space S.

- The probability distribution $\mathbb{P}_X$ on $R_X = \{X(s) : s \in S\}$ 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 R_X is countable, then X (and also its distribution $\mathbb{P}_X$) is called discrete and

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

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 $R_X \subset \mathbb{R}$ is uncountable, then X (and also its distribution $\mathbb{P}_X$) is called continuous whenever there exists a probability density function (PDF) $f_X : \mathbb{R}^n \to \mathbb{R}$ such that

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

for all a<b.

Definition 2.2 (Joint and Marginal Distributions). For two discrete random variables X, Y we call $\mathbb{P}_{X,Y}(x,y) := \mathbb{P}(X = x, Y = y)$ the joint PMF of X and Y. Two random variables are called jointly continuous if they have a joint PDF which satisfies

$$\mathbb{P}(X \in [a,b], Y \in [c,d]) = \int_a^b \int_c^d f_{X,Y}(x,y)\mathrm{d}y\mathrm{d}x\,,$$

for all $a < b$ and $c < d$. Given the joint distribution of X, Y, the (marginal) distribution of X is computed by integrating out or summing over the other variables. If X is discrete this means that the PMF of X is

$$P_X(x) = \sum_{y \in R_Y} P_{X,Y}(x,y)\,,$$

If X, Y are jointly continuous, then the PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\mathrm{d}y\,.$$

For the marginal distribution of Y the formulas are analogously obtained by integrating out or summing the variable corresponding to 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 a continuous random variable with PDF f_X, then the following relationships between CDF and PDF hold

$$F_X'(x) = f_X(x)\,, \qquad F_X(x) = \int_{-\infty}^{x} f_X(y)\mathrm{d}y\,.$$

If X is discrete with values in $R_X \subset \mathbb{R}$ and PMF pP_X, then for every $x \in R_X$ we have

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

The CDF is particularly useful when computing the distribution of $\min\{X, Y\}$ or $\max\{X, Y\}$ for two independent random variables X and Y because

$$\mathbb{P}(\max\{X,Y\} \leq z) = \mathbb{P}(X \leq z, Y \leq z) = \mathbb{P}(X \leq z)\mathbb{P}(Y \leq z) = F_X(z)F_Y(z)\,.$$

3 Expectation and variance

Definition 3.1 (Expectation). For a discrete real random variable X with PMF $P_X : R_X \to [0,1]$ we define its expectation through

$$\mathbb{E}X = \sum_{x \in R_X} x\, P_X(x)\,.$$

For a continuous real random variable X with PDF f_X we define its expectation through

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

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

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

$$\mathbb{E}g(X) = \sum_{x \in R_X} g(x)\mathbb{P}(X = x) = \sum_{x \in R_X} g(x) P_X(x)\,.$$

- For a continuous random variable $X : S \to \mathbb{R}$ with PDF f_X and $g : \mathbb{R} \to \mathbb{R}$ we have

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

- If the discrete random variables X, Y have joint PMF $P_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y)$ and $g : \mathbb{R}^2 \to \mathbb{R}$, then

$$\mathbb{E}\,g(X,Y) = \sum_{x \in R_X, y \in R_Y} g(x,y) P_{X,Y}(x,y)$$

- If the random variables X, Y have a joint PDF $f_{X,Y}$ and $g : \mathbb{R}^2 \to \mathbb{R}$, then

$$\mathbb{E}\,g(X,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\mathrm{d}x\mathrm{d}y$$

- 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 of random variables). Random variables X, Y are called independent if

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \tag{4.1}$$

holds for all possible choices of A, B.

Theorem 4.6 (Product rule). The random variables X, Y are independent if and only if their joint PMF (in the discrete case) factorises

$$P_{X,Y}(x, y) = P_X(x)P_Y(y)$$

or their PDF (in the continuous case) factorises

$$f_{X,Y}(x, y) = f_X(x)f_Y(y) .$$

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 PDFs f_X and f_Y, respectively. Then $X + Y$ has PDF

$$f_{X+Y}(z) = \int f_X(x)f_Y(z - x)\mathrm{d}x .$$

Lemma 4.9 (Bienaymé formula). Let X, Y be uncorrelated, i.e. $\mathrm{Cov}(X, Y) = 0$ (e.g. when they are independent). Then

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

Examples of Distributions

- Discrete uniform distribution: $\mathbb{P}(A) = \frac{|A|}{|S|}$ on finite sample space S with PMF $\mathbb{P}(\{s\}) = \frac{1}{|S|}$.
- Bernoulli distribution: $\mathbb{P} = \mathrm{Bernoulli}(p)$ on $\{0, 1\}$ with $\mathbb{P}(\{0\}) = 1 - p$ and $\mathbb{P}(\{1\}) = p$ for $p \in (0, 1)$.

- Binomial distribution: $\mathbb{P} = \text{Binomial}(n,p)$ 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: $\mathbb{P} = \text{Poisson}(\lambda)$ 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 $S \subset \mathbb{R}^n$: $\mathbb{P}(A) = \frac{\text{Vol}A}{\text{Vol}S}$ with PDF $f(x) = \frac{1}{\text{Vol}S} \mathbb{1}_S(x)$. In particular, we have the uniform distribution on the interval $[a,b] \subset \mathbb{R}$ with PDF

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

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

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

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

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

  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 is a distribution on $\mathbb{R}^2$ with (joint) PDF

$$f_X : \mathbb{R}^2 \to \mathbb{R}\,, \qquad f_X(x) = \frac{\mathrm{e}^{-\frac{1}{2}(x-m)^T A^{-1}(x-m)}}{2\pi\sqrt{\det A}}\,, \qquad X = (X_1, X_2)\,,\ x = (x_1, x_2) \in \mathbb{R}^2\,.$$