1228 Probabilistic Graphical Models
H. Zhu and J. Boedecker
April 30, 2026

Mathematical Background

Contents

1 Probability theory

1.1 Basic concepts

In this course we will adhere to the Bayesian interpretation of probability, according to which probabilities encode degrees of belief about events in the world and data are used to strengthen, update, or weaken those degrees of belief. For example, if A stands for an event, then $\mathbf{P}(A \mid K)$ stands for a person's subjective belief in event A given a body of knowledge K. In defining probability expressions, we often simply write $\mathbf{P}(A)$, leaving out the symbol K. However, when the background information undergoes changes, we need to identify specifically the assumptions that account for our beliefs and explicitly articulate K (or some of its elements).

In the Bayesian formalism, belief measures obey the three basic axioms of probability calculus:

- $0 \leq \mathbf{P}(A) \leq 1$,
- $\mathbf{P}(\text{sure proposition}) = 1$,
- $\mathbf{P}(A \text{ or } B) = \mathbf{P}(A) + \mathbf{P}(B)$ if A and B are mutually exclusive.

The third axiom states that the belief assigned to any set of events is the sum of the beliefs assigned to its nonintersecting components. Because any event A can be written as the union of the joint events $(A \wedge B)$ and $(A \wedge \neg B)$, their associated probabilities are given by

$$\mathbf{P}(A) = \mathbf{P}(A, B) + \mathbf{P}(A, \neg B),$$

where $\mathbf{P}(A, B)$ is short for $\mathbf{P}(A \wedge B)$. More generally, if $\{B_1, \ldots, B_n\}$ is a set of exhaustive and mutually exclusive propositions, then $\mathbf{P}(A)$ can be computed from $\mathbf{P}(A, B_i), i = 1, \ldots, n$, by using the sum

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}(A, B_i),$$

which has come to be known as the *law of total probability*. The operation of summing up probabilities over all B_i is also called *marginalizing* over B; and the resulting probability, $\mathbf{P}(A)$, is called the *marginal probability* of A.

The basic expressions in the Bayesian formalism are statements about *conditional probabilities*. For example, $\mathbf{P}(A \mid B)$, which specify the belief in A under the assumption that B is known with absolute certainty. If $\mathbf{P}(A \mid B) = \mathbf{P}(A)$, we say that A and B are *independent*, since our belief in A remains unchanged upon learning the truth of B. If $\mathbf{P}(A \mid B, C) = \mathbf{P}(A \mid C)$, we say that A and B are *conditionally independent* given C; that is, once we know C, learning B would not change our belief in A. Bayesian philosophers see the conditional relationship as more basic than that of joint events, *i.e.*, more compatible with the organization of human knowledge. In this view, B serves as a pointer to a context or frame of knowledge, and $A \mid B$ stands for an event A in the context specified by B. Consequently, empirical knowledge invariably will be encoded in conditional probability statements, whereas belief in joint events will be computed from those statements via the product

$$\mathbf{P}(A, B) = \mathbf{P}(A \mid B)\, \mathbf{P}(B). \tag{1.1}$$

A useful generalization of the product rule (1.1) is the *chain rule* formula. It states that if we have a set of n events, $E_1, \ldots, E_n$, then the probability of the joint event $(E_1, \ldots, E_n)$ can be written as a product of n conditional probabilities:

$$\mathbf{P}(E_1, \ldots, E_n) = \mathbf{P}(E_1)\, \mathbf{P}(E_2 \mid E_1) \cdots \mathbf{P}(E_n \mid E_{n-1}, \ldots, E_1).$$

The heart of Bayesian inference lies in the celebrated inversion formula, *i.e.*, the *Bayes' theorem*

$$\mathbf{P}(H \mid e) = \frac{\mathbf{P}(e \mid H)\, \mathbf{P}(H)}{\mathbf{P}(e)}, \tag{1.2}$$

which states that the belief we accord a hypothesis H upon obtaining evidence e can be computed by multiplying our previous belief $\mathbf{P}(H)$ by the *likelihood* $\mathbf{P}(e \mid H)$ that e will materialize if H is true. This $\mathbf{P}(H \mid e)$ is sometimes called the *posterior probability*, and $\mathbf{P}(H)$ is called the *prior probability*. The denominator $\mathbf{P}(e)$ of (1.2) hardly enters into consideration because it is merely a normalizing constant.

1.2 Random variables and expectations

By a *variable* we will mean an attribute, measurement or inquiry that may take on one of several possible values from a specified domain. If we have probabilities attached to the possible values that a variable may attain, we will call that variable

a *random variable*. Most of our analysis will concern a finite set V of random variables (also called *partitions*) where each variable $X \in V$ may take on values from a finite *domain* $\mathbf{dom}\, X$. We will use capital letters X, Y, Z for variable names and lowercase letters x, y, z as generic symbols for specific values taken by the corresponding variables. Clearly, the statement $X = x$ defines a set of exhaustive and mutually exclusive events, one for each value of x.

In most of our discussions, we will not make notational distinction between variables and sets of variables, because a set of variables essentially defines a compound variable whose domain is the Cartesian product of the domains of the individual constituents in the set. Thus, if Z stands for the set $\{X, Y\}$, then z stands for pairs (x, y) such that $x \in \mathbf{dom}\, X$ and $y \in \mathbf{dom}\, Y$. When the distinction between variables and sets of variables requires special emphasis, indexed letters $X_1, \ldots, X_n$ will be used to represent individual variables. Besides, we shall consistently use the abbreviation $\mathbf{P}(x)$ for the probabilities $\mathbf{P}(X = x)$, $x \in \mathbf{dom}\, X$.

When the values of a random variable X are real numbers, *i.e.*, $x \in \mathbf{R}$, X is called a *real* random variable; one can then define the *mean* or *expected value* of X as

$$\mathbf{E}\, X = \sum_x x\, \mathbf{P}(x) \tag{1.3}$$

and the conditional mean of X, given event $Y = y$, as

$$\mathbf{E}(X \mid y) = \sum_x x\, \mathbf{P}(x \mid y).$$

The expectation of any function g of X is defined as

$$\mathbf{E}\, g(X) = \sum_x g(x)\, \mathbf{P}(x).$$

In particular, the function $g(X) = (X - \mathbf{E}\, X)^2$ has received much attention; its expectation is called the *variance* of X, denoted $\mathbf{var}\, X$;

$$\mathbf{var}\, X = \mathbf{E}\left((X - \mathbf{E}\, X)^2\right).$$

We are often interested in the square root of the variance $\mathbf{var}\, X$, which is called the *standard deviation* of the random variable X, denoted as

$$\sigma(X) = \sqrt{\mathbf{var}\, X}.$$

When function g is a two-variable function with $g(X, Y) = (X - \mathbf{E}\, X)(Y - \mathbf{E}\, Y)$, its expectation is known as the *covariance* of X and Y,

$$\mathbf{cov}(X, Y) = \mathbf{E}((X - \mathbf{E}\, X)(Y - \mathbf{E}\, Y)),$$

and which is often normalized to yield the *correlation coefficient*

$$\rho(X, Y) = \frac{\mathbf{cov}(X, Y)}{\sigma(X)\sigma(Y)}$$

and the *regression coefficient* (of X on Y)

$$r(X,Y) = \rho(X,Y)\frac{\sigma(X)}{\sigma(Y)} = \frac{\mathbf{cov}(X,Y)}{\mathbf{var}\, Y}.$$

The foregoing definitions apply to discrete random variables, *i.e.*, variables that take on finite or denumerable sets of values on $\mathbf{R}$. The treatment of expectation and correlation is more often applied to continuous random variables, which are characterized by a *density function* $p(x)$ defined as follows:

$$\mathbf{P}(a \leq X \leq b) = \int_a^b p(x)\ dx$$

for any $a, b \in \mathbf{R}, a < b$. If X is discrete, then $p(x)$ coincides with the probability function $\mathbf{P}(x)$, once we interpret the integral through the translation

$$\int_{-\infty}^{\infty} p(x)\ dx \quad \Longleftrightarrow \quad \sum_x \mathbf{P}(x).$$

This translation should be kept in mind whenever summation is used through the course. For example, the expected value of a continuous random variable X can be transformed from (1.3) to

$$\mathbf{E}\, X = \int_{-\infty}^{\infty} xp(x)\ dx,$$

with analogous translations for the variance, correlation, and so forth.

Example. *Gaussian distribution.* A random variable X has a Gaussian distribution with mean μ and variance σ^2, denoted $\mathcal{N}(\mu, \sigma^2)$, if it has the density function

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

A Gaussian distribution has a bell-like curve, where the mean parameter μ controls the location of the peak, that is, the value for which the Gaussian gets its maximum value. The variance parameter σ^2 determines how peaked the Gaussian is: the smaller the variance, the more peaked the Gaussian. A standard Gaussian is one with mean 0 and variance 1. Figure 1 shows the density function of a few different Gaussian distributions.

More technically, the density function of Gaussian distribution is specified as an exponential, where the expression in the exponent corresponds to the square of the number of standard deviations σ that x is away from the mean μ. The probability of x decreases exponentially with the square of its deviation from the mean, as measured in units of its standard deviation.

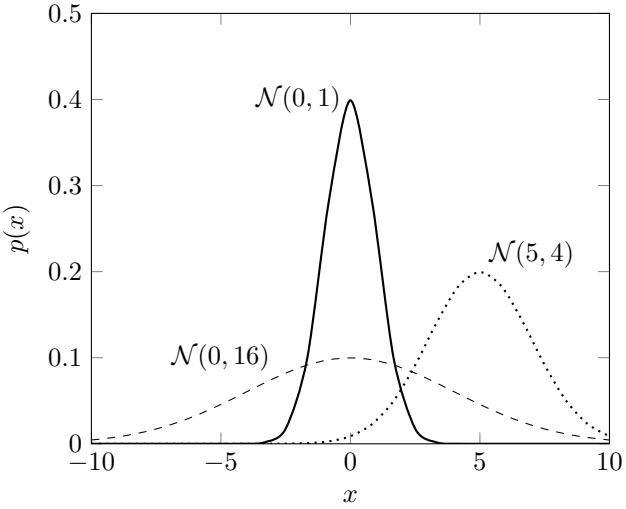

Figure 1 Density function of three example Gaussian distributions.

1.3 Set conditional independence and graphoids

Let $V = \{V_1, V_2, \ldots\}$ be a finite set of variables and let X, Y, Z stand for any three subsets of variables in V. The sets X and Y are said to be *conditionally independent* given Z if and only if $\mathbf{P}(x \mid y, z) = \mathbf{P}(x \mid z)$ for all y, z that $\mathbf{P}(y, z) > 0$ holds. In words, learning the value of Y does not provide additional information about X, once we know Z. We will use the notation $(X \perp\!\!\!\perp Y \mid Z)$ to denote the conditional independence of X and Y given Z. Unconditional independence (also called *marginal independence*) will be denoted by $(X \perp\!\!\!\perp Y \mid \emptyset)$, which says $\mathbf{P}(x \mid y) = \mathbf{P}(x)$ for all y that $\mathbf{P}(y) > 0$ holds. Note that $(X \perp\!\!\!\perp Y \mid Z)$ implies the conditional independence of all pairs of variables $V_i \in X$ and $V_j \in Y$, but the converse is not necessarily true.

In the following we list some properties satisfied by the conditional independence relation $(X \perp\!\!\!\perp Y \mid Z)$:

- *Symmetry*: $(X \perp\!\!\!\perp Y \mid Z) \implies (Y \perp\!\!\!\perp X \mid Z)$.
- *Decomposition*: $(X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z)$.
- *Weak union*: $(X \perp\!\!\!\perp YW \mid Z) \implies (X \perp\!\!\!\perp Y \mid ZW)$.
- *Contraction*: $(X \perp\!\!\!\perp Y \mid Z)$ & $(X \perp\!\!\!\perp W \mid ZY) \implies (X \perp\!\!\!\perp YW \mid Z)$.
- *Intersection*: $(X \perp\!\!\!\perp W \mid ZY)$ & $(X \perp\!\!\!\perp Y \mid ZW) \implies (X \perp\!\!\!\perp YW \mid Z)$.

Note that intersection is only valid for strictly positive probability distributions. These properties are called *graphoid axioms* and the proof of them can be derived from the definition of conditional independence and the basic axioms of probability theory.

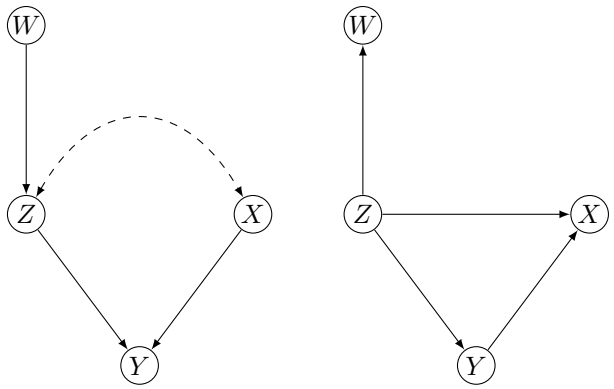

Figure 2 *Left.* A graph containing both directed and bidirected edges. *Right.* A directed acyclic graph with the same skeleton as the left graph.

2 Graphs

A graph G consists of a set V of *vertices* (or *nodes*) and a set E of *edges* (or *links*) that connect some pairs of vertices. The vertices in our graphs will correspond to variables, and the edges will denote a certain relationship that holds in pairs of variables, the interpretation of which will vary with the application. Two vertices connected by an edge are called *adjacent*.

Each edge in a graph can be either directed (marked by a single arrowhead on the edge), or undirected (unmarked links). In some applications we will also use 'bidirected' edges to denote the existence of unobserved common causes (sometimes called *confounders*). These edges will be marked as dotted curved arcs with two arrowheads as shown in figure 2. If all edges are directed (as in figure 2, right), we then have a *directed graph*. If we strip away all arrowheads from the edges in a graph G, the resultant undirected graph is called the *skeleton* of G. A *path* in a graph is a sequence of edges (*e.g.*, $((W,Z),(Z,Y),(Y,X),(X,Z))$ in figure 2) such that each edge starts with the vertex ending the preceding edge. In other words, a path is any unbroken, nonintersecting route traced out along the edges in a graph, which may go either along or against the arrows. If every edge in a path is an arrow that points from the first to the second vertex of the pair, we have a *directed path*. In the left graph of figure 2, for example, the path $((W,Z),(Z,Y))$ is directed, but the paths $((W,Z),(Z,Y),(Y,X))$ and $((W,Z),(Z,X))$ are not. If there exists a path between two vertices in a graph, then the two vertices are said to be *connected*; else they are *disconnected*.

Directed graphs may include directed cycles (*e.g.*, $X \to Y, Y \to X$), representing mutual causation or feedback processed, but not self-loops (*e.g.*, $X \to X$). A graph (like the two in figure 2) that contains no directed cycles in called *acyclic*. A graph that is both directed and acyclic (figure 2, right) is called a *directed acyclic graph* (DAG), and such graphs will occupy much of our discussion through the

course. We make free use of the terminology of kinship (*e.g.*, *parents*, *children*, *descendants*, *ancestors*, *spouses*) to denote various relationships in a graph. These kinship relations are defined along the full arrows in the graph, including arrows that form directed cycles but ignoring bidirected and undirected edges. In the left graph of figure 2, for example, Y has two parents (X and Z), three ancestors (X, Z, and W), and no children, while X has no parents (hence, no ancestors), one spouse (Z), and one child (Y). A family in a graph is a set of nodes containing a node and all its parents. For example, $\{W\}$, $\{Z, W\}$, $\{X\}$, and $\{Y, Z, X\}$ are the families in the graphs of figure 2.

A node in a directed graph is called a *root* if it has no parents and a *sink* if it has no children. Every DAG has at least one root and at least one sink. A connected DAG in which every node has at most one parent is called a *tree*, and a tree in which every node has at most one child is called a *chain*. A graph in which every pair of nodes is connected by an edge is called *complete*. The graphs in figure 2, for instance, is connected but not complete, because the pairs (W, X) and (W, Y) are note adjacent.

Bibliography

Our introduction about probability theory is mostly based on [Pea09, §1.1], with some additional information from [KF09, §2.1]. Many textbooks on the subject, *e.g.*, [Fel50, HPS71], or the appendix to [Sup70], can be referred to for additional mathematical machinery in probability.

The basic probability axioms introduced at the beginning of this chapter deviate a bit from the standard statement, which can be found in the textbook [KB18].

The graphoid axioms listed in this chapter were first introduced in [Daw79] and [Spo80] in a slightly different form, and were independently proposed by Pearl and Paz [PP87] to characterize the relationships between graphs and informational relevance. Geiger and Pearl [GP93] present an in-depth analysis. The intuitive interpretation of the graphoid axioms is discussed in [Pea88, page 85].

Our brief discussion about the terminology of graphs is adapted from [Pea09, §1.2.1]. A more detailed introduction on graphs related to the topic of probabilistic graphical models can be found in [KF09]. The textbook [Wes01] specifically for graph theory can be referred to for additional mathematical machinery.

The notation introduced in this chapter and will be used in the following chapters are mostly based on [BV04]. Some notation related to probability are taken from those used in [Pea09] and [KF09].

References

[BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[Daw79] A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(1):1–31, 1979.

[Fel50] W. Feller. *Probability Theory and Its Applications*. John Wiley & Sons, 1950.

[GP93] D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21(4):2001–2021, 1993.

[HPS71] P. G. Hoel, S. C. Port, and C. J. Stone. *Introduction to Probability Theory*. Houghton Mifflin Company, 1971.

[KB18] A. N. Kolmogorov and A. T. Bharucha-Reid. *Foundations of the Theory of Probability: Second English Edition*. Courier Dover Publications, 2018.

[KF09] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[Pea09] J. Pearl. *Causality*. Cambridge University Press, 2nd edition, 2009.

[PP87] J. Pearl and A. Paz. Graphoids: A graph-based logic for reasoning about relevance relations. *Advances in Artificial Intelligence-II*, pages 357–363, 1987.

[Spo80] W. Spohn. Stochastic independence, causal independence, and shieldability. *Journal of Philosophical Logic*, 9:73–99, 1980.

[Sup70] P. Suppes. *A Probabilistic Theory of Causality*. North-Holland Publishing Co., 1970.

[Wes01] D. B. West. *Introduction to Graph Theory*. Prentice hall Upper Saddle River, 2nd edition, 2001.