Probability & Statistics – Xác Suất & Thống Kê

Nguyễn Quản Bá Hồng*

Ngày 20 tháng 12 năm 2024

Tóm tắt nội dung

This text is a part of the series *Some Topics in Advanced STEM & Beyond*:
URL: https://nqbh.github.io/advanced_STEM/.
Latest version:

- *Probability & Statistics – Xác Suất & Thống Kê*.
  PDF: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/probability_statistics/NQBH_probability_statistics.pdf.
  TEX: URL: https://github.com/NQBH/advanced_STEM_beyond/blob/main/probability_statistics/NQBH_probability_statistics.tex.

Mục lục

*A Scientist & Creative Artist Wannabe. E-mail: nguyenquanbahong@gmail.com. Bến Tre City, Việt Nam.

1 Wikipedia

Relationship among AL, ML, & DL: DL $\subset$ ML $\subset$ AI.

1.1 Wikipedia/probability

"*Probability* is the branch of mathematics concerning events & numerical descriptions of how likely they are to occur. The probability of an event is a number $\in [0, 1]$; the larger the probability, the more likely an event is to occur. A simple example is the tossing of a fair (unbiased) coin. Since the coin is fair, the 2 outcomes ("heads" & "tails") are both equally probable; the probability of "heads" equals the probability of "tails"; & since no other outcomes are possible, the probability of either "heads" or "tails" is $\frac{1}{2}$ (which could also be written as 0.5 or 50%).

These concepts have been given an axiomatic mathematical formalization in *probability theory*, which is used widely in areas of study e.g. statistics, mathematics, science, finance, gambling, AI, ML, computer science, game theory, & philosophy to, e.g., draw inferences about the expected frequency of events. Probability theory is also used to describe the underlying mechanics & regularities of complex systems.

1.1.1 Interpretations

Main article: Wikipedia/probability interpretations. When dealing with random experiments – i.e., experiments that are random & well-defined – in a purely theoretical setting (like tossing a coin), probabilities can be numerically described by the number of desired outcomes, divided by the total number of all outcomes. This is referred to as *theoretical probability* (in contrast to empirical probability, dealing with probabilities in the context of real experiments). E.g., tossing a coin twice will yield "head-head", "head-tail", "tail-head", & "tail-tail" outcomes. The probability of getting an outcome of "head-head" is 1 out of 4 outcomes, or, in numerical terms, $\frac{1}{4}, 0.25, 25\%$. However, when it comes to practical application, there are 2 major competing categories of probability interpretations, whose adherents hold different views about the fundamental nature of probability:

- Objectivists assign numbers to describe some objective or physical state of affairs. The most popular version of objective probability is frequentist probability, which claims that the probability of a random event denotes the *relative frequency of occurrence* of an experiment's outcome when the experiment is repeated indefinitely. This interpretation considers probability to be the relative frequency "in the long run" of outcomes. A modification of this is propensity probability, which interprets probability as the tendency of some experiment to yield a certain outcome, even if it is performed only once.
- Subjectivists assign numbers per subjective probability, i.e., as a degree of belief. The degree of belief has been interpreted as "the price at which you would buy or sell a bet that pays 1 unit of utility of E, 0 if not E", although that interpretation is not universally agreed upon. The most popular version of subjective probability is Bayesian probability, which includes expert knowledge as well as experimental data to produce probabilities. The expert knowledge is represented by some (subjective) prior probability distribution. These data are incorporated in a likelihood function. The product of the prior & the likelihood, when normalized, results in a posterior probability distribution that incorporates all the information known to date. By Aumann's agreement theorem, Bayesian agents whose prior beliefs are similar will end up with similar posterior beliefs. However, sufficiently different priors can lead to different conclusions, regardless of how much information the agents share.

1.1.2 Etymology

The word *probability* derives from the Latin *probabilitas*, which can also mean "probity", a measure of the authority of a witness in a legal case in Europe, & often correlated with the witness's nobility. In a sense, this differs much from the modern meaning of *probability*, which in contrast is a measure of the weight of empirical evidence, & is arrived at from inductive reasoning & statistical inference.

1.1.3 History

Main article: Wikipedia/history of probability, Wikipedia/history of statistics. The scientific study of probability is a modern development of mathematics. Gambling shows that there has been an interest in quantifying the ideas of probability throughout history, but exact mathematical descriptions arose much later. There are reasons for the slow development of the mathematics

of probability. Whereas games of chance provided the impetus for the mathematical study of probability, fundamental issues are still obscured by superstitions. [...]

1.1.4 Theory

Main article: Wikipedia/probability theory. Like other theories, the theory of probability is a representation of its concepts in formal terms – i.e., in terms that can be considered separately from the meaning. These formal terms are manipulated by the rules of mathematics & logic, & any results are interpreted or translated back into the problem domain.

There have been at least 2 successful attempts to formalize probability, namely the Kolmogorov formulation & the Cox formulation. In Kolmogorov's formulation (see also probability space), sets are interpreted as events & probability as a measure on a class of sets. In Cox's theorem, probability is taken as a primitive (i.e., not further analyzed), & the emphasis is on constructing a consistent assignment of probability values to propositions. In both cases, the laws of probability are the same, except for technical details.

There are other methods for quantifying uncertainty, e.g., the Dempster–Shafer theory or possibility theory, but those are essentially different & not compatible with the usually-understood laws of probability.

1.1.5 Applications

Probability theory is applied in everyday life in risk assessment & modeling. The insurance industry & markets use actuarial science to determine pricing & make trading decisions. Governments apply probabilistic methods in environmental regulation, entitlement analysis, & financial regulation.

An example of the use of probability theory in equity trading is the effect of the perceived probability of any widespread Middle East conflict on oil prices, which have ripple effects in the economy as a whole. An assessment by a commodity trader that a war is more likely can send that commodity's prices up or down, & signals other traders of that opinion. Accordingly, the probabilities are neither assessed independently nor necessarily rationally. The theory of behavioral finance emerged to describe the effect of such groupthink on pricing, on policy, & on peace & conflict.

In addition to financial assessment, probability can be used to analyze trends in biology (e.g., disease spread) as well as ecology (e.g., biological Punnett squares). As with finance, risk assessment can be used as a statistical tool to calculate the likelihood of undesirable events occurring, & can assist with implementing protocols to avoid encountering such circumstances. Probability is used to design games of chance so that casinos can make a guaranteed profit, yet provide payouts to players that are frequent enough to encourage continued play.

Another significant application of probability theory in everyday life is reliability. Many consumer products, e.g., automobiles & consumer electronics, use reliability theory in product design to reduce the probability of failure. Failure probability may influence a manufacturer's decisions on a product's warranty.

The cache language model & other statistical language models that are used in natural language processing are also examples of applications of probability theory.

1.1.6 Mathematical treatment

[...]

1. **Independent events.**
2. **Mutually exclusive events.**
3. **Not (necessarily) mutually exclusive events.**
4. **Conditional probability.**
5. **Inverse probability.**
6. **Summary of probabilities.**

1.1.7 Relation to randomness & probability in quantum mechanics

Main article: Wikipedia/randomness. [...]" – Wikipedia/probability

1.2 Wikipedia/probability theory

"*Probability theory* or *probability calculus* is the branch of mathematics concerned with probability. Although there are several different probability interpretations, probability theory treats the concept in a rigorous mathematical manner by expressing it through a set of axioms. Typically these axioms formalize probability in terms of a probability space, which assigns a measure taking values $\in [0, 1]$, termed the probabilistic measure, to a set of outcomes called the sample space. Any specified subset of the sample space is called an event.

Central subjects in probability theory include discrete & continuous random variables, probability distributions, & stochastic processes (which provide mathematical abstractions of non-deterministic or uncertain processes or measured quantities that may either be single occurrences or evolve over time in a random fashion). Although it is not possible to perfectly predict random

events, much can be said about their behavior. 2 major results in probability theory describing such behavior are the law of large numbers & the central limit theorem.

As a mathematical foundation for statistics, probability theory is essential to many human activities that involve quantitative analysis of data. Methods of probability theory also apply to descriptions of complex systems given only partial knowledge of their state, as in statistical mechanics or sequential estimation. A great discovery of 20th-century physics was the probabilistic nature of physical phenomena at atomic scales, described in quantum mechanics.

1.2.1 History of probability

1.2.2 Treatment

1.2.3 Classical probability distributions

1.2.4 Convergence of random variables

" – Wikipedia/probability theory

1.3 Wikipedia/universal approximation theorem

"In the mathematical theory of artificial neural networks, *universal approximation theorems* are theorems of the following form:

Theorem 1 (Universal approximation)**.** *Given a family of neural networks, for each function f from a certain function space, there exists a sequence of neural networks $\phi_1, \phi_2, \ldots$ from the family, s.t. $\phi_n \to f$ as $n \to \infty$ according to some criterion, i.e., the family of neural networks is dense in the function space.*

The most popular version states that feedforward networks with non-polynomial activation functions are dense in the space of continuous functions between 2 Euclidean spaces, w.r.t. the compact convergence topology.

Universal approximation theorems are existence theorems: They simply state that there *exists* such a sequence $\phi_n \to f$, & do not provide any way to actually find such a sequence. They also do not guarantee any method, e.g., backpropagation, might actually find such a sequence. Any method for searching the space of neural networks, including backpropagation, might find a converging sequence, or not (i.e., the backpropagation might get stuck in a local optimum).

Universal approximation theorems are limit theorems: They simply state that for any f & a criteria of closeness $\epsilon > 0$, if there are *enough* neurons in a neural network, then there exists a neural network with that many neurons that does approximate f to within ϵ. There is no guarantee that any finite size, say, 10000 neurons, is enough.

1.3.1 Setup

Artificial neural networks are combinations of multiple simple mathematical functions that implement more complicated functions from (typically) real-valued vectors to real-valued vectors. The spaces of multivariate functions that can be implemented by a network are determined by the structure of the network, the set of simple functions, & its multiplicative parameters. A great deal of theoretical work has gone into characterizing these function spaces.

Most universal approximation theorems are in 1 of 2 classes. The 1st quantifies the approximation capabilities of neural networks with an arbitrary number of artificial neurons ("*arbitrary width*" case) & the 2nd focuses on the case with an arbitrary number of hidden layers, each containing a limited number of artificial neurons ("*arbitrary depth*" case). In addition to these 2 classes, there are also universal approximation theorems for neural networks with bounded number of hidden layers & a limited number of neurons in each layer ("*bounded depth & bounded width*" case).

1.3.2 History

Arbitrary width.

Arbitrary depth.

Bounded depth & bounded width.

Quantitative bounds.

Kolmogorov network.

Variants.

1.3.3 Arbitrary-width case

A spate of papers in the 1980s–1990s, from George Cybenko & Kurt Hornik etc., established several universal approximation theorems for arbitrary width & bounded depth. Most often quoted:

Theorem 2 (Universal approximation theorem)**.** *Let $C(X, \mathbb{R}^m)$ denote the set of continuous functions from a subset X of a Euclidean $\mathbb{R}^n$ space to a Euclidean space $\mathbb{R}^m$. Let $\sigma \in C(\mathbb{R}, \mathbb{R})$. Note that $(\sigma \circ \mathbf{x})_i = \sigma(x_i)$, so $\sigma \circ \mathbf{x}$ denotes σ applied to each component of $\mathbf{x}$. Then σ is not polynomial iff $\forall m, n \in \mathbb{N}$, compact $K \subseteq \mathbb{R}^n$, $f \in C(K, \mathbb{R}^m)$, $\varepsilon > 0$ there exist $k \in \mathbb{N}$, $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k, C \in \mathbb{R}^{m \times k}$ s.t. $\sup_{\mathbf{x} \in K} \|f(\mathbf{x}) - g(\mathbf{x})\| < \varepsilon$ where $g(\mathbf{x}) \coloneqq C \cdot (\sigma \circ (A \cdot \mathbf{x} + b))$.*

Also, certain non-continuous activation functions can be used to approximate a sigmoid function, which then allows the above theorem to apply to those functions. E.g., the step function works. In particular, this shows that a perceptron network with a single infinitely wide hidden layer can approximate arbitrary functions.

Such an f can also be approximated by a network of greater depth by using the same construction for the 1st layer & approximating the identity function with later layers. Proof sketch has not specified how one might use a ramp function to approximate arbitrary functions in $C_0(\mathbb{R}^n, \mathbb{R})$. A sketch of the proof is that one can 1st construct flat bump functions, intersect them to obtain spherical bump functions that approximate the Dirac delta function δ, then use those to approximate arbitrary functions in $C_0(\mathbb{R}^n, \mathbb{R})$. The original proofs, e.g., the one by Cybenko, use methods from functional analysis, including the Hahn–Banach & Riesz –Markov–Kakutani representation theorems.

Notice also that the neural network is only required to approximate within a compact set K. The proof does not describe how the function would be extrapolated outside of the region.

The problem with polynomials may be removed by allowing the outputs of the hidden layers to be multiplied together (the "pi-sigma $\pi\sigma$ networks"), yielding the generalization:

Theorem 3 (Universal approximation theorem for pi-sigma $\pi\sigma$ networks)**.** *With any nonconstant activation function, a 1-hidden-layer pi-sigma network is a universal approximator.*

1.3.4 Arbitrary-depth case

The "dual" versions of the theorem consider networks of bounded width & arbitrary depth. A variant of the universal approximation theorem was proved for the arbitrary depth case by Zhou Lu et al. in 2017. They showed that networks of width $n+4$ with ReLU activation functions can approximate any Lebesgue-integrable function on n-dimensional input space w.r.t. L^1 distance if network depth is allowed to grow. It was also shown that if the width was $\leq n$, this general expressive power to approximate any Lebesgue integrable function was lost. ReLU networks with width $n+1$ were sufficient to approximate any continuous function of n-dimensional input variables. The following refinement, specifies the optimal minimum width for which such an approximation is possible & is due to.

Theorem 4 (Universal approximation theorem (L_1 distance, ReLU activation, arbitrary depth, minimal width))**.** *For any Bochner–Lebesgue p-integrable function $f : \mathbb{R}^n \to \mathbb{R}^m$, & any $\varepsilon > 0$, there exists a fully connected ReLU network F of width exactly $d_m \coloneqq \max\{n+1, m\}$ satisfying $\int_{\mathbb{R}^n} \|f(\mathbf{x}) - F(\mathbf{x})\|^p \, \mathrm{d}\mathbf{x} < \varepsilon$. Moreover, there exists a function $f \in L^p(\mathbb{R}^n, \mathbb{R}^m)$ & some $\varepsilon > 0$, for which there is no fully connected ReLU network of width $< d_m \coloneqq \max\{n+1, m\}$ satisfying the above approximation bound.*

Remark 1. *If the activation is replaced by leaky-ReLU, & the input is restricted in a compact domain, then the exact minimum width is $d_m \coloneqq \max\{m, n, 2\}$.*

Quantitative refinement: In the case where $f : [0,1]^n \to \mathbb{R}$, i.e., $m = 1$, & σ is the ReLU activation function, the exact depth & width for a ReLU network to achieve ε error is also known. If, moreover, the target function f is smooth, then the required number of layer & their width can be exponentially smaller. Even if f is not smooth, the curse of dimensionality can be broken if f admits additional "compositional structure".

Together, the central result of

- Kidger, Patrick; Lyons, Terry (July 2020). *Universal Approximation with Deep Narrow Networks.* Conference on Learning Theory. arXiv:1905.08539

yields the following universal approximation theorem for networks with bounded width:

Theorem 5 (Universal approximation theorem (uniform non-affine activation, arbitrary depth, constrained width))**.** *Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^d$. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be any non-affine continuous function which is continuously differentiable at at least 1 point, with nonzero derivative at that point. Let $\mathcal{N}^{\sigma}_{d,D:d+D+2}$ denote the space of feed-forward neural networks with d input neurons, D output neurons, & an arbitrary number of hidden layers each with $d+D+2$ neurons, s.t. every hidden neuron has activation function σ & every output neuron has the identity as its activation function, with input layer ϕ & output layer ρ. Then given any $\varepsilon > 0$ & any $f \in C(\mathcal{X}, \mathbb{R}^D)$, there exists $\hat{f} \in \mathcal{N}^{\sigma}_{d,D:d+D+2}$ s.t. $\sup_{\mathbf{x} \in \mathcal{X}} \|\hat{f}(\mathbf{x}) - f(\mathbf{x})\| < \varepsilon$, i.e., $\mathcal{N}$ is dense in $C(\mathcal{X}; \mathbb{R}^D)$ w.r.t. the topology of uniform convergence.*

Quantitative refinement: The number of layers & width of each layer required to approximate f to ε precision known; moreover, the result hold true when $\mathcal{X}$ & $\mathbb{R}^D$ are replaced with any non-positively curved Riemannian manifold.

Certain necessary conditions for the bounded width, arbitrary depth case have been established, but there is still a gap between the known sufficient & necessary conditions.

1.3.5 Bounded depth & bounded width case

The 1st result on approximation capabilities of neural networks with bounded number of layers, each containing a limited number of artificial neurons was obtained by Maiorov & Pinkus. Their remarkable result revealed that such networks can be universal approximators & for achieving this property 2 hidden layers are enough.

Theorem 6 (Universal approximation theorem)**.** *There exists an activation σ which is analytic, strictly increasing & sigmoidal & has the following property: For any $f \in C([0,1]^d)$ & $\varepsilon > 0$ there exist constants $d_i, c_{ij}, \theta_{ij}, \gamma_i$, & vectors $\mathbf{w}^{ij} \in \mathbb{R}^d$ for which*

$$\left| f(\mathbf{x}) - \sum_{i=1}^{6d+3} d_{i\sigma} \left(\sum_{j=1}^{3d} c_{ij} \sigma(\mathbf{w}^{ij} \cdot \mathbf{x} - \theta_{ij}) - \gamma_i \right) \right| < \varepsilon, \ \forall \mathbf{x} = (x_1, \ldots, x_d) \in [0,1]^d.$$

This is an existence result: activation functions providing universal approximation property for bounded depth bounded width networks exist. Using certain algorithmic & computer programming techniques, Guliyev & Ismailov efficiently constructed such activation functions depending on a numerical parameter. The developed algorithm allows one to compute the activation functions at any point of the real axis instantly. For the algorithm & the corresponding computer code see

- Guliyev, Namig; Ismailov, Vugar (November 2018). *"Approximation capability of two hidden layer feedforward neural networks with fixed weights"*. Neurocomputing. 316: 262–269. arXiv:2101.09181. doi:10.1016/j.neucom.2018.07.075. S2CID 52285996.

The theoretical result can be formulated as follows.

Theorem 7 (Universal approximation theorem)**.** [...]

" – Wikipedia/universal approximation theorem

2 Probability – Xác Suất

Community – Cộng đồng. Andrey Nikolaevich Kolmogorov.
Resources – Tài nguyên.

1. Simon J. D. Prince. *Computer Vision: Models, Learning, & Inference.*
2. [Jay03]. E. T. Jaynes. *Probability Theory: The Logic of Science.*
3. [Kal21]. Olav Kallenberg. *Probability Theory & Stochastic Modeling*
4. [Kle20]. Achim Klenke. *Probability Theory: A Comprehensive Course.* [16 Amazon ratings][1744 citations]

   Amazon review. This popular textbook, now in a revised & expanded 3e, presents a comprehensive course in modern probability theory.

   Probability plays an increasingly important role not only in mathematics, but also in physics, biology, finance, & computer science, helping to understand phenomena e.g. magnetism, genetic diversity, & market volatility, & also to construct efficient algorithms. Starting with very basic, this textbook covers a wide variety of topics in probability, including many not usually found in introductory books, e.g.:

   - limit theorems for sums of random variables
   - martingales
   - percolation
   - Markov chains & electrical networks
   - construction of stochastic processes
   - Poisson point process & infinite divisibility
   - large deviation principles & statistical physics
   - Brownian motion
   - stochastic integrals & stochastic differential equations.

   Presentation is self-contained & mathematically rigorous, with material on probability theory interspersed with chapters on measure theory to better illustrate power of abstract concepts.

   3e has been carefully extended & includes new features, e.g. concise summaries at end of each sect & additional questions to encourage self-reflection, as well as updates to figures & computer simulations. With a wealth of examples & $>$ 290 exercises, as well as biographical details of key mathematicians, it will be of use to students & researchers in mathematics, statistics, physics, computer science, economics, & biology.

About the Author. ACHIM KLENKE is a professor at Johannes Gutenberg University in Mainz, Germany. He is known for his work on interacting particle systems, stochastic analysis, & branching processes, in particular for his pioneering work with LEONID MYTNIK on infinite rate mutually catalytic branching processes.

Preface to 3e. New in 3e: sections close with a short "takeaways" block where highlights of section are summarized sometimes on an informal level without full rigor. In some places "reflection" blocks have been added. They are of different levels of difficulty indicated by number of clubsuits.

Preface to 1e. This book is based on 2 4-hour courses on advanced probability theory have held in recent years at universities of Cologne & Mainz. Implicitly assumed: reader has a certain familiarity with basic concepts of probability theory, although formal framework will be fully developed in this book.

Aim: to present central objects & concepts of probability theory: random variables, independence, laws of large numbers, & central limit theorems, martingales, exchangeability, & infinite divisibility, Markov chains & Markov processes, as well as their connection with discrete potential theory, coupling, ergodic theory, Brownian motion & Itô integral (including stochastic differential equations), Poisson point process, percolation, & theory of large deviations.

Measure theory & integration are necessary prerequisites for a systematic probability theory. Develop it only to point to which it is needed for our purposes: construction of measures & integrals, Radon–Nikodym theorem & regular conditional distributions, convergence theorems for functions (Lebesgue) & measures (Prohorov), & construction of measures in product spaces. Chapters on measure theory do not come as a block at beginning (although they are written s.t. this would be possible; i.e., independent of probabilistic chapters) but are rather interlaced with probabilistic chapters that are designed to display power of abstract concepts in more intuitive world of probability theory. E.g., study percolation theory at point where we barely have measures, random variables, & independence; not even integral is needed. As only exception, *systematic* construction of independent random variables is deferred to Chap. 14. Although it is rather a matter of taste, hope: this setup helps to motivate reader throughout measure-theoretical chapters.

Those readers with a solid measure-theoretical education can skip in particular 1st & 4th chapters & might wish only to look up this or that.

In 1st 8 chapters, lay foundations that will be needed in all subsequent chapters. After that, there are 7 more or less independent parts, consisting of Chaps. 9–20, & 23. Chap. 21 on Brownian motion makes reference to Chaps. 9–15. Again, after that, 3 blocks consisting of Chaps. 22, 24, & 25, 26 can be read independently.

- 1. Basic Measure Theory. Introduce classes of sets that allow for a systematic treatment of events & random observations in framework of probability theory. Furthermore, construct measures, in particular probability measures, on such classes of sets. Define random variables as measurable maps.
  - 1.1. Classes of Sets. Let $\Omega \neq \emptyset$ be a nonempty set & let $\mathcal{A} \subset 2^\Omega$ (set of all subsets of Ω) be a class of subsets of Ω. Later, Ω will be interpreted as space of elementary events & $\mathcal{A}$ will be system of observable events. Introduce names for classes of subsets of Ω that are stable under certain set operations & establish simple relations between such classes.

    Definition 1. *A class of sets $\mathcal{A}$ is called:*

    $\cap$-closed (closed under intersections) or a π-system if $A \cap B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$.

    σ-$\cap$-closed (closed under countable intersections) if $\bigcap_{n=1}^\infty A_n \in \mathcal{A}$ for any choice of countably many sets $A_1, A_2, \ldots \in \mathcal{A}$.

    $\cup$-closed (closed under unions) if $A \cup B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$.

    σ-$\cup$-closed (closed under countable unions) if $\bigcup_{n=1}^\infty A_n \in \mathcal{A}$ for any choice of countably many sets $A_1, A_2, \ldots \in \mathcal{A}$.

    $\setminus$-closed (closed under differences) if $A \setminus B \in \mathcal{A}$ whenever $A, B \in \mathcal{A}$

    closed under complements if $A^c := \Omega \setminus A \in \mathcal{A}$ for any set $A \in \mathcal{A}$.

    Definition 2 (σ-algebra)**.** *A class of sets $\mathcal{A} \subset 2^\Omega$ is called a* σ-algebra *if it fulfills 3 conditions:*

    (i) $\Omega \in \mathcal{A}$.

    (ii) $\mathcal{A}$ is closed under complements.

    (iii) $\mathcal{A}$ is closed under countable unions.

    Sometimes a σ-algebra is also named a σ-field. Can define probabilities on σ-algebras in a consistent way. Hence these are natural classes of sets to be considered as *events* in probability theory.

    Immediate consequences of de Morgan's rule:

    Theorem 8. *If $\mathcal{A}$ is closed under complements, then have equivalences: $\mathcal{A}$ is $\cap$-closed $\Leftrightarrow$ $\mathcal{A}$ is $\cup$-closed, $\mathcal{A}$ is σ-$\cap$-closed $\Leftrightarrow$ $\mathcal{A}$ is σ-$\cup$-closed.*

    Theorem 9. *Assume $\mathcal{A}$ is $\setminus$-closed.*

    (i) $\mathcal{A}$ is $\cap$-closed.

    (ii) If in addition $\mathcal{A}$ is σ-$\cup$-closed, then $\mathcal{A}$ is σ-$\cap$-closed.

    (iii) Any countable (resp., finite) union of sets in $\mathcal{A}$ can be expressed as a countable (resp., finite) disjoint union of sets in $\mathcal{A}$.

    Definition 3. *A class of sets $\mathcal{A} \subset 2^\Omega$ is called an* algebra *if 3 conditions are fulfilled:*

    (i) $\Omega \in \mathcal{A}$.

    (ii) $\mathcal{A}$ is $\setminus$-closed.

    (iii) $\mathcal{A}$ is $\cup$-closed.

If $\mathcal{A}$ is an algebra, then obviously $\emptyset = \Omega\backslash\Omega \in \mathcal{A}$. However, in general, this property is weaker than (i) in Def. 1.6.

Theorem 10. *A class of sets $\mathcal{A} \subset 2^\Omega$ is an algebra iff 3 properties hold:*
(i) $\Omega \in \mathcal{A}$.
(ii) $\mathcal{A}$ is closed under complements.
(iii) $\mathcal{A}$ is closed under intersections.

Definition 4 (Ring)**.** *A class of sets $\mathcal{A} \in 2^\Omega$ is called a* ring *if 3 conditions hold:*
(i) $\emptyset \in \mathcal{A}$.
(ii) $\mathcal{A}$ is $\backslash$-closed.
(iii) $\mathcal{A}$ is $\cup$-closed.

A ring is called a *σ-ring* if it is also σ-$\cup$-closed.

Definition 5 (Semiring)**.** *A class of sets $\mathcal{A} \subset 2^\Omega$ is called a* semiring *if*
(i) $\emptyset \in \mathcal{A}$,
(ii) for any 2 sets $A, B \in \mathcal{A}$, difference set $B\backslash A$ is a finite union of mutually disjoint sets in $\mathcal{A}$,
(iii) $\mathcal{A}$ is $\cap$-closed.

Definition 6 ((Dynkin's) λ-system)**.** *A class of sets $\mathcal{A} \subset 2^\Omega$ is called a* λ-system (or Dynkin's λ-system) *if*
(i) $\Omega \in \mathcal{A}$,
(ii) for any 2 sets $A, B \in \mathcal{A}$ with $A \subset B$, difference set $B\backslash A \in \mathcal{A}$,
(iii) $\biguplus_{n=1}^\infty A_n \in \mathcal{A}$ for any choice of countably many pairwise disjoint sets $A_1, A_2, \ldots \in \mathcal{A}$.

Example 1.
(i) For any nonempty set Ω, classes $\mathcal{A} = \{\emptyset, \Omega\}$ & $\mathcal{A} = 2^\Omega$ are trivial examples of algebras, σ-algebras & λ-systems. On the other hand, $\mathcal{A} = \{\emptyset\}$ & $\mathcal{A} = 2^\Omega$ are trivial examples of semirings, rings, & σ-rings.
(ii) Let $\Omega = \mathbb{R}$. Then $\mathcal{A} = \{A \subset \mathbb{R} : A \text{ is countable}\}$ is a σ-ring.
(iii) $\mathcal{A} = \{(a, b] : a, b \in \mathbb{R}, a \leq b\}$ is a semiring on $\Omega = \mathbb{R}$ (but is not a ring).
(iv) Class of finite unions of bounded intervals is a ring on $\Omega = \mathbb{R}$ (but is not an algebra).
(v) Class of finite unions of arbitrary (also unbounded) intervals is an algebra on $\Omega = \mathbb{R}$ (but is not a σ-algebra). [...]
(ix) Every σ-algebra is a λ-system.

Theorem 11 (Relations between classes of sets)**.**
(i) Every σ-algebra also is a λ-system, an algebra & a σ-ring.
(ii) Every σ-ring is a rng, & every ring is a semiring.
(iii) Every algebra is a ring. An algebra on a finite set Ω is a σ-algebra.

- 1.2. Set Functions.
- 1.3. Measure Extension Theorem.
- 1.4. Measurable Maps.
- 1.5. Random Variables.

- 2. Independence.
  - 2.1. Independence of Events.
  - 2.2. Independent Random Variables.
  - 2.3. Kolmogorov's 0–1 Law.
  - 2.4. Example: Percolation.

- 3. Generating Functions. A fundamental principle of mathematics to map a class of objects that are of interest into a class of objects where computations are easier. This map can be 1-1, as with linear maps & matrices, or it may map only some properties uniquely, as with matrices & determinants.

  In probability theory, in 2nd category fall quantities e.g. median, mean, & variance of random variables. In 1st category, have characteristic functions, Laplace transforms, & probability generating functions. These are useful mostly because addition of independent random variables leads to multiplication of transforms. Before introduce characteristic functions (& Laplace transforms) later, want to illustrate basic idea with probability generating functions that are designed for $\mathbb{N}_0$-valued random variables.

  In Sect. 3.1, give basic defs & derive simple properties. Sects. 3.2–3.3 are denoted to 2 applications: Poisson approximation theorem & a simple investigation of Galton–Watson branching processes.

  - 3.1. Def & Examples.

    Definition 7 (Probability generating function)**.** *Let X be an $\mathbb{N}_0$-valued random variable. The (probability)* generating function *(p.g.f.) of $\mathbf{P}_X$ (or, loosely speaking, of X) is map $\psi_{\mathbf{P}_X} = \psi_X$ defined by (with understanding that $0^0 = 1$)*

$$\psi_X : [0,1] \to [0,1],\ z \mapsto \sum_{n=0}^\infty \mathbf{P}[X = n] z^n. \tag{1}$$

Theorem 12.
(i) ψ_X is continuous on $[0,1]$ & infinitely often continuously differentiable on $(0,1)$. For $n \in \mathbb{N}$, the nth derivative $\psi_X^{(n)}$ fulfills

$$\lim_{z \uparrow 1} \psi_X^{(n)}(z) = \sum_{n=0}^{\infty} \mathbf{P}[X = k] \cdot k(k-1) \cdots (k-n+1), \tag{2}$$

where both sides can equal ∞.
(ii) Distribution $\mathbf{P}_X$ of X is uniquely determined by ψ_X.
(iii) For any $r \in (0,1)$, ψ_X is uniquely determined by countably many values $\psi_X(x_i), x_i \in [0,r], i \in \mathbb{N}$. If series in (1) converges for some $z > 1$, then statement is also true for any $r \in (0,r)$ & have

$$\lim_{x \uparrow 1} \psi_X^{(n)}(x) = \psi_X^{(n)}(1) < \infty, \ \forall n \in \mathbb{N}.$$

In this case, ψ_X is uniquely determined by derivatives $\psi_X^{(n)}(1)$, $n \in \mathbb{N}$.
Statements follow from elementary theory of power series. For (iii), see [Rud76, Thm. 8.5].

Problem 1. *Come up with an example for X s.t. series in (3.1) does not converge for any $z > 1$ but $\lim_{x \uparrow 1} \psi_X'(x)$ exists & is finite.*

Theorem 13 (Multiplicativity of generating functions)**.** *If $X_1, \ldots, X_n$ are independent & $\mathbb{N}_0$-valued random variables, then $\psi_{X_1 + \cdots + X_n} = \prod_{i=1}^n \psi_{X_i}$.*

- 3.2. Poisson Approximation.
- 3.3. Branching Processes.

- 4. Integral.
  - 4.1. Construction & Simple Properties.
  - 4.2. Monotone Convergence & Fatou's Lemma.
  - 4.3. Lebesgue Integral vs. Riemann Integral.
- 5. Moments & Laws of Large Numbers. Most important characteristic quantities of random variables are median, expectation, & variance. For large n, expectation describes typical approximate value of arithmetic mean $\frac{\sum_{i=1}^n X_i}{n}$ of i.i.d. random variables (law of large numbers). In Chap. 15, see how variance determines size of typical deviations of arithmetic mean from expectation.
  - 5.1. Moments. $(\Omega, \mathcal{A}, \mathbf{P})$: a probability space.

Definition 8. *Let X be a real-valued random variable.*
(i) If $X \in L^1(\mathbf{P})$, then X is called integrable *& call $\mathbf{E}[X] \coloneqq \int X \, \mathrm{d}\mathbf{P}$* expectation *or* mean *of X. If $\mathbf{E}[X] = 0$, then X is called* centered*. More generally, also write $\mathbf{E}[X] = \int X \, \mathrm{d}\mathbf{P}$ if only X^- or X^+ is integrable.*
(ii) If $n \in \mathbb{N}, X \in L^n(\mathbf{P})$, then quantities

$$m_k \coloneqq \mathbf{E}[X^k], \ M_k \coloneqq \mathbf{E}[|X|^k], \ \forall k = 1, \ldots, n, \tag{3}$$

*are called k*th moments *& k*th absolute moments*, resp., of X.*
(iii) If $X \in L^2(\mathbf{P})$, then X is called square integrable *& $\mathrm{Var}[X] \coloneqq \mathbf{E}[X^2] - E[X]^2$ is* variance *of X. Number $\sigma \coloneqq \sqrt{\mathrm{Var}[X]}$ is called* standard deviation *of X. Formally, sometimes write $\mathrm{Var}[X] = \infty$ if $\mathbf{E}[X^2] = \infty$.*
(iv) If $X, Y \in L^2(\mathbf{P})$, then define covariance *of X, Y by $\mathrm{Cov}[X,Y] \coloneqq \mathbf{E}[(X - \mathbf{E}[X])(Y - \mathbf{E}[Y])]$. X, Y are called* uncorrelated *if $\mathrm{Cov}[X,Y] = 0$ &* correlated *otherwise.*
Def (ii) is sensible since, $X \in L^n(\mathbf{P})$ implies that $M_k < \infty$, $\forall k = 1, \ldots, n$. If $X, Y \in L^2(\mathbf{P})$, then $XY \in L^1(\mathbf{P})$ since $|XY| \leq X^2 + Y^2$. hence def (iv) makes sense & have $\mathrm{Cov}[X,Y] = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$. In particular, $\mathrm{Var}[X] = \mathrm{Cov}[X,X]$. Collect most important rules of expectations in a theorem. All of these properties are direct consequences of corresponding properties of integral.

Theorem 14 (Rules for expectations)**.** *Let $X, Y, X_n, Z_n, n \in \mathbb{N}$, be real integrable random variables on $(\Omega, \mathcal{A}, \mathbf{P})$.*
(i) If $\mathbf{P}_X = \mathbf{P}_Y$, then $\mathbf{E}[X] = \mathbf{E}[Y]$.
(ii) (Linearity) Let $c \in \mathbb{R}$. Then $cX \in L^1(\mathbf{P})$, $X + Y \in L^1(\mathbf{P})$ as well as $\mathbf{E}[cX] = c\mathbf{E}[X]$, $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$.
(iii) If $X \geq 0$ almost surely, then $\mathbf{E}[X] = 0 \Leftrightarrow X = 0$ almost surely.
(iv) (Monotonicity) If $X \leq Y$ almost surely, then $\mathbf{E}[X] \leq \mathbf{E}[Y]$ with equality iff $X = Y$ almost surely.
(v) (Triangle inequality) $|\mathbf{E}[X]| \leq \mathbf{E}[|X|]$.
(vi) If $X_n \geq 0$ almost surely $\forall n \in \mathbb{N}$, then $\mathbf{E}\left[\sum_{n=1}^{\infty} X_n\right] = \sum_{n=1}^{\infty} \mathbf{E}[X_n]$.
(vii) If $Z_n \uparrow Z$ for some Z, then $\mathbf{E}[Z] = \lim_{n \to \infty} \mathbf{E}[Z_n] \in (-\infty, \infty]$.
Again probability theory comes into play when independence enters stage, i.e., when we exit realm of linear integration theory.

Theorem 15 (Independent $L(\mathbf{P})$-random variables are uncorrelated)**.** *Let $X, Y \in L^1(\mathbf{P})$ be independent. Then $(XY) \in L^1(\mathbf{P}), \mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. In particular, independent square integrable random variables are uncorrelated.*

Theorem 16 (Wald's identity)**.** *Let $T, X_1, X_2, \dots$ be independent real random variables ion $L^1(\mathbf{P})$. Let $\mathbf{P}[T \in \mathbb{N}_0] = 1$ & assume $X_1, X_2, \dots$ are identically distributed. Define $S_T \coloneqq \sum_{i=1}^{T} X_i$. Then $S_T \in L^1(\mathbf{P}), \mathbf{E}[S_T] = \mathbf{E}[T]\mathbf{E}[X_1]$.*

Collect some basic properties of variance:

Theorem 17. *Let $X \in L^2(\mathbf{P})$. Then:*

(i) $\mathrm{Var}[X] = \mathbf{E}[(X - \mathbf{E}[X])^2] \geq 0$.

(ii) $\mathrm{Var}[X] = 0 \Leftrightarrow X = \mathbf{E}[X]$ *almost surely.*

(iii) Map $f : \mathbb{R} \to \mathbb{R}, x \mapsto \mathbf{E}[(X-x)^2]$ is minimal at $x_0 = \mathbf{E}[X]$ with $f(\mathbf{E}[X]) = \mathrm{Var}[X]$.

Theorem 18. *Map $\mathrm{Cov} : L^2(\mathbf{P}) \times L^2(\mathbf{P}) \to \mathbb{R}$ is a positive semidefinite symmetric bilinear form & $\mathrm{Cov}[X, Y] = 0$ if Y is almost surely constant. Detailed version of this concise statement is: Let $X_1, \dots, X_m, Y_1, \dots, Y_n \in L^2(\mathbf{P})$ & $\alpha_1, \dots, \alpha_m, \beta_1, \dots, \beta_n \in \mathbb{R}$ as well as $d, e \in \mathbb{R}$. Then*

$$\mathrm{Cov}\left[d + \sum_{i=1}^{m} \alpha_i X_i, e + \sum_{j=1}^{n} \beta_j Y_j\right] = \sum_{i,j} \alpha_i \beta_j \mathrm{Cov}[X_i, Y_j]. \tag{4}$$

In particular, $\mathrm{Var}[\alpha X] = \alpha^2 \mathrm{Var}[X]$ *&* Bienaymé formula *holds*

$$\mathrm{Var}\left[\sum_{i=1}^{m} X_i\right] + \sum_{i,j=1, i \neq j}^{m} \mathrm{Cov}[X_i, X_j]. \tag{5}$$

For uncorrelated $X_1, \dots, X_m$, have $\mathrm{Var}[\sum_{i=1}^m X_i] = \sum_{i=1}^m \mathrm{Var}[X_i]$.

Theorem 19 (Cauchy–Schwarz inequality)**.** *If $X, Y \in L^2(\mathbf{P})$, then $(\mathrm{Cov}[X, Y])^2 \leq \mathrm{Var}[X]\mathrm{Var}[Y]$. Equality holds iff there are $a, b, c \in \mathbb{R}$ with $|a| + |b| + |c| > 0$ & s.t. $aX + bY + c = 0$ a.s.*

Example 2.

(i) Let $p \in [0, 1], X \sim \mathrm{Ber}_p$. Then $\mathbf{E}[X^2] = \mathbf{E}[X] = \mathbf{P}[X = 1] = p$ & thus $\mathrm{Var}[X] = p(1-p)$.

(ii) Let $n \in \mathbb{N}, p \in [0, 1]$ & X be binomially distributed, $X \sim b_{n,p}$. Then $\mathbf{E}[X] = \cdots = np$, $\mathbf{E}[X(X-1)] = n(n-1)p^2$, hence $\mathbf{E}[X^2] = \mathbf{E}[X(X-1)] + \mathbf{E}[X] = n^2p^2 + np(1-p)$ & thus $\mathrm{Var}[X] = np(1-p)$. Statement can be derived more simply than by direct computation if make use of fact: $b_{n,p} = b_{1,p}^{*n}$ (see Example 3.4(ii)). I.e., see Thm. 2.31, $\mathbf{P}_X = \mathbf{P}_{Y_1 + \cdots + Y_n}$, where $Y_1, \dots, Y_n$ are independent & $Y_i \sim \mathrm{Ber}_p$ for any $i = 1, \dots, n$. Hence $\mathbf{E}[X] = n\mathbf{E}[Y_1] = np, \mathrm{Var}[X] = n\mathrm{Var}[Y_1] = np(1-p)$.

(iii) Let $\mu \in \mathbb{R}, \sigma^2 > 0$, & let X be normally distributed, $X \sim \mathcal{N}_{\mu,\sigma^2}$. Then $\mathbf{E}[X] = \cdots = \mu$. Similarly, get $\mathrm{Var}[X] = \mathbf{E}[X^2] - \mu^2 = \cdots = \sigma^2$.

(iv) Let $\theta > 0$ & X be exponentially distributed, $X \sim \exp_\theta$. Then $\mathbf{E}[X] = \cdots = \frac{1}{\theta}$, $\mathrm{Var}[X] = \theta^{-2}$.

Theorem 20 (Blackwell–Girshick)**.** *Let $T, X_1, X_2, \dots$ be independent real random variables in $L^2(\mathbf{P})$. Let $\mathbf{P}[T \in \mathbb{N}_0] = 1$ & let $X_1, X_2, \dots$ be identically distributed. Define $S_T \coloneqq \sum_{i=1}^{T} X_i$. Then $S_T \in L^2(\mathbf{P})$ & $\mathrm{Var}[S_T] = \mathbf{E}[X_1]^2\mathrm{Var}[T] + \mathbf{E}[T]\mathrm{Var}[X_1]$.*

In its proof, where did use independence of T? Check if $\mathbf{E}[X_i] = 0$, $\forall i \in \mathbb{N}$, then instead of independence of T, enough to postulate $\{T \leq n\}$ is independent of $X_{n+1}, X_{n+2}, \dots$ $\forall n$.

Takeaways. Moments are important characteristics of probability distributions. Have seen a formula for 1st & 2nd moment of a sum of random variables, even if number of summands is random itself. Independent random variables are uncorrelated. In this case, formulas for 2nd moments of sums are particularly simple.

Problem 2. *Let X be a nonnegative random variable with finite 2nd moment. Use Cauchy–Schwarz inequality for X & $\mathbf{1}_{\{X>0\}}$ in order to show Paley–Zygmund inequality*

$$\mathbf{P}[X > 0] \geq \frac{\mathbf{E}[X]^2}{\mathbf{E}[X^2]}. \tag{6}$$

Problem 3. *Let X be an integrable real random variable whose distribution $\mathbf{P}_X$ has a density f (w.r.t. Lebesgue measure λ). Show $\mathbf{E}[X] = \int_{\mathbb{R}} x f(x) \lambda(\mathrm{d}x)$.*

Problem 4. *Let $X \sim \beta_{r,s}$ be a Beta-distributed random variable with parameters $r, s > 0$. Show $\mathbf{E}[X^n] = \prod_{k=0}^{n-1} \frac{r+k}{r+s+k}$, $\forall n \in \mathbb{N}$.*

Problem 5. *Let $X_1, X_2, \dots$ be i.i.d. nonnegative random variables. By Borel–Cantelli lemma, show: (a)* $\limsup_{n\to\infty} \frac{1}{n} X_n$ *... (b) For any $c \in (0, 1)$, $\sum_{n=1}^{\infty} e^{X_n} c^n$ ...*

◦ 5.2. Weak Law of Large Numbers.

Theorem 21 (Markov inequality, Chebyshev inequality)**.** *Let X be a real random variable & let $f : [0, \infty) \to [0, \infty)$ be monotone increasing. Then for any $\varepsilon > 0$ with $f(\varepsilon) > 0$,* Markov inequality *holds*

$$\mathbf{P}[|X| \geq \varepsilon] \leq \frac{\mathbf{E}[f(|X|)]}{f(\varepsilon)}. \tag{7}$$

In special case $f(x) = x^2$, get $\mathbf{P}[|X| \geq \varepsilon] \leq \varepsilon^{-2}\mathbf{E}[X^2]$. In particular, if $X \in L^2(\mathbf{P})$, Chebyshev inequality *holds:*

$$\mathbf{P}[|X - \mathbf{E}[X]| \geq \varepsilon] \leq \varepsilon^{-2}\mathrm{Var}[X]. \tag{8}$$

- 5.3. Strong Law of Large Numbers.
- 5.4. Speed of Convergence in Strong LLN.
- 5.5 Poisson Process.

- 6. Convergence Theorems.
  - 6.1. Almost Sure & Measure Convergence.
  - 6.2. Uniform Integrability.
  - 6.3. Exchanging Integral & Differentiation.
- 7. L^p-Spaces & Radon–Nikodym Theorem.
  - 7.1. Defs.
  - 7.2. Inequalities & Fischer–Riesz Theorem.
  - 7.3. Hilbert Spaces.
  - 7.4. Lebesgue's Decomposition Theorem.
  - 7.5. Supplement: Signed Measures.
  - 7.6. Supplement: Dual Spaces.
- 8. Conditional Expectations.
  - 8.1. Elementary Conditional Probabilities.
  - 8.2. Conditional Expectations.
  - 8.2. Regular Conditional Distribution.
- 9. Martingales.
  - 9.1. Processes, Filtrations, Stopping Times.
  - 9.2. Martingales.
  - 9.3. Discrete Stochastic Integral.
  - 9.4. Discrete Martingale Representation Theorem & CRR Model.
- 10. Optional Sampling Theorems.
  - 10.1. Doob Decomposition & Square Variation.
  - 10.2. Optional Sampling & Optional Stopping.
  - 10.3. Uniform Integrability & Optional Sampling.
- 11. Martingale Convergence Theorems & Their Applications.
  - 11.1. Doob's Inequality.
  - 11.2. Martingale Convergence Theorems.
  - 11.3. Example: Branching Process.
- 12. Backwards Martingales & Exchangeability.
  - 12.1. Exchangeable Families of Random Variables.
  - 12.2. Backwards Martingales.
  - 12.3. De Finetti's Theorem.
- 13. Convergence of Measures.
  - 13.1. A Topology Primer.
  - 13.2. Weak & Vague Convergence.
  - 13.3. Prohorov's Theorem.
  - 13.4. Application: A Fresh Look at de Finetti's Theorem.
- 14. Probability Measures on Product Spaces.
  - 14.1. Product Spaces.
  - 14.2. Finite Products & Transition Kernels.
  - 14.3. Kolmogorov's Extension Theorem.
  - 14.4. Markov Semigroups.
- 15. Characteristic Functions & Central Limit Theorem.
  - 15.1. Separating Classes of Functions.
  - 15.2. Characteristic Functions: Examples.
  - 15.3. Lévy's Continuity Theorem.
  - 15.4. Characteristic Functions & Moments.
  - 15.5. Central Limit Theorem.
  - 15.6. Multidimensional Central Limit Theorem.

- 16. Infinitely Divisible Distributions.
  - 16.1. Lévy–Khinchin Formula.
  - 16.2. Stable Distributions.
- 17. Markov Chains.
  - 17.1. Defs & Construction.
  - 17.2. Discrete Markov Chains: Examples.
  - 17.3. Discrete Markov Processes in Continuous Time.
  - 17.4. Discrete Markov Chains: Recurrence & Transience.
  - 17.5. Application: Recurrence & Transience of Random Walks.
  - 17.6. Invariant Distributions.
  - 17.7. Stochastic Ordering & Coupling.
- 18. Convergence of Markov Chains.
  - 18.1. Periodicity of Markov Chains.
  - 18.2. Coupling & Convergence Theorem.
  - 18.3. Markov Chain Monte Carlo Method.
  - 18.4. Speed of Convergence.
- 19. Markov Chains & Electrical Networks.
  - 19.1. Harmonic Functions.
  - 19.2. Reversible Markov Chains.
  - 19.3. Finite Electrical Networks.
  - 19.4. Recurrence & Transience.
  - 19.5. Network Reduction.
  - 19.6. Random Walk in a Random Environment.
- 20. Ergodic Theory.
  - 20.1. Defs.
  - 20.2. Ergodic Theorems.
  - 20.3. Examples.
  - 20.4. Application: Recurrence of Random Walks.
  - 20.5. Mixing.
  - 20.6. Entropy.
- 21. Brownian Motion.
  - 21.1. Continuous Versions.
  - 21.2. Construction & Path Properties.
  - 21.3. Strong Markov Property.
  - 21.4. Supplement: Feller Processes.
  - 21.5. Construction via L^2-Approximation.
  - 21.6. Space $C([0,\infty))$.
  - 21.7. Convergence of Probability Measures on $C([0,\infty))$.
  - 21.8. Donsker's Theorem.
  - 21.9. Pathwise Convergence of Branching Processes.
  - 21.10. Square Variation & Local Martingales.
- 22. Law of Iterated Logarithm.
  - 22.1. Iterated Logarithm for Brownian Motion.
  - 22.2. Skorohod's Embedding Theorem.
  - 22.3. Hartman–Wintner Theorem.
- 23. Large Deviations.
  - 23.1. Cramér's Theorem.
  - 23.2. Large Deviations Principle.
  - 23.3. Sanov's Theorem.
  - 23.4. Varadhan's Lemma & Free Energy.
- 24. Poisson Point Process.
  - 24.1. Random Measures.
  - 24.2. Properties of Poisson Point Process.

- 24.3. Poisson–Dirichlet Distribution.

- 25. Itô Integral.
  - 25.1. Iô Integral w.r.t. Brownian Motion.
  - 25.2. Iô Integral w.r.t. Diffusions.
  - 25.3. Itô Formula.
  - 25.4. Dirichlet Problem & Brownian Motion.
  - 25.5. Recurrence & Transience of Brownian Motion.

- 26. Stochastic Differential Equations. Stochastic differential equations describe time evolution of certain continuous Markov processes with values in $\mathbb{R}^n$. In contrast wih classical differential equations, in addition to derivative of function, there is a term that describes random fluctuations that are coded as an Itô integral w.r.t. a Brownian motion. Depending on how seriously take concrete Brownian motion as driving force of noise, speak of strong & weak solutions. In Sect. 26.1, develop theory of strong solutions under Lipschitz conditions for coefficients. In Sect. 26.2, develop so-called *(local) martingale problem* as a method of establishing weak solutions. In Sect. 26.3, present some examples in which method of duality can be used to prove weak uniqueness.

  As stochastic differential equations are a very broad subject, & since things quickly become very technical, only excursively touch some of most important results, partly without proofs, & illustrate them with examples.

  Problem 6 (ODEs/PDEs vs. SDEs)**.** *Cf. ODEs/PDEs vs. SDEs. Difficulties? Numerics? Rigorous?*

  - 26.1. Strong Solutions. Consider a SDE of type

$$\begin{cases} X_0 = \xi, \\ dX_t = \sigma(t, X_t)dW_t + b(t, X_t)dt, \end{cases} \tag{9}$$

    where $W = (W^1, \ldots, W^m)$: an m-dimensional Brownian motion, ξ: an $\mathbb{R}^n$-valued random variable with distribution μ that is independent of W, $\sigma(t,x) = (\sigma_{ij}(t,x))_{i,j=1}^{n,m}$: a real $n \times m$ matrix & $b(t,x) = (b_i(t,x))_{i=1}^n$: an n-dimensional vector. Assume maps $(t,x) \mapsto \sigma_{ij}(t,x), (t,x) \mapsto b_i(t,x)$ are measurable. By a solution of (9), understand a continuous adapted stochastic process X with values in $\mathbb{R}^n$ that satisfies integral equation (26.2)

$$X_t = \xi + \int_0^t \sigma(s, X_s)\, \mathrm{d}W_s + \int_0^t b(s, X_s)\, \mathrm{d}s \ \mathbf{P}\text{-a.s. } \forall t \geq 0. \tag{10}$$

    Written in full, this is

$$X_t^i = \xi^i + \sum_{j=1}^m \int_0^t \sigma_{ij}(s, X_s)\, \mathrm{d}W_s^j + \int_0^t b_i(s, X_s)\, \mathrm{d}s, \ \forall i = 1, \ldots, n. \tag{11}$$

    Problem arises: To which filtration $\mathbb{F}$ do we wish X to be adapted? Should it be filtration that is generated by ξ & W, or do we allow $\mathbb{F}$ to be larger? Already for ODEs, depending on equation, uniqueness of solution may fail (although existence is usually not a problem); e.g., $f' = |f|^{\frac{1}{3}}$. If $\mathbb{F}$ is larger than filtration generated by W, then can define further random variables that select one out of a variety of possible solutions. Thus have more possibilities for solutions than if $\mathbb{F} = \sigma(W)$. Indeed, in some situations for existence of a solution, necessary to allow a larger filtration. Roughly speaking, X is a strong solution of (9) if (26.2) holds & if X is adapted to $\mathbb{F} = \sigma(W)$. On other hand, X is a weak solution if X is adapted to a larger filtration $\mathbb{F}$ w.r.t. which W is still a martingale. Weak solutions will be dealt with in Sect. 26.2.
  - 26.2. Weak Solutions & Martingale Problem.
  - 26.3. Weak Uniqueness via Duality.

5. [Var01]. S. R. S. Varadhan. *Probability Theory.*

3 Statistics – Thống Kê

Community – Cộng đồng.
Resources – Tài nguyên.

1.

4 Stochastic – Ngẫu Nhiên

Community – Cộng đồng. Caroline Geiersbach, Michael Hintermüller.
Resources – Tài nguyên.

1.

5 Data Science (DS)

Community – Cộng đồng.
Resources – Tài nguyên.

1.

5.1 Wikipedia/Data Science

"*Data science* is an interdisciplinary academic field that uses statistics, scientific computing, scientific methods, processing, scientific visualization, algorithms & systems to extract or extrapolate knowledge & insights from potentially noisy, structured, or unstructured data.

Data science also integrates domain knowledge from the underlying application domain (e.g., natural sciences, information technology, & medicine). Data science is multifaceted & can be described as a science, a research paradigm, a research method, a discipline, a workflow, & a profession.

Data science is "a concept to unify statistics, data analysis, informatics, & their related methods" to "understand & analyze actual phenomena" with data. It uses techniques & theories drawn from many fields within the context of mathematics, statistics, computer science, information science, & domain knowledge. However, data science is different from computer science & information science. Turing Award winner Jim Gray imagined data science as a "4th paradigm" of science (empirical, theoretical, computational, & now data-driven) & asserted that "everything about science is changing because of the impact of information technology" & the data deluge.

A *data scientist* is a professional who creates programming code & combines it with statistical knowledge to create insights from data.

5.1.1 Foundations

Data science is an interdisciplinary field focused on extracting knowledge from typically large data sets & applying the knowledge & insights from that data to solve problems in a wide range of application domains. The field encompasses preparing data for analysis, formulating data science problems, analyzing data, developing data-driven solutions, & presenting findings to inform high-level decisions in a broad range of application domains. As such, it incorporates skills from computer science, statistics, information science, mathematics, data visualization, information visualization, data sonification, data integration, graphic design, complex systems, communication & business. Statistician Nathan Yau, drawing on Ben Fry, also links data science to human-computer interaction: users should be able to intuitively control & explore data. In 2015, the American Statistical Association identified database management, statistics & machine learning, & distributed & parallel systems as the 3 emerging foundational professional communities.

Relationship to statistics. Many statisticians, including Nate Silver, have argued that data science is not a new field, but rather another name for statistics. Others argue that data science is distinct from statistics because it focuses on problems & techniques unique to digital data. Vasant Dhar writes that statistics emphasizes quantitative data & description. In contrast, data science deals with quantitative & qualitative data (e.g., from images, text, sensors, transactions, customer information, etc.) & emphasizes prediction & action. Andrew Gelman of Columbia University has described statistics as a non-essential part of data science.

Stanford professor David Donoho writes that data science is not distinguished from statistics by the size of datasets or use of computing & that many graduate programs misleadingly advertise their analytics & statistics training as the essence of a data-science program. He describes data science as an applied field growing out of traditional statistics.

5.1.2 Etymology

5.1.3 Data science & data analysis

Data science & data analysis are both important disciplines in the field of data management & analysis, but they differ in several key ways. While both fields involve working with data, data science is more of an interdisciplinary field that involves the application of statistical, computational, & machine learning methods to extract insights from data & make predictions, while data analysis is more focused on the examination & interpretation of data to identify patterns & trends.

Data analysis typically involves working with smaller, structured datasets to answer specific questions or solve specific problems. This can involve tasks such as data cleaning, data visualization, & exploratory data analysis to gain insights into the data & develop hypotheses about relationships between variables. Data analysts typically use statistical methods to test these hypotheses & draw conclusions from the data. E.g., a data analyst might analyze sales data to identify trends in customer behavior & make recommendations for marketing strategies.

Data science, on the other hand, is a more complex & iterative process that involves working with larger, more complex datasets that often require advanced computational & statistical methods to analyze. Data scientists often work with unstructured data such as text or images & use machine learning algorithms to build predictive models & make data-driven decisions. In addition to statistical analysis, data science often involves tasks such as data preprocessing, feature engineering, & model selection. E.g., a data scientist might develop a recommendation system for an e-commerce platform by analyzing user behavior patterns & using machine learning algorithms to predict user preferences.

While data analysis focuses on extracting insights from existing data, data science goes beyond that by incorporating the development & implementation of predictive models to make informed decisions. Data scientists are often responsible for collecting & cleaning data, selecting appropriate analytical techniques, & deploying models in real-world scenarios. They work at the intersection fo mathematics, computer science, & domain expertise to solve complex problems & uncover hidden patterns in large datasets.

Despite these differences, data science & data analysis are closely related fields & often require similar skills sets. Both fields require a solid foundation in statistics, programming, & data visualization, as well as the ability to communicate findings effectively to both technical & non-technical audiences. Both fields benefit from critical thinking & domain knowledge, as understanding the context & nuances of the data is essential for accurate analysis & modeling.

In summary, data analysis & data science are distinct yet interconnected disciplines within the broader field of data management & analysis. Data analysis focuses on extracting insights & drawing conclusions from structured data, while data science involves a more comprehensive approach that combines statistical analysis, computational methods, & machine learning to extract insights, build predictive models, & drive data-driven decision-making. Both fields use data to understand patterns, make informed decisions, & solve complex problems across various domains.

5.1.4 Cloud computing for data science

Cloud computing can offer access to large amounts of computational power & storage. In big data, where volumes of information are continually generated & processed, these platforms can be used to handle complex & resource-intensive analytical tasks.

Some distributed computing frameworks are designed to handle big data workloads. These frameworks can enable data scientists to process & analyze large datasets in parallel, which can reducing processing times.

5.1.5 Ethical consideration in data science

Data science involve collecting, processing, & analyzing data which often including personal & sensitive information. Ethical concerns include potential privacy violations, bias perpetuation, & negative societal impacts.

Machine learning models can amplify existing biases present in training data, leading to discriminatory or unfair outcomes." – Wikipedia/data science

6 Deep Learning (DL)

Community – Cộng đồng. Yann LeCun, Yoshua Bengio, Geoffrey Hinton.
Resources – Tài nguyên.

1. [LBH15]. Yann LeCun, Yoshua Bengio, Geoffrey Hinton. *Deep Learning.*

2. [NP23]. Phan-Minh Nguyen, Huy Tuan Pham. *A rigorous framework for the mean field limit of multilayer neural networks.*

Những năm gần đây, sự phát triển của các hệ thống tính toán cùng lượng dữ liệu khổng lồ được thu thập bởi các hãng công nghệ lớn đã giúp machine learning tiến thêm 1 bước dài. 1 lĩnh vực mới được ra đời được gọi là *học sâu* (deep learning, DL). Deep learning đã giúp máy tính thực thi những việc vào 10 năm trước tưởng chừng là không thể: phân loại cả ngàn vật thể khác nhau trong các bức ảnh, tự tạo chú thích cho ảnh, bắt chước giọng nói & chữ viết, giao tiếp với con người, chuyển đổi ngôn ngữ, hay thậm chí cả sáng tác văn thơ & âm nhạc." – [Tiệ25, p. 15]

Φ: Start with simple things – Luôn bắt đầu từ những điều đơn giản. Khi bắt tay vào giải quyết 1 bài toán ML hay bất cứ bài toán nào, nên bắt đầu từ các thuật toán đơn giản. Không phải chỉ có các thuật toán phức tạp mới có thể giải quyết được vấn đề. Các thuật toán phức tạp thường có yêu cầu cao về khả năng tính toán & đôi khi nhạy cảm với cách chọn tham số. Ngược lại, các thuật toán đơn giản giúp ta nhanh chóng có 1 bộ khung cho mỗi bài toán. Kết quả của các thuật toán đơn giản cũng mang lại cái nhìn sơ bộ về sự phức tạp của mỗi bài toán. Việc cải thiện kết quả sẽ được thực hiện dần ở các bước sau. " – [Tiệ25, p. 17]

Approach. Để giải quyết mỗi bài toán ML, cần chọn 1 mô hình phù hợp. Mô hình này được mô tả bởi bộ các tham số ta cần đi tìm. Thông thường, lượng tham số có thể lên tới hàng triệu & được tìm bằng cách giải 1 bài toán tối ưu. Khi viết về các thuật toán ML, VKTiệp sẽ bắt đầu từ các ý tưởng trực quan. Các ý tưởng này được mô hình hóa dưới dạng 1 bài toán tối ưu. Các suy luận toán học & ví dụ mẫu trên Python sẽ giúp hiểu rõ hơn về nguồn gốc, ý nghĩa, & cách sử dụng mỗi thuật toán. Xen kẽ giữa các thuật toán ML, trình bày các kỹ thuật tối ưu cơ bản, với hy vọng giúp hiểu rõ hơn bản chất của vấn đề.

Audiences. Cuốn sách được thực hiện hướng tới nhiều nhóm độc giả khác nhau. Nếu không thực sự muốn đi sâu vào phần toán, vẫn có thể tham khảo mã nguồn & cách sử dụng các thư viện. Nhưng để sử dụng các thư viện 1 cách hiệu quả, cũng cần hiểu nguồn gốc của mô hình & ý nghĩa của các tham số. Còn nếu thực sự muốn tìm hiểu nguồn gốc, ý nghĩa của các thuật toán, có thể học được nhiều điều từ cách xây dựng & tối ưu các mô hình.

Python. Python là 1 ngôn ngữ lập trình miễn phí, có thể được cài đặt dễ dàng trên các nền tảng hệ điều hành khác nhau. Có rất nhiều thư viện hỗ trợ ML cũng như DL trên Python. Có 2 thư viện Python chính thường được sử dụng là `numpy`, `scikit-learn`.

- `numpy` www.numpy.org là 1 thư viện phổ biến giúp xử lý các phép toán liên quan đến các mảng nhiều chiều, hỗ trợ các hàm gần gũi với đại số tuyến tính. Cách xử lý các mảng nhiều chiều.

- `scikit-learn/sklearn` scikit-learn.org: 1 thư viện chứa đầy đủ các thuật toán ML cơ bản & rất dễ sử dụng. Tài liệu của scikit-learn cũng là 1 nguồn tham khảo chất lượng cho MLer. Scikit-learn được dùng để kiểm chứng các suy luận toán học & các mô hình được xây dựng thông qua `numpy`.

Inevitability of mathematics in ML. Có rất nhiều thư viện giúp tạo ra các sản phẩm ML/DL mà không yêu cầu nhiều kiến thức toán. Hướng tới việc giúp hiểu bản chất toán học đằng sau mỗi mô hình trước khi áp dụng các thư viện sẵn có. Việc sử dụng thư viện + yêu cầu kiến thức nhất định về việc lựa chọn mô hình & điều chỉnh các tham số.

7 Machine Learning (ML)

Resources – Tài nguyên.

1. Andrew Ng. *Machine Learning Course* on Coursera.
2. Machine Learning Mastery: Making Developers Awesome at Machine Learning: `https://machinelearningmastery.com`.
   - Machine Learning Mastery/8 Inspirational Applications of Deep Learning.
3. Machine Learning cơ bản: `https://machinelearningcoban.com/`.
4. [Tiệ25]. Vũ Hữu Tiệp. *Machine Learning Cơ Bản*.
   Mã nguồn cuốn ebook "Machine Learning Cơ Bản": `https://github.com/tiepvupsu/ebookMLCB`.

Definition 9. *"Machine learning (ML) is a field of study in AI concerned with the development & study of statistical algorithms that can learn from data & generalize to unseen data, & thus perform tasks without explicit instructions. Quick progress in the fields of deep learning, beginning in 2010s, allowed neural networks to surpass many previous approaches in performance." – Wikipedia/machine learning*

Định nghĩa 1. *"Học máy (machine learning, ML) là 1 tập con của trí tuệ nhân tạo. Machine learning là 1 lĩnh vực nhỏ trong Khoa học Máy tính, có khả năng tự học hỏi dựa trên dữ liệu được đưa vào mà không cần phải được lập trình cụ thể: "Machine Learning is the subfield of computer science, that "gives computers the ability to learn without being explicitly programmed" – Wikipedia." – [Tiệ25, p. 15]*

"ML finds application in many fields, including natural language processing, computer vision, speech recognition, email filtering, agriculture, & medicine. The application of ML to business problems is known as predictive analysis.

Statistics & mathematical optimization/mathematical programming methods comprise the foundations of machine learning. Data mining is related field of study, focusing on exploratory data analysis (EDA) via unsupervised learning.

From a theoretical viewpoint, probably approximately correct (PAC) learning provides a framework for describing machine learning." – Wikipedia/machine learning

Relationships of ML to AI. As a scientific endeavor, machine learning grew out of the quest for AI. In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons & other models e.g. ADALINE that were later found to be reinventions of the generalized linear models of statistics. Probabilistic reasoning was also employed, especially in automated medical diagnosis. However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI & machine learning. Probabilistic systems were plagued by theoretical & practical problems of data acquisition & representation.

7.1 Wikipedia/Machine Learning

"*Machine learning* (ML) is a field of study in AI concerned with the development & study of statistical algorithms that can learn from data & generalize to unseen data, & thus perform tasks without explicit instructions. Quick progress in the field of deep learning, beginning in 2010s, allowed neutral networks to surpass many previous approaches in performance.

ML finds application in many fields, including natural language processing, computer vision, speed recognition, email filtering, agriculture, & medicine. The application of ML to business problems is known as predictive analytics.

Statistics & mathematical optimization/programming methods comprise the foundations of machine learning. Data mining is a related field of study, focusing on exploratory data analysis (EDA) via unsupervised learning.

From a theoretical viewpoint, probably approximately correct (PAC) learning provides a framework for describing machine learning.

7.1.1 History

7.1.2 Relationships to other fields

AI. As a scientific endeavor, machine learning grew out of the quest for AI. In the early days of AI as an academic discipline, some researchers were interested in having machines learn from data. They attempted to approach the problem with various symbolic methods, as well as what were then termed "neural networks"; these were mostly perceptrons & other models ADALINE

that were later found to be reinventions of the generalized linear models of statistics. Probablistic reasoning was also employed, especially in automated medical diagnosis.

However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI & machine learning. Probabilistic systems were plagued by theoretical & practical problems of data acquisition & representation. By 1980, expert systems had come to dominate AI, & statistics was out of favor. Work on symbolic/knowledge-based learning did continue within AI, leading to inductive logic programming (ILP), but the more statistical line of research was now outside the field of AI proper, in pattern recognition & information retrieval. Neural networks research had been abandoned by AI & computer science around the same time. This line, too, was continued outside the AI/CS field, as "connectionism", by researchers from other disciplines including John Hopfield, David Rumelhart, & Geoffrey Hinton. Their main success came in the mid-1980s with the reinvention of backpropagation.

Machine learning (ML), reorganized & recognized as its own field, started to flourish in the 1990s. The field changed its goal from achieving AI to tackling solvable problems of a practical nature. It shifted focus away from the symbolic approaches it had inherited from AI, & toward methods & models borrowed from statistics, fuzzy logic, & probability theory.

Data compression. Main: Wikipedia/data compression/ML. There is a close connection between ML & compression. A system that predicts the posterior probabilities of a sequence given its entire history can be used for optimal data compression (by using arithmetic coding on the output distribution). Conversely, an optimal compressor can be used for prediction (by finding the symbol that compresses best, given the previous history). This equivalence has been used as a justification for using data compression as a benchmark for "general intelligence".

An alternative view can show compression algorithms implicitly map strings into implicit feature space vectors, & compression-based similarity measures compute similarity within these feature spaces. For each compressor $C(\cdot)$ we define an associated vector space $\mathfrak{N}$ s.t. $C(\cdot)$ maps an input string x, corresponding to the vector norm $\|\widetilde{x}\|$. An exhaustive examination of the feature spaces underlying all compression algorithms is precluded by space; instead, feature vectors chooses to examine 3 representative lossless compression methods, LZW, LZ77, & PPM.

According to AIXI theory, a connection more directly explained in Hutter Prize, the best possible compression of x is the smallest possible software that generates x. E.g., in that model, a zip file's compressed size includes both the zip file & the unzipping software, since you cannot unzip it without both, but there may be an even smaller combined form.

Examples of AI-powered audio/video compression software include NVIDIA Maxine, AIVC. Examples of software that can perform AI-powered image compression include OpenCV, TensorFlow, MATLAB's Imaging Processing Toolbox (IPT) & High-Fidelity Generative Image Compression.

In unsupervised machine learning, k-mean clustering can be utilized to compress data by grouping similar data points into clusters. This technique simplifies handling extensive datasets that lack predefined labels & finds widespread use in fields such as image compression.

Data compression aims to reduce the size of data files, enhancing storage efficiency & speeding up data transmission. k-means clustering, an unsupervised machine learning algorithm, is employed to partition a dataset into a specified number of clusters, k, each represented by the centroid of its points. This process condenses extensive datasets into a more compact set of representative points. Particularly beneficial in image & signal processing, k-means clustering aids in data reduction by replacing groups of data points with their centroids, thereby preserving the core information of the original data while significantly decreasing the required storage space.

Large language models (LLMs) are also capable of lossless data compression, as demonstrated by DeepMind's research with the Chinchilla 70B model. Developed by DeepMind, Chinchilla 70B effectively compressed data, outperforming conventional methods such as Portable Network Graphics (PNG) for images & Free Lossless Audio Codec (FLAC) for audio. It achieved compression of image & audio data to 43.4% & 16.4% of their original sizes, respectively.

Data mining. ML & data mining often employ the same methods & overlap significantly, but while ML focuses on prediction, based on *known* properties learned from the training data, data mining focuses on the discovery of (previously) *unknown* properties in the data (this is the analysis step of knowledge discovery in databases). Data mining uses many ML methods, but with different goals; on the other hand, ML also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve learner accuracy. Much of the confusion between these 2 research communities (which do often have separate conferences & separate journals, ECML PKDD being a major exception) comes from the basic assumptions they work with: in ML, performance is usually evaluated w.r.t. the ability to *reproduce known* knowledge, while in knowledge discovery & data mining (KDD) the key task is the discovery of previously *unknown* knowledge. Evaluated w.r.t. known knowledge, an uninformed (unsupervised) method will easily be outperformed by other supervised methods, while in a typical KDD task, supervised methods cannot be used due to the unavailability of training data.

ML also has intimate ties to optimization: Many learning problems are formulated as minimization of some loss function on a training set of examples. Loss functions express the discrepancy between the predictions of the model being trained & the actual problem instances (e.g., in classification, one wants to assign a label to instances, & models are trained to correctly predict the preassigned labels of a set of examples).

Generalization. Characterizing the generalization of various learning algorithms is an active topic of current research, especially for deep learning algorithms.

Statistics. ML & statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics draws population inferences from a sample, while ML finds generalizable predictive patterns. According to Michael I. Jordan, the ideas of ML, from methodological principles to theoretical tools, have had a long pre-history in statistics. he also suggested the term data science as a placeholder to call the overall field.

Conventional statistical analyzes require the a priori selection of a model most suitable for the study data set. In addition, only significant or theoretically relevant variables based on previous experience are included for analysis. In contrast, ML is not built on a pre-structured model; rather, the data shape the model by detecting underlying patterns. The more variables (input) used to train the model, the more accurate the ultimate model will be.

Leo Breiman distinguished 2 statistical modeling paradigms: data model & algorithmic model, wherein "algorithm model" means more or less the ML algorithms like Random Forest.

Some statisticians have adopted methods from ML, leading to a combined field that they call *statistical learning*.

Statistical physics. Analytical & computational techniques derived from deep-rooted physics of disordered systems can be extended to large-scale problems, including ML, e.g., to analyze the weight space of deep neural networks. Statistical physics is thus finding applications in the area of medical diagnostics.

7.1.3 Theory

Main: Wikipedia/computational learning theory & Wikipedia/statistical learning theory. A core objective of a learner is to generalize from its experience. Generalization in this context is the ability of a learning machine to perform accurately on new, unseen examples/tasks after having experienced a learning data set. The training examples come from some generally unknown probability distribution (considered representative of the space of occurrences) & the learner has to build a general model about this space that enables it to produce sufficiently accurate predictions in new cases.

The computational analysis of ML algorithms & their performance is a branch of theoretical computer science known as computational learning theory via the Probably Approximately Correct Learning (PAC) model. Because training sets are finite & the future is uncertain, learning theory usually does not yield guarantees of the performance of algorithms. Instead, probabilistic bounds on the performance are quite common. The bias-variance decompositon is 1 way to quantify generalization error.

For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of he function underlying the data. If the hypothesis is less complex than the function, then the model has under fitted the data. If the complexity of the model is increased in response, then the training error decreases. But if the hypothesis is too complex, then the model is subject to overfitting & generalization will be poorer.

In addition to performance bounds, learning theorists study the time complexity & feasibility of learning. In computational learning theory, a computation is considered feasible if it can be done in polynomial time. There are 2 kinds of time complexity results: Positive results show that a certain class of functions can be learned in polynomial time. Negative results show that certain classes cannot be learned in polynomial time.

7.1.4 Approaches

ML approaches are traditionally divided into 3 broad categories, which corresponding to learning paradigms, depending on the nature of the "signal" or "feedback" available to the learning system:

- Supervised learning: The computer is presented with example inputs & their desired outputs, given by a "teacher", & the goal to learn a general rule that maps inputs to outputs.
- Unsupervised learning: No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means toward an end (feature learning).
- Reinforcement learning: A computer program interacts with a dynamic environment in which it must perform a certain goal (e.g. driving a vehicle or playing a game against an opponent). As it navigates its problem space, the program is provided feedback that's analogous to rewards, which it tries to maximize.

Although each algorithm has advantages & limitations, no single algorithm works for all problems.

Supervised learning.

Semi-supervised learning.

Reinforcement learning.

Dimensionality reduction.

Other types. " – Wikipedia/machine learning

8 Artificial Intelligence (AI)

Resources – Tài nguyên.

1. [BV14]. Lê Hoài Bắc, Tô Hoài Việt. *Cơ Sở Trí Tuệ Nhân Tạo.*

2. [Aou14]. Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence.*

3. [Aou19]. Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence – Chạy Dua Với Robot: Học Tập Thời Trí Tuệ Nhân Tạo.*

9 Miscellaneous

Tài liệu

[Aou14] Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence*. MIT Publisher, 2014, p. 187.

[Aou19] Joseph E. Aoun. *Robot-Proof: Higher Education in the Age of Artificial Intelligence – Chạy Dua Với Robot: Học Tập Thời Trí Tuệ Nhân Tạo*. Trịnh Huy Nam dịch. Nhà Xuất Bản Thế Giới, 2019, p. 241.

[BV14] Lê Hoài Bắc and Tô Hoài Việt. *Cơ Sở Trí Tuệ Nhân Tạo*. Nhà Xuất Bản Khoa Học & Kỹ Thuật, 2014, p. 229.

[Jay03] E. T. Jaynes. *Probability theory*. The logic of science, Edited and with a foreword by G. Larry Bretthorst. Cambridge University Press, Cambridge, 2003, pp. xxx+727. ISBN: 0-521-59271-2. DOI: 10.1017/CBO9780511790423. URL: https://doi.org/10.1017/CBO9780511790423.

[Kal21] Olav Kallenberg. *Foundations of modern probability*. Third. Vol. 99. Probability Theory and Stochastic Modelling. Springer, Cham, [2021] ©2021, p. 946. ISBN: 978-3-030-61871-1; 978-3-030-61870-4. DOI: 10.1007/978-3-030-61871-1. URL: https://doi.org/10.1007/978-3-030-61871-1.

[Kle20] Achim Klenke. *Probability theory—a comprehensive course*. Universitext. Third edition [of 2372119]. Springer, Cham, [2020] ©2020, pp. xiv+716. ISBN: 978-3-030-56402-5; 978-3-030-56401-8. DOI: 10.1007/978-3-030-56402-5. URL: https://doi.org/10.1007/978-3-030-56402-5.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep Learning". In: *Nature* 521 (2015), pp. 436–444. DOI: 10.1038/nature14539. URL: https://doi.org/10.1038/nature14539.

[NP23] Phan-Minh Nguyen and Huy Tuan Pham. "A rigorous framework for the mean field limit of multilayer neural networks". In: *Math. Stat. Learn.* 6.3-4 (2023), pp. 201–357. ISSN: 2520-2316. DOI: 10.4171/msl/42. URL: https://doi.org/10.4171/msl/42.

[Rud76] Walter Rudin. *Principles of mathematical analysis*. Third. International Series in Pure and Applied Mathematics. McGraw-Hill Book Co., New York-Auckland-Düsseldorf, 1976, pp. x+342.

[Tiệ25] Vũ Khắc Tiệp. *Machine Learning Cơ Bản*. 2025, p. 422.

[Var01] S. R. S. Varadhan. *Probability theory*. Vol. 7. Courant Lecture Notes in Mathematics. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, 2001, pp. viii+167. ISBN: 0-8218-2852-5. DOI: 10.1090/cln/007. URL: https://doi.org/10.1090/cln/007.