# ECE 590.17: Lecture 1 – History of Graphical Models & Inference

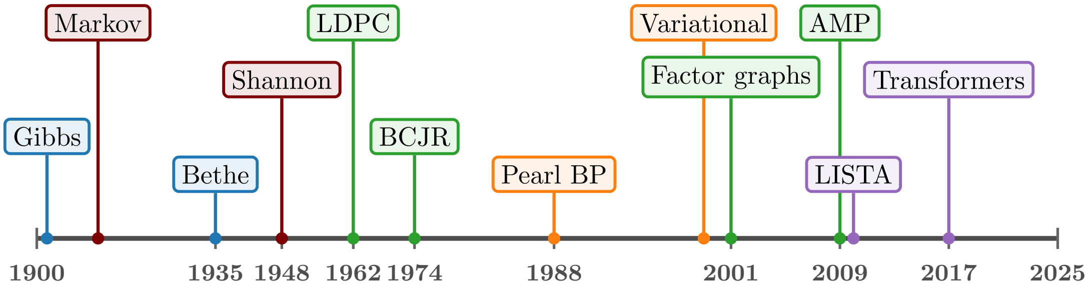
Graphical Models and Inference for Machine Learning

Duke University, Spring 2026

**History:** Written by Henry Pfister (2026).

**Last Modified:** 01/07/2026

## Outline of lecture:



## 1.1 Introduction

Graphical Models and message-passing inference sit at a crossroads of several intellectual traditions that, for much of the twentieth century, developed in parallel. One thread comes from probability and statistics: the idea that uncertainty should be represented quantitatively, and that evidence should be combined systematically to update beliefs. A second thread comes from statistical physics: large interacting systems led physicists to express distributions as products of local interactions (energies) and to seek approximations that become exact on tree-like structures. A third thread comes from communications and signal processing: decoding and estimation problems need effective low-complexity inference algorithms. Many of these would later be recognized as special cases of sum–product and max–product message passing.

The modern unification arrived when researchers realized that the essential object is not a specific application (coding, filtering, vision, or AI), but the factorization of a global function (typically a probability distribution or likelihood) into local terms. Once the factorization is made explicit, the computational story follows: on trees, exact marginals and MAP decisions can be computed by local messages; on graphs with cycles, the same updates become powerful heuristics and, in many cases, principled approximations with variational interpretations (e.g., Bethe free energy and related constructions). This viewpoint also clarifies why different communities invented closely related algorithms under different names: forward–backward, Viterbi, Kalman filtering, iterative decoding, and belief propagation are all instances of the same underlying "compute global quantities from local structure" paradigm.

In the last decade, the factor-graph perspective has become even more central because modern machine learning increasingly blends model-based structure with learned components. Deep unfolding treats an iterative inference method as a computation graph and learns parameters that preserve the algorithm's inductive bias while improving speed and robustness. Graph neural networks and Transformers can be read as learned message passing on graphs (sometimes sparse and structured, sometimes dense and data-dependent) making "messages on graphs" a shared language spanning probabilistic inference, optimization, and representation learning. With this historical lens, the milestones below are not a linear story but a braided one: repeated rediscoveries of the same core idea, followed by periodic syntheses that turned separate techniques into a common framework.

## 1.2 Milestones at a glance

A. **Early probabilistic thinking and "combining evidence"**

- 1763–1812: Bayesian inference becomes a quantitative program (Bayes/Laplace tradition; precursor to all later probabilistic modeling).
- 1902: Gibbs formalizes ensembles and the Gibbs distribution (prototype for factorized probabilistic models).
- 1906: Markov introduces dependent sequences (Markov chains), central to HMMs and state-space inference.
- 1913: Wigmore develops evidence charts (tree-like decomposition of claims/evidence; early "graph of proof").
- 1967–1976: Dempster–Shafer theory (belief functions; alternative calculus of evidence and combination rules).
- 1948/1927: Shannon (classical) and von Neumann (quantum) entropies anchor information measures used throughout inference/learning.

B. **Statistical physics roots of graphical inference**

- 1925: Ising model (canonical lattice interaction model; later MRF archetype).
- 1935: Bethe approximation / Bethe lattice ideas (tree-like approximations; ancestor of BP/Bethe free energy).
- 1951: Kikuchi/cluster variational methods (systematic loop corrections to Bethe).
- 1953: Metropolis Monte Carlo (stochastic computation with Gibbs distributions).
- 1970: Hastings generalizes Metropolis (Metropolis–Hastings MCMC).
- 1971–1984: Hammersley–Clifford and Geman–Geman connect MRFs, Gibbs measures, and practical image restoration.

C. **Graphical models in statistics and AI (MRFs, Bayes nets, junction trees)**

- 1971+: MRF ↔ Gibbs equivalence and conditional independence graphs (formal backbone for undirected models).
- 1988: Pearl popularizes belief networks + "belief propagation" for trees/polytrees.
- 1988: Lauritzen–Spiegelhalter junction tree (exact inference by clique tree / triangulation).
- 1989–1990: HUGIN system and local computation architectures operationalize junction-tree style inference.
- 2006–2009: Textbook-era consolidation (PRML; PGM) standardizes notation, inference recipes, and learning.

D. **Signal processing + coding precursors (dynamic programming on graphs)**

- 1960: Kalman filter (recursive Bayesian inference in linear-Gaussian state-space models).
- 1962: Gallager introduces LDPC and iterative decoding ideas (early form of BP in coding).
- 1967: Viterbi algorithm (max-product on trellises for convolutional codes).
- 1974: BCJR / forward–backward on a trellis gives marginal posterior probabilities.
- 1981: Tanner graphs (explicit bipartite graph representation of codes).

E. **LDPC/turbo era: modern BP "rediscovered" and industrialized**

- 1993: Turbo codes show near-capacity iterative decoding works in practice.
- 1996: MacKay–Neal reignite LDPC with sparse-graph decoding and modern experiments.
- 1996: Wiberg thesis gives unified view of decoding on graphs (sum-product/min-sum).
- 1998: McEliece–MacKay–Cheng explicitly connect turbo decoding to Pearl's BP.

F. **Unification: factor graphs and the generalized distributive law**

- 2000: Aji–McEliece Generalized Distributive Law unifies many algorithms as one template.
- 2001: Kschischang–Frey–Loeliger tutorial formalizes factor graphs + sum-product as a universal algorithm.
- 2001: Forney "codes on graphs" normal realizations (system theory / duality viewpoint).
- 2004: Loeliger Signal Processing Magazine article brings factor graphs into mainstream engineering language.

G. **Loopy BP, variational views, and approximate inference toolkits**

- 1977: EM algorithm (alternating inference/learning; later reframed as message passing/variational).
- 1999: Variational methods for graphical models (mean-field / structured approximations).
- 2001: Expectation propagation (moment-matching message passing beyond mean-field).
- 2001+: BP fixed points as stationary points of Bethe free energy (physics–inference bridge).
- 2005: Variational message passing (VMP) provides an automatic coordinate-ascent/message framework.

H. **Approximate message passing (AMP) and state evolution**

- 2002/2003: Kabashima's low-complexity BP-style multiuser detection for dense CDMA is precursor to later AMP algorithms [23, 24].

- 2009: AMP-style algorithms for compressed sensing (fast BP-inspired iterations for dense linear models).
- 2010/2011: State evolution becomes rigorous (dense-graph analog of density evolution).
- 2010+: GAMP extends AMP to nonlinear output channels and generalized priors.

I. **Modern ML connections: neural message passing, GNNs, and learned inference**

- 2009: Early GNN formalism frames learning as fixed-point message passing on graphs.
- 2010: LISTA (Learned ISTA) — Gregor & LeCun introduce deep unfolding for sparse coding by unrolling ISTA into a fixed-depth trainable network [16].
- 2015+: Encoder–decoder architectures (e.g., U-Net) popularize multi-scale computational graphs (conceptual cousin of multi-resolution inference).
- 2016+: Graph convolution / message-passing neural networks (GNNs) mainstream in deep learning.
- 2016+: "Learned BP" and unrolled inference improve decoding and inference pipelines.
- 2017: Transformer architecture becomes the dominant sequence model. Can be interpreted as learned message passing on a fully connected token graph [46].
- 2021: "Algorithm unrolling" survey (Monga–Li–Eldar) consolidates unfolding across signal processing and inverse problems [37].
- 2025: "Transformers are Graph Neural Networks" formalizes view of Transformers as message-passing GNNs on complete graphs with positional encodings for structure [22].

## 1.3 Milestones with brief annotations and references

### 1.3.1 A. Early probabilistic thinking and "combining evidence"

**1763–1812: Bayesian inference as a quantitative program.** While Bayesian reasoning predates graphical models, the core idea—updating beliefs by likelihood and prior—is the semantic content that later gets *compiled* into factor graphs and message passing. A convenient modern anchor is Pearl's synthesis in [39] (see also its historical discussion). [39]

https://www.google.com/books/edition/Probabilistic_Reasoning_in_Intelligent_S/AvNID7LyMusC

**1902: Gibbs distributions and ensembles.** Gibbs formalized probability distributions over configurations with energy decompositions; this is the conceptual template for factorized models and local interactions. [14] https://www.gutenberg.org/files/50992/50992-pdf.pdf

**1906: Markov chains (dependence with tractable structure).** Markov's dependent sequences provide the mathematical foundation for HMMs/state-space models, where inference becomes dynamic programming/message passing along a chain. [33] https://www.maths.usyd.edu.au/u/eseneta/senetamcfinal.pdf

**1913: Wigmore evidence charts (early "graphs of proof").** Legal scholar J. H. Wigmore developed diagrammatic methods to structure arguments/evidence, anticipating later graph-based representations of reasoning (though not probabilistic in the modern sense). [50] https://archive.org/details/principlesofjudi00wigm

**1967–1976: Dempster–Shafer theory (combination of evidence).** Dempster introduces upper/lower probabilities induced by multivalued mappings, and Shafer develops belief functions and combination rules; this line influenced AI uncertainty frameworks adjacent to Bayesian nets. [9, 43] https://www.glennshafer.com/books/amte.html

**1927/1948: von Neumann and Shannon entropies.** von Neumann's quantum entropy and Shannon's classical entropy formalize information measures that later reappear in variational inference, free energies, and learning objectives. [48, 44] http://eudml.org/doc/59231 https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

### 1.3.2 B. Statistical physics roots of graphical inference

**1925: Ising model (local interactions on a graph).** The Ising model is a canonical pairwise Markov random field on a lattice; it motivated both approximation methods (Bethe/Kikuchi) and MCMC sampling methods used in inference. [19] https://en.wikipedia.org/wiki/Ising_model

**1935: Bethe approximation (tree-like free energy).** Bethe's work on superlattices and the Bethe lattice approximation are historical roots of BP: tree approximations yield tractable local recursions that later map onto sum-product updates. [6] https://www.stat.phys.kyushu-u.ac.jp/B4/Papers/1935-Bethe.pdf

**1951: Kikuchi/cluster variational methods.** Kikuchi generalizes Bethe by using larger clusters to correct for loops—a precursor to modern region-graph/generalized BP ideas. [26] https://link.aps.org/doi/10.1103/PhysRev.81.988

**1953: Metropolis Monte Carlo.** Metropolis et al. introduce the accept/reject sampling method for Gibbs measures, making probabilistic computation feasible when exact summation is impossible. [35] https://bayes.wustl.edu/Manual/EquationOfState.pdf

**1970: Hastings generalization (Metropolis–Hastings).** Hastings broadens Metropolis to general proposal kernels, turning MCMC into a general-purpose inference engine for high-dimensional models. [18] https://academic.oup.com/biomet/article-abstract/57/1/97/284580

**1971–1984: MRF ↔ Gibbs + practical vision inference.** The Hammersley–Clifford theorem underpins the equivalence between undirected graphical models and Gibbs distributions (under positivity), and Geman–Geman popularize Gibbs/MRF methods in Bayesian image restoration. [17, 13] https://www.dam.brown.edu/people/geman/Homepage/Image%20processing,%20image%20analysis,%20Markov%20random%20fields,%20and%20MCMC/stochastic%20relaxation.pdf

### 1.3.3 C. Graphical models in statistics and AI (MRFs, Bayes nets, junction trees)

**1971+: Conditional independence graphs and undirected models.** The formal notion that graph separation encodes conditional independence (for suitable distributions) motivates MRF modeling and sets the stage for exact/approximate inference algorithms. [17] https://en.wikipedia.org/wiki/Hammersley%E2%80%93Clifford_theorem

**1988: Pearl's belief networks + "belief propagation".** Pearl's book consolidates directed graphical models and the message-passing algorithm for trees/polytrees that later becomes the canonical starting point for BP. [39]

https://www.google.com/books/edition/Probabilistic_Reasoning_in_Intelligent_S/AvNID7LyMusC

**1988: Junction tree / clique tree exact inference.** Lauritzen–Spiegelhalter show how triangulation and clique trees enable exact marginalization in general graphs (at exponential cost in treewidth), connecting graph structure to computational complexity. [30] https://en.wikipedia.org/wiki/Junction_tree_algorithm

**1989–1990: HUGIN and local computation architectures.** The HUGIN line of work makes junction-tree propagation concrete in software architectures and emphasizes "local computation" as the guiding engineering principle. [2, 20] https://www.ijcai.org/Proceedings/89-2/Papers/037.pdf https://www.stats.ox.ac.uk/~steffen/papers/jensenetal90.pdf

**2006–2009: Consolidation into ML curricula.** Modern texts unify Bayesian nets, MRFs, approximate inference, and learning (EM/variational/MCMC), becoming the shared language for ML and signal processing communities. [7, 28] https://books.google.com/books/about/Pattern_Recognition_and_Machine_Learning.html?id=kTNoQgAACAAJ https://mitpress.mit.edu/9780262013192/probabilistic-graphical-models/

### 1.3.4 D. Signal processing + coding precursors (dynamic programming on graphs)

**1960: Kalman filtering as recursive Bayesian inference.** The Kalman filter is an early flagship instance of structured probabilistic inference: it passes sufficient statistics forward in time for linear-Gaussian state-space models. [25] https://en.wikipedia.org/wiki/Kalman_filter

**1967: Viterbi algorithm (max-product on a trellis).** Viterbi decoding is dynamic programming for MAP sequence estimation; in modern terms it is max-product message passing on a chain-structured factor graph. [47] https://en.wikipedia.org/wiki/Viterbi_algorithm

**1974: BCJR / forward–backward (sum-product on a trellis).** BCJR computes posterior symbol probabilities via forward/backward recursions; this is sum-product on a chain and a direct ancestor of factor-graph inference in communications. [3] https://en.wikipedia.org/wiki/BCJR_algorithm

**1962: Gallager LDPC and iterative decoding ideas.** Gallager introduces LDPC codes and decoding methods that (in retrospect) foreshadow BP-style iterative message passing on sparse graphs. [12] https://en.wikipedia.org/wiki/Low-density_parity-check_code

**1981: Tanner graphs.** Tanner's bipartite graph representation makes code constraints explicit and provides the natural substrate for iterative local-message decoding. [45] https://en.wikipedia.org/wiki/Tanner_graph

### 1.3.5 E. LDPC/turbo era: modern BP "rediscovered" and industrialized

**1993: Turbo codes.** Turbo codes demonstrate that iterated local computations can approach Shannon limits in practice, catalyzing broad interest in iterative decoding/inference. [5] https://en.wikipedia.org/wiki/Turbo_code

**1996: Modern LDPC revival (MacKay–Neal).** MacKay and Neal's work helps reintroduce LDPC codes to the community with modern experimental validation and the sparse-graph viewpoint aligned with BP. [32] https://www.inference.org.uk/mackay/

**1996: Wiberg's "decoding on graphs" unification.** Wiberg frames decoding/inference algorithms on general graphs (sum-product/min-sum) and clarifies when and why message passing is exact (trees) vs. approximate (loops). [49] https://people.kth.se/~bjornwiberg/

**1998: Turbo decoding as BP (explicit bridge to AI).** McEliece–MacKay–Cheng make the connection explicit: turbo decoding can be understood as Pearl-style belief propagation on a graph with cycles. [34] https://en.wikipedia.org/wiki/Belief_propagation

### 1.3.6 F. Unification: factor graphs and the generalized distributive law

**2000: Generalized distributive law (GDL).** Aji–McEliece show that many algorithms (FFT, Viterbi/BCJR, decoding, etc.) arise from one algebraic pattern: distribute products over sums (or more general semirings) via local message passing. [1] https://authors.library.caltech.edu/records/sw1pm-bwj40/files/AJIieeetit00.pdf

**2001: Factor graphs + sum-product tutorial.** Kschischang–Frey–Loeliger provide the canonical factor-graph formulation and the modern "sum-product algorithm" presentation that made BP a general inference primitive. [29] https://vision.unipv.it/IA2/Factor%20graphs%20and%20the%20sum-product%20algorithm.pdf

**2001: Codes on graphs / normal realizations (Forney).** Forney emphasizes normal realizations and duality, strengthening the links between coding theory, graph-based realizations, and inference/partition functions. [11] https://ieeexplore.ieee.org/document/910573

**2004: Factor graphs as mainstream engineering language.** Loeliger's Signal Processing Magazine article popularizes Forney-style factor graphs and positions them as a unified language for estimation/decoding algorithms. [31] https://people.kth.se/~tjtkoski/factorgraphs.pdf

### 1.3.7 G. Loopy BP, variational views, and approximate inference toolkits

**1977: EM as alternating inference/learning.** EM is a cornerstone for latent-variable learning; it later gets reinterpreted through variational objectives and message passing (E-step as inference, M-step as parameter update). [8] https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

**1999: Variational methods for graphical models.** Jordan et al. systematize variational approximations (mean-field, structured) that trade exactness for tractable optimization and often yield message updates reminiscent of BP. [21] https://en.wikipedia.org/wiki/Variational_Bayesian_methods

**2001: Expectation propagation (EP).** EP generalizes message passing via moment matching; it became a major alternative to BP/mean-field for approximate Bayesian inference. [36] https://en.wikipedia.org/wiki/Expectation_propagation

**2001+: BP fixed points and Bethe free energy.** Yedidia–Freeman–Weiss connect BP fixed points to stationary points of a Bethe free energy, making the statistical-physics lineage explicit and enabling a variational interpretation of loopy BP. [52] https://www.merl.com/publications/docs/TR2001-22.pdf

**2005: Variational message passing (VMP).** VMP formalizes a family of coordinate-ascent variational updates as local messages in conjugate-exponential models, supporting automated inference engines. [51] https://en.wikipedia.org/wiki/Variational_message_passing

### 1.3.8 H. High-dimensional inference: AMP and state evolution

**2002+: Dense-graph BP in CDMA as an AMP precursor (Kabashima).** Kabashima formulated synchronous CDMA multiuser detection as inference on a dense (complete bipartite) graph and showed that belief propagation can be made practical by applying central-limit and self-averaging approximations to the dense messages, yielding an efficient iterative detector and a dynamical analysis linked to statistical mechanics [23, 24]. This is part of the prehistory of approximate message passing for dense linear systems.

**2009: AMP for compressed sensing (dense-graph BP limit).** Donoho–Maleki–Montanari propose BP-inspired iterative thresholding with an "Onsager" correction, yielding accurate, scalable inference for random linear models. [10] https://www.pnas.org/doi/10.1073/pnas.0909892106

**2010/2011: Rigorous state evolution (dense analog of density evolution).** Bayati–Montanari give a rigorous analysis showing AMP's iterates are tracked by a scalar recursion (state evolution), explaining its predictive performance and connecting to spin-glass methods. [4] https://arxiv.org/abs/1001.3448

**2010+: GAMP extends AMP to generalized channels.** Rangan generalizes AMP to nonlinear/quantized output channels and broader priors, framing it as approximate loopy BP with tractable state evolution in large random systems. [40] https://arxiv.org/abs/1010.5141

### 1.3.9 I. Modern ML connections: neural BP, GNNs, learned inference

**2009: GNNs as learned fixed-point message passing.** Early GNN formulations treat learning on graphs as iterated information diffusion (messages) to a fixed point, conceptually mirroring BP but with learned update rules. [42] https://dl.acm.org/doi/abs/10.1109/tnn.2008.2005605

**2010: LISTA and the deep unfolding paradigm.** Gregor and LeCun introduced LISTA (Learned ISTA) as a trainable, fixed-depth network obtained by unrolling the ISTA iterations for sparse coding and learning the linear transforms/thresholds to match the optimal solution in far fewer steps [16]. This kicked off a broad "deep unfolding" (a.k.a. algorithm unrolling) literature in which classical iterative inference/optimization algorithms become structured neural architectures with learnable parameters.

**2015+: Encoder–decoder architectures (e.g., U-Net) and multi-scale computation graphs.** Although not probabilistic message passing per se, U-Net popularizes structured multiresolution computation graphs; this connects naturally to multi-scale inference ideas and hierarchical factorization in practice. [41] https://arxiv.org/abs/1505.04597

**2016/2017: Graph convolutions and message passing neural networks.** Kipf–Welling and Gilmer et al. mainstream the idea that many successful graph models are "message passing" systems with learned aggregation and update maps. [27, 15] https://arxiv.org/abs/1609.02907 https://arxiv.org/abs/1704.01212

**2016+: Learned decoding / unrolled inference.** Unrolling BP as a differentiable computation graph (and learning parameters/weights) bridges classical inference and deep learning, improving decoding and inspiring broader "algorithm unrolling" for inverse problems. [38] https://arxiv.org/abs/1607.04793

**2017: Transformers as learned message passing.** The Transformer introduced self-attention as the core mechanism for sequence modeling in "Attention is all you need", enabling each token to aggregate information from all other tokens with learned, data-dependent weights. From the factor-graph viewpoint, this resembles learned message passing on a on a fully connected token graph whose edge weights are computed by attention. [46]

**2021: Survey / consolidation of algorithm unrolling.** Monga, Li, and Eldar provide a widely cited tutorial-style survey of algorithm unrolling, emphasizing interpretability, efficiency, and hybrid model-based/data-driven design principles across signal and image processing [37].

**2025: Explicit connections between Transformers and GNNs.** Joshi argues that Transformers can be viewed as message passing GNNs operating on fully connected graphs of tokens, with positional encodings injecting inductive bias about structure/order [22]. This provides a clean conceptual bridge between graphical-model inference and modern foundation models.

## 1.4 Why start with history?

> Study the past if you would divine the future." - Confucius

It is neither practical nor wise to try and understand all of the topics mentioned above. One can only hope to capture a flavor from the immense volume of past work. One theme that seems to repeat is that the message-passing frameworks associated with graphical models seem to be useful from a macroscopic perspective. Many of the most powerful machine learning algorithms known have a macroscopic structure that aligns with our understanding of graphical models. However, their amazing performance was unlocked only by learning very complicated factors from massive amounts of data. Early successes in graphical models were mainly for problems where we had very good theoretical models of the probabilistic interactions (e.g., digital communication is Gaussian noise). Speech recognition achieved moderate success based on hidden Markov models with Gaussian mixture observations but real success was only achieved by training neural networks to model output distributions from massive amounts of data. It remains an open question whether a better understanding of generic models like transformers will allow us to reduce their enormous complexity.

## 1.5 Probability Review*

Students who take this class often have a variety of backgrounds, including electrical engineering, statistics, computer science, math, physics, and other engineering and science disciplines. In order to be successful in this course it is necessary that you are *comfortable* working with probability at the undergraduate level. This includes conditioning, expectation, discrete and continuous randoms as well as familiarity with multivariate Gaussian distributions, Markov chains, and convergence of random variables. That said, this course is designed such that you do *not* need probability at the graduate level (i.e., measure-theoretic probability).

- **Probability space**
  - sample space $\Omega$ of all possible outcomes
  - event space $\mathcal{F}$ of events defined on the sample space (e.g., $A, B \in \mathcal{F}$ are events)
  - probability measure $\mathbb{P}$ that satisfies three axioms

$$\mathbb{P}[\Omega] = 1, \quad \mathbb{P}[A] \geq 0, \quad \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] \quad \text{for all } A, B \in \mathcal{F} \text{ with } A \cap B = \emptyset$$

  - Formally, a (real) random variable $X$ is a function from $\Omega$ to $\mathbb{R}$ which specifies the value of $X$ when outcome $\omega$ occurs (e.g., $X(\omega) = \mathbf{1}_A(\omega)$ is the indicator rv of event $A$)

- **Example:** There is a 1% chance I have a certain disease. I take a test for this disease which is 90% accurate. i.e.

$$\mathbb{P}[\text{ positive} \mid \text{disease}] = \mathbb{P}[\text{ negative} \mid \text{no disease}] = 0.9$$

  Given the test is positive, what is the probability I have the disease?

  - Let $A$ be the event I have the disease and $B$ be the event that the test is positive.

$$\mathbb{P}[A] = 0.01, \quad \mathbb{P}[B \mid A] = \mathbb{P}[B^c \mid A^c] = 0.9$$

$$\begin{aligned}\mathbb{P}[A \mid B] &= \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B]} = \frac{\mathbb{P}[B \mid A]\mathbb{P}[A]}{\mathbb{P}[B \mid A]\mathbb{P}[A] + \mathbb{P}[B \mid A^c]\mathbb{P}[A^c]} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} = \frac{9}{118} = \frac{1}{12}\end{aligned}$$

  - Even though the test is highly accurate, the probability I have the disease given a positive outcome is relatively small. This is because the prior probability that I have the disease is very small so the most likely explanation of a positive result is that it's a false positive.

- **Notation**
  - Random variables are denoted by uppercase: $X, Y, Z$
  - Deterministic (i.e., non-random) values denoted by lower case: $x, y, z$
  - Support (or alphabet) of random variable denoted by calligraphic font $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$

- **Discrete random variables**:
  - The **probability mass function (pmf)** of a discrete random variable is $X$ with support $\mathcal{X}$ is given by

$$p_X(x) = \mathbb{P}[X = x] \quad \text{for all } x \in \mathcal{X}$$

  To simplify notation, it is common to write $p(x)$ where the association with the random variable $X$ is implied by the argument of the function.

- ○ The joint pmf of random variables $(X, Y)$ is given by

$$p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y] \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

To simplify notation, it is common to write $p(x, y)$ where the association with the pair $(X, Y)$ is implied by the argument of the function.

- ○ The marginal distributions of $X$ and $Y$ respectively are given by

$$p_X(x) = \sum_{y \in \mathcal{Y}} p_{X,Y}(x, y), \qquad p_Y(y) = \sum_{x \in \mathcal{X}} p_{X,Y}(x, y)$$

- **Independence:**

  - ○ Events $A$ and $B$ are independent if and only if their joint probability equals the product of their probabilities

$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B]$$

  - ○ Random variables $X$ and $Y$ are **independent** if and only if the joint probability is equal to the product of the marginals:

$$p_{X,Y}(x, y) = p_X(x)p_Y(y) \quad \text{for all } (x, y) \in \mathcal{X} \times \mathcal{Y}$$

- **Example:** Let $X$ and $Y$ be independent random variables supported on $\mathcal{X} = \mathcal{Y} = \{1, \dots, M\}$. What is the probability that $X$ equals $Y$?

$$\begin{aligned}\mathbb{P}[X = Y] &= \sum_{m=1}^{M} p_X(m)\mathbb{P}[X = Y \mid X = m] \\ &= \sum_{m=1}^{M} p_X(m)\mathbb{P}[Y = m] \\ &= \sum_{m=1}^{M} p_X(m)p_Y(m)\end{aligned}$$

- **Example:** Provide an example of three random variables $X, Y, Z$ that are pair-wise independent but not independent.

- **Independent and identically distributed (i.i.d.):** A sequence of random variables $X_1, X_2, \dots$ is i.i.d. if the variables are independent and share a common marginal distribution $p(x)$, i.e.,

$$p_{X_1,\dots,X_n}(x_1, \dots, x_n) = \prod_{i=1}^{n} p(x_i)$$

- **Expectation:**

  - ○ The expected value of a random variable $X$ is given by

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\, p_X(x)$$

  - ○ The expected value of a function $f : \mathcal{X} \to \mathbb{R}$ is given by

$$\mathbb{E}[f(X)] = \sum_{x \in \mathcal{X}} f(x)\, p_X(x)$$

- ∘ Warning: the expectation $\mathbb{E}[f(X)]$ not a random variable. Instead, it is a functional of the *distribution* of $X$. Sometimes it is written as $\mathbb{E}_{p_X}[f]$ to make this relationship clear.
- ∘ The conditional expectation of $X$ given a particular realization $\{Y = y\}$ is

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x \, p_{X|Y}(x \mid y)$$

This is a deterministic (nonrandom) quantity that is a function of $y$.

- ∘ The conditional expectation of $X$ given $Y$ is

$$\mathbb{E}[X \mid Y] = \sum_{x \in \mathcal{X}} x \, p_{X|Y}(x \mid Y)$$

This this is a random variable because it is a function of the random variable $Y$.

- **Variance:** The variance of a random variable $X$ is given by

$$\mathsf{Var}(X) = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] = \mathbb{E}\big[X^2\big] - (\mathbb{E}[X])^2$$

and the conditional variance $\mathsf{Var}(X \mid Y)$ is a random variable that, given $Y = y$, equals the variance of $X \sim p_{X|Y}(x \mid y)$.

$$\mathsf{Var}(X \mid Y = y) = \mathbb{E}\big[(X - \mathbb{E}[X \mid Y = y])^2 \mid Y = y\big]$$

- **Example:** The law of total variance states that

$$\mathsf{Var}(X) = \mathbb{E}[\mathsf{Var}(X \mid Y)] + \mathsf{Var}(\mathbb{E}[X \mid Y])$$

- **Law of large numbers (LLN):** If a sequence of random variables $X_1, X_2, \dots$ is i.i.d. with finite absolute first moment $\mathbb{E}[|X_1|] < \infty$, then the long-term average converges to the mean:

$$\frac{1}{n} \sum_{i=1}^{n} X_i \to \mathbb{E}[X]$$

with probability one as $n \to \infty$.

- **Central limit theorem (CLT):** If a sequence of random variables $X_1, X_2, \dots$ is i.i.d. with mean $\mu = \mathbb{E}[X_1]$ and finite variance $\sigma^2 = \mathsf{Var}(X_1) < \infty$ then the fluctuation of the long-term average about the the mean has a Gaussian distribution

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \to \mathcal{N}(0, \sigma^2)$$

in distribution as $n \to \infty$.

- **Example:** Let $X_1, X_2, \dots,$ be a sequence of i.i.d. Bernoulli($p$) variables, $p \in (0, 1)$, and let $S_n = X_1 + \dots + X_n$ be the sum of the first $n$ terms.

  - ∘ $S_n$ has a binomial distribution with parameters $n$ and $p$, and probability mass function

$$p_{S_n}(s) = \binom{n}{k} p^s (1 - p)^{n-s}, \qquad s \in \{0, 1, \dots, n\}$$

and cumulative distribution function

$$F_{S^n}(s) = \mathbb{P}[S_n \le s] = \sum_{k \le t} p_{S_n}(k), \qquad s \in (-\infty, \infty)$$

- By law of large numbers, $\frac{1}{n}S_n \to p$ as $n \to \infty$. This implies that the cdf converges to a step function

$$F_{\frac{1}{n}S^n}(t) = \mathbb{P}\big[n^{-1}S_n \leq t\big] \to \begin{cases} 1, & t > p \\ 0, & t < p \end{cases}$$

- By central limit theorem, $Z_n = (S_n - \mathbb{E}[S_n])/\sqrt{n\,\mathsf{Var}(S_1)}$ converges in distribution to a $\mathcal{N}(0,1)$ random variable as $n \to \infty$. This is equivalent to saying that the cdf converges to the cdf of a standard Gaussian variable:

$$F_{Z_n}(z) = \mathbb{P}\left[\frac{S_n - n\alpha}{\sqrt{n\alpha(1-\alpha)}} \leq z\right] \to \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}\, du$$

- **Warning:** Convergence is distribution can occur to any distribution and it does not always imply that the pmf $p_{Z_n}(\cdot)$ converges to the pdf of the Gaussian distribution.

## References

[1] Srinivas M. Aji and Robert J. McEliece. The generalized distributive law. *IEEE Transactions on Information Theory*, 46(2):325–343, 2000.

[2] S. K. Andersen, F. V. Jensen, and K. G. Olesen. HUGIN—a shell for building bayesian belief universes for expert systems. 1989.

[3] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Transactions on Information Theory*, 20(2):284–287, 1974.

[4] Mohsen Bayati and Andrea Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 2011. Often accessed via the arXiv preprint 1001.3448.

[5] Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. In *Proceedings of ICC*, 1993.

[6] H. A. Bethe. Statistical theory of superlattices. *Proceedings of the Royal Society A*, 150:552–575, 1935.

[7] Christopher M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[9] Arthur P. Dempster. Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2):325–339, 1967.

[10] David L. Donoho, Arian Maleki, and Andrea Montanari. Message-passing algorithms for compressed sensing. *Proceedings of the National Academy of Sciences*, 106(45):18914–18919, 2009.

[11] Jr. G. David Forney. Codes on graphs: Normal realizations. *IEEE Transactions on Information Theory*, 47(2):520–548, 2001.

[12] Robert G. Gallager. Low-density parity-check codes. *IRE Transactions on Information Theory*, 8(1):21–28, 1962.

[13] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

[14] J. Willard Gibbs. *Elementary Principles in Statistical Mechanics.* Charles Scribner's Sons, 1902.

[15] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. *arXiv preprint*, 2017. ICML 2017.

[16] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

[17] J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971. Often cited via circulated/technical reports; the "Hammersley–Clifford theorem" became a standard reference point for MRF ↔ Gibbs equivalence.

[18] W. K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[19] Ernst Ising. Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 1925. Foundational model later central to MRFs and statistical physics inference.

[20] Finn V. Jensen, Steffen L. Lauritzen, and Kristian G. Olesen. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quarterly*, 4:269–282, 1990.

[21] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[22] Chaitanya K. Joshi. Transformers are graph neural networks. *arXiv preprint*, 2025.

[23] Yoshiyuki Kabashima. A statistical-mechanical approach to CDMA multiuser detection: propagating beliefs in a densely connected graph. *arXiv preprint*, 2002.

[24] Yoshiyuki Kabashima. A CDMA multiuser detection algorithm on the basis of belief propagation. *Journal of Physics A: Mathematical and General*, 36(43):11111–11121, 2003.

[25] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82:35–45, 1960.

[26] Ryoichi Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81:988–1003, 1951.

[27] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint*, 2016. Published as ICLR 2017.

[28] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

[29] Frank R. Kschischang, Brendan J. Frey, and Hans-Andrea Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[30] Steffen L. Lauritzen and David J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B*, 50(2):157–224, 1988.

[31] Hans-Andrea Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21(1):28–41, 2004.

[32] David J. C. MacKay and Radford M. Neal. Near shannon limit performance of low density parity check codes. *Electronics Letters*, 1996. Seminal modern rediscovery/popularization of LDPC decoding by message passing.

[33] A. A. Markov. Extension of the law of large numbers to dependent quantities (original 1906 work on "chains"). *Izvestiya Fiziko-matematicheskogo obschestva pri Kazanskom universitete*, 1906. Commonly cited as Markov's first paper on what are now called Markov chains; bibliographic details vary across translations.

[34] Robert J. McEliece, David J. C. MacKay, and Jung-Fu Cheng. Turbo decoding as an instance of pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications*, 16(2):140–152, 1998.

[35] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[36] Thomas P. Minka. *Expectation Propagation for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[37] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[38] Eliya Nachmani, Yair Beery, and David Burshtein. Learning to decode linear codes using deep learning. *arXiv preprint*, 2016.

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

[40] Sundeep Rangan. Generalized approximate message passing for estimation with random linear mixing. *arXiv preprint*, 2010.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint*, 2015. MICCAI 2015.

[42] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[43] Glenn Shafer. *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[44] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3,4):379–423, 623–656, 1948.

[45] R. Michael Tanner. A recursive approach to low complexity codes. *IEEE Transactions on Information Theory*, 27(5):533–547, 1981.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[47] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[48] J. von Neumann. Thermodynamik quantenmechanischer gesamtheiten. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, 1927:273–291, 1927.

[49] Niclas Wiberg. *Codes and Decoding on General Graphs.* PhD thesis, Linköping University, 1996.

[50] John Henry Wigmore. *The Principles of Judicial Proof.* Little, Brown, and Company, 1913.

[51] John Winn and Christopher M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 2005. Commonly cited from the VMP technical report and JMLR-era dissemination; bibliographic variants exist.

[52] Jonathan S. Yedidia, William T. Freeman, and Yair Weiss. Understanding belief propagation and its generalizations. Technical report, MERL, 2001.