

Markov Networks

Aayush Bajaj

August 31, 2026

some dependencies have no natural direction. two neighbouring pixels tend to share a label; two friends tend to vote alike; two adjacent atoms couple their spins. forcing an arrow onto these symmetric interactions — as a bayesian network must — misrepresents them. the **markov network** (markov random field, mrf) is the undirected alternative, born in statistical physics: edges express mutual compatibility, parametrised not by conditional probabilities but by unnormalised *potentials*, with a global normalising constant picking up the bill.

the trade is exactly that: local, symmetric, easy-to-change factors — in exchange for a partition function that is np-hard to evaluate

Contents

1	the gibbs distribution	1
1.1	the voting example	2
1.2	hammersley-clifford	2
2	factor graphs	2
3	energy functions	3
4	independence semantics	3
4.1	bayesian vs markov networks	3
5	inference	4
6	conditional random fields	4
7	results	4
8	see also	5

1 the gibbs distribution

a markov network over variables $\mathbf{X}$ is an undirected graph H plus a set of **potential functions** $\psi_k(\mathbf{D}_k)$ — non-negative factors, each over a clique $\mathbf{D}_k$ of the graph. the joint is their normalised product, the **gibbs distribution**:

a clique is a set of pairwise-connected nodes; a maximal clique is one no superset of which is a clique

$$P(\mathbf{x}) = \frac{1}{Z} \prod_k \psi_k(\mathbf{d}_k), \quad Z = \sum_{\mathbf{x}} \prod_k \psi_k(\mathbf{d}_k). \quad (1)$$

Z is the **partition function**. unlike cpts, potentials need not sum to anything — $\psi(a, b) = 100$ just says the configuration is 100 times more compatible than $\psi = 1$ — which makes them easy to elicit and easy to tweak locally, but pushes all the normalisation into Z , a sum over every instantiation.

1.1 the voting example

four colleagues; friends influence each other's vote, and the friendships form a cycle:

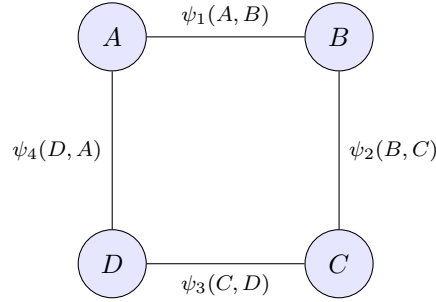


Figure 1: a pairwise Markov network. each edge carries a potential scoring agreement between its endpoints; there is no direction and no cpt — just compatibilities, normalised globally by Z .

with agreement potentials like $\psi_1(a, b) = 30$, $\psi_1(a, \bar{b}) = 5$, $\psi_1(\bar{a}, b) = 1$, $\psi_1(\bar{a}, \bar{b}) = 10$, the joint over votes is the product of the four edge tables over Z . note the graph has a cycle — no Bayesian network could use this topology directly.

1.2 Hammersley–Clifford

the correspondence between graph and distribution is a theorem, not a definition. a distribution P is **positive** if $P(\mathbf{x}) > 0$ for every instantiation. the **Hammersley–Clifford theorem**: a positive distribution satisfies the Markov properties of graph H (below) *iff* it is a Gibbs distribution over the cliques of H . positivity is essential — Bayesian networks happily encode deterministic zeros, but here zeros can break the equivalence.

the same graph supports many factorisations: potentials over maximal cliques, pairwise potentials over edges, or higher-order mixtures. the graph alone therefore *underspecifies* the parametrisation, which is where factor graphs earn their keep.

2 factor graphs

a **factor graph** makes the factorisation explicit: a bipartite graph with round nodes for variables and square nodes for factors, an edge whenever the factor mentions the variable.

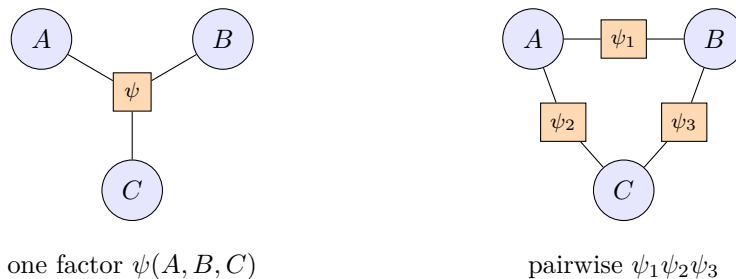


Figure 2: two factor graphs over the same three-variable clique: one ternary factor (left) versus three pairwise factors (right). the plain Markov network — a triangle — cannot distinguish them.

factor graphs are also the natural home of **belief propagation**: messages flow variable→factor and factor→variable, and the formulation covers directed and undi-

rected models uniformly.

3 energy functions

positive potentials can always be written $\psi(\mathbf{d}) = e^{-\epsilon(\mathbf{d})}$, turning the gibbs distribution into its physics form:

$$P(\mathbf{x}) = \frac{1}{Z} e^{-E(\mathbf{x})}, \quad E(\mathbf{x}) = \sum_k \epsilon_k(\mathbf{d}_k), \quad (2)$$

with E the **energy** — a sum of local terms. maximising probability is minimising energy, products become sums (numerically kinder), and mpe/map queries become combinatorial optimisation. this **map-mrf** framing runs much of classical low-level vision: for binary image denoising, take a grid-shaped pairwise network with a data term pulling each pixel towards its observed value and a smoothness term charging disagreeing neighbours,

$$E(\mathbf{x}) = \sum_i \alpha [x_i \neq o_i] + \sum_{(i,j) \in \text{grid}} \beta [x_i \neq x_j], \quad (3)$$

and search for a low-energy labelling. the assignment space is astronomically large, so the search is stochastic:

- **iterated conditional modes (icm)** — flip each variable to its locally best value; pure greedy descent, stops in the nearest local minimum.
- **metropolis** — propose a flip, always accept downhill, accept uphill with probability $e^{-\Delta E/T}$; the uphill acceptances are what let the search escape local minima.
- **simulated annealing** — metropolis with temperature T lowered on a schedule: explore early, commit late.

metropolis moves are also the correctness engine of mcmc — the same acceptance rule that escapes minima leaves the gibbs distribution invariant

4 independence semantics

undirected independence is plain graph **separation**: $\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$ is asserted whenever every path from $\mathbf{X}$ to $\mathbf{Y}$ passes through $\mathbf{Z}$ — no valves, no collider exception, just blocking by removal. like d-separation it is sound and (for particular parametrisations) incomplete. two local corollaries:

- deleting an edge asserts a conditional independence between its endpoints;
- the **markov blanket** of a node is exactly its neighbours: $X \perp \text{rest} \mid \text{Neighbours}(X)$ — no spouse clause, because there are no colliders.

4.1 bayesian vs markov networks

aspect	bayesian network	markov network
edges	directed, acyclic	undirected, cycles fine
parameters	cpts — easy to elicit, locally normalised	potentials — easy to adjust, globally normalised by Z
independence	d-separation (with explaining away)	separation (monotone blocking)
generativity	ancestral sampling by topological order — trivial	sampling generally requires mcmc
normalisation	automatic ($Z = 1$)	Z is np-hard in general
reading factors	factorisation explicit in the dag	ambiguous without a factor graph

neither language subsumes the other: a dag’s collider independencies ($B \perp C$, not $B \perp C \mid D$) have no undirected representation, and a cycle’s independencies ($A \perp C \mid B, D$ and $B \perp D \mid A, C$) have no dag. converting a bayesian network to a markov i-map is **moralisation** — marry co-parents, drop arrows — as drawn on the exact inference page; the marriage edges are precisely the marginal parent-independencies sacrificed.

5 inference

variable elimination and the jointree algorithm run unchanged on the potentials — indeed the interaction graph on which elimination is analysed *is* a markov network, and the markov-network community’s “clique tree algorithm” is the jointree algorithm under another name. the one genuinely new object is Z : eliminating every variable from the potential pool computes it, at the usual $O(n \exp w)$ — so conditional queries $P(\mathbf{q} \mid \mathbf{e})$, where Z cancels in the ratio, are no harder than in bayesian networks, but the *absolute* probability of a single configuration, trivial in a bayesian network, costs a full inference here.

6 conditional random fields

a **crf** is a markov network over labels $\mathbf{Y}$ whose potentials are functions of an observed input $\mathbf{x}$ — modelling $P(\mathbf{Y} \mid \mathbf{x})$ directly and refusing to model $\mathbf{x}$ at all:

$$P(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_k \psi_k(\mathbf{y}_k; \mathbf{x}). \quad (4)$$

this is the **discriminative** move: spend all capacity on the decision boundary, none on the input distribution. linear-chain crfs superseded hmms for sequence labelling (the potentials may inspect the whole input, not just the aligned emission), and grid crfs did the same to mrfs for segmentation. the price of discrimination: no generative story, so no synthetic data and no principled handling of missing inputs.

the classic generative/discriminative pair is naive bayes vs logistic regression — the same model class, trained to different objectives; with little data the generative bias wins, with much data the discriminative fit does

7 results

- **gibbs distribution** — $P = \frac{1}{Z} \prod_k \psi_k$ over cliques of H ; potentials non-negative, unnormalised.
- **hammersley–clifford** — for positive P : markov w.r.t. $H \iff$ gibbs over H ’s cliques.
- **separation soundness** — graph separation implies conditional independence in every gibbs distribution over H ; not complete.
- **markov blanket** — in a markov network, a node’s blanket is its neighbour set.
- **moralisation** — marrying co-parents and dropping directions yields a minimal markov i-map of a bayesian network.
- **inference** — ve/jointree apply verbatim; computing Z is #p-hard in general; conditionals dodge Z .
- **map-mrf** — map inference as energy minimisation; icm, metropolis, simulated annealing as the stochastic search ladder.

8 see also

- bayesian networks — the directed counterpart
- exact inference — elimination and jointrees, shared machinery
- approximate inference — gibbs sampling and loopy belief propagation live on these graphs
- computer vision — crfs and superpixel segmentation in the wild