

ECE 590.17: Lecture 4 – Machine Learning and Statistical Physics

Graphical Models and Inference for Machine Learning
Duke University, Spring 2026

History: Written by Henry Pfister (2026).

Last Modified: 02/10/2026

Outline of lecture:

4.1	Machine Learning and Statistical Physics	1
4.2	The Boltzmann distribution	1
4.3	Free energy and thermodynamic quantities	2
4.4	Energy-based models and factor graphs	3
4.5	Example: Ising model	3
4.6	Worked example: 1D Ising chain via transfer matrices	4
4.7	Limiting behavior and intuition	5
4.8	Optimization viewpoint	5
4.9	Summary	6
4.10	Exercises	6

4.1 Machine Learning and Statistical Physics

Many models used in machine learning and inference can be expressed in terms of energy:

$$\mu(x) = \frac{1}{Z(\beta)} \exp(-\beta E(x)), \quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}.$$

This is precisely the Boltzmann distribution introduced in statistical physics. It is a natural distribution to use when configurations with lower energy (or lower cost, or higher likelihood) are more probable. This section reviews the basic statistical physics quantities (free energy, entropy, internal energy) using probabilistic language and connects them to inference, optimization, and factor graphs. More advanced topics (phase transitions and computational hardness) are deferred to a later lecture. In contrast to earlier notes / lectures on factor graphs, we use $\mathcal{X} = \mathcal{X}_0^n$ to denote the entire configuration space and \mathcal{X}_0 to represent one component.

4.2 The Boltzmann distribution

Given a configuration space \mathcal{X} and an energy function $E(x)$, one gets the Boltzmann distribution

$$\mu_\beta(x) = \frac{1}{Z(\beta)} e^{-\beta E(x)}, \quad Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}.$$

The inverse temperature $\beta = 1/T$ controls average energy. Low temperature (high β) prefers low-energy states and high temperature allows the mass to spread into high-energy states. For example, a small β gives a diffuse distribution whereas large β concentrates the mass on minimizers of E . Two limiting cases are useful:

- $\beta \rightarrow 0$: μ_β approaches the uniform distribution.

- $\beta \rightarrow \infty$: μ_β concentrates on ground states $x \in \arg \min E(x)$.

These limits connect probabilistic inference and combinatorial optimization.

Maximizing entropy under an energy constraint.

Lemma 1. *Among all distributions $p(x)$ on \mathcal{X} with fixed mean energy $\sum_x p(x)E(x) = U$, the distribution that maximizes Shannon entropy is the Boltzmann distribution $p(x) \propto e^{-\beta E(x)}$ for some Lagrange multiplier β .*

Proof. Maximize $-\sum_x p(x) \ln p(x)$ subject to $\sum_x p(x) = 1$ and $\sum_x p(x)E(x) = U$. The Lagrangian is

$$\mathcal{L}(p, \lambda, \beta) = -\sum_x p(x) \ln p(x) + \lambda \left(\sum_x p(x) - 1 \right) - \beta \left(\sum_x p(x)E(x) - U \right).$$

Setting the derivative w.r.t. $p(x')$ to zero gives $-\ln p(x') - 1 + \lambda - \beta E(x') = 0$, giving

$$p(x') = \exp(\lambda - 1) \exp(-\beta E(x')).$$

Normalization implies $\exp(\lambda - 1) = 1/Z(\beta)$ and yields $p(x) = e^{-\beta E(x)}/Z(\beta)$. \square

Moment generating function of the energy. Under $\mu_\beta(x) = e^{-\beta E(x)}/Z(\beta)$, the moment generating function of E is

$$M_E(t) = \mathbb{E}_{\mu_\beta}[e^{tE}] = \frac{1}{Z(\beta)} \sum_x e^{-\beta E(x) + tE(x)} = \frac{Z(\beta - t)}{Z(\beta)}.$$

Thus, the cumulant generating function (or log moment generating function) is

$$K_E(t) = \ln M_E(t) = \ln Z(\beta - t) - \ln Z(\beta).$$

Differentiating at $t = 0$ yields the cumulants at fixed β :

$$K'_E(0) = -\frac{\partial}{\partial \beta} \ln Z(\beta) = \mathbb{E}_{\mu_\beta}[E], \quad K''_E(0) = \frac{\partial^2}{\partial \beta^2} \ln Z(\beta) = \text{Var}_{\mu_\beta}(E).$$

4.3 Free energy and thermodynamic quantities

For abstract systems like this, we define the free energy and free entropy by

$$F(\beta) := -\frac{1}{\beta} \ln Z(\beta), \quad \Phi(\beta) := \ln Z(\beta).$$

The internal energy is the expected energy

$$U(\beta) = \mathbb{E}_{\mu_\beta}[E(X)] = -\frac{\partial}{\partial \beta} \ln Z(\beta),$$

and the entropy (measured in nats) is

$$\begin{aligned} S(\beta) &= -\sum_{x \in \mathcal{X}} \mu_\beta(x) \ln \mu_\beta(x) \\ &= -\sum_x \mu_\beta(x) (-\beta E(x) - \ln Z(\beta)) \end{aligned}$$

$$\begin{aligned}
&= \beta \sum_x \mu_\beta(x) E(x) + \ln Z(\beta) \sum_x \mu_\beta(x) \\
&= \beta U(\beta) + \ln Z(\beta).
\end{aligned}$$

Equivalently,

$$\ln Z(\beta) = S(\beta) - \beta U(\beta).$$

This shows how normalization depends on the balance of typical energy and entropy: increasing entropy (more configurations with non-negligible probability) or decreasing average energy both increase $\ln Z(\beta)$. The trade-off depends on β .

Thus, the free energy decomposes as

$$F(\beta) = U(\beta) - \frac{1}{\beta} S(\beta).$$

Convexity. The log-partition function $\ln Z(\beta)$ is convex in β because

$$\frac{\partial^2}{\partial \beta^2} \ln Z(\beta) = \text{Var}_{\mu_\beta}(E(X)) \geq 0.$$

Equivalently, $-\ln Z(\beta)$ (and thus $\beta F(\beta)$) is concave. The variance of energy is the curvature of $\ln Z(\beta)$.

4.4 Energy-based models and factor graphs

In factor graph form, a distribution is defined by factors $f_a(x_{\partial a}) > 0$ and

$$\mu(x) = \frac{1}{Z} \prod_{a \in F} f_a(x_{\partial a}).$$

Define local energies $E_a(x_{\partial a}) = -\ln f_a(x_{\partial a})$ and

$$E(x) = \sum_{a \in F} E_a(x_{\partial a}).$$

Then μ is exactly a Boltzmann distribution with $\beta = 1$. This makes statistical physics language natural for inference on factor graphs and reveals a variational interpretation of belief propagation. The Bethe free energy, introduced later, provides an approximation for $\ln Z$ and a variational view of BP fixed points.

4.5 Example: Ising model

Let $\sigma \in \{\pm 1\}^n$ denote a vector of electron *spins* whose pairwise interactions are defined by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V} = [n]$ and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Then, the Ising model is defined by the factorization

$$f_i(\sigma_i) = \exp(\beta h_i^{\text{ext}} \sigma_i), \quad f_{ij}(\sigma_i, \sigma_j) = \exp(\beta J_{ij} \sigma_i \sigma_j),$$

so that

$$E(\sigma) = - \sum_{(i,j) \in \mathcal{E}} J_{ij} \sigma_i \sigma_j - \sum_i h_i^{\text{ext}} \sigma_i.$$

When all $J_{ij} > 0$, the distribution favors aligned spins and the system is called *ferromagnetic*. When all $J_{ij} < 0$, the distribution favors anti-aligned spins and the system is called *antiferromagnetic*.

The unary terms h_i^{ext} bias individual spins and are called *external fields*. Below is a small factor graph with external fields.

For a given configuration, the *effective field* acting on spin i is

$$h_i^{\text{eff}} := h_i^{\text{ext}} + \sum_{j \in \partial i} J_{ij} \sigma_j,$$

where ∂i is the neighbor set of i in \mathcal{G} . This quantity the external field and the influence of all neighboring spins.

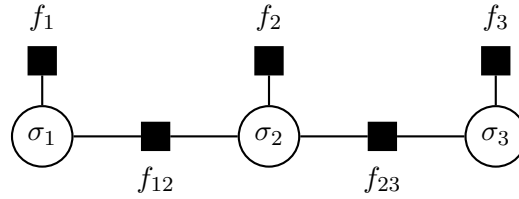


Figure 4.1: A 3-spin chain with pairwise couplings and unary external fields.

The total magnetization and total energy per spin are

$$M(\sigma) = \frac{1}{|\mathcal{V}|} \sum_i \sigma_i, \quad U(\sigma) = \frac{E(\sigma)}{|\mathcal{V}|}.$$

As β grows, μ_β shifts from a near-uniform distribution to one concentrated around highly aligned configurations (neighboring spins agree with high probability).

4.6 Worked example: 1D Ising chain via transfer matrices

Consider a chain of n spins $\sigma_i \in \{\pm 1\}$ with uniform coupling J and uniform external field h :

$$E(\sigma) = -J \sum_{i=1}^{n-1} \sigma_i \sigma_{i+1} - h \sum_{i=1}^n \sigma_i.$$

Define the 2×2 transfer matrix indexed by $\sigma, \tau \in \{\pm 1\}$ as

$$T_{\sigma, \tau} = \exp\left(\beta J \sigma \tau + \frac{\beta h}{2} (\sigma + \tau)\right).$$

Then the partition function can be written as

$$Z_n(\beta) = \sum_{\sigma_1, \sigma_n} u_{\sigma_1} (T^{n-1})_{\sigma_1, \sigma_n} u_{\sigma_n}, \quad u_\sigma = \exp\left(\frac{\beta h}{2} \sigma\right).$$

Explicitly,

$$T = \begin{bmatrix} e^{\beta(J+h)} & e^{-\beta J} \\ e^{-\beta J} & e^{\beta(J-h)} \end{bmatrix}.$$

For large n , $Z_n(\beta)$ is dominated by the largest eigenvalue of T . The eigenvalues of T are

$$\lambda_{\pm} = e^{\beta J} \left(\cosh(\beta h) \pm \sqrt{\sinh^2(\beta h) + e^{-4\beta J}} \right).$$

Thus, in the thermodynamic limit, the free energy per spin is

$$\lim_{n \rightarrow \infty} F_n(\beta) = \lim_{n \rightarrow \infty} -\frac{1}{\beta n} \ln Z_n(\beta) = -\frac{1}{\beta} \ln \lambda_+.$$

Define the spin–spin correlation function by

$$C_{ij} = \mathbb{E}_{\mu_\beta}[\sigma_i \sigma_j] - \mathbb{E}_{\mu_\beta}[\sigma_i] \mathbb{E}_{\mu_\beta}[\sigma_j].$$

For the 1D chain, C_{ij} decays exponentially with distance, $C_{ij} \sim \exp(-|i - j|/\xi)$ for large $|i - j|$. The correlation length ξ is defined by the eigenvalue gap:

$$\xi^{-1} = \ln\left(\frac{\lambda_+}{\lambda_-}\right),$$

because correlations decay as $\exp(-|i - j|/\xi)$ along the chain.

4.7 Limiting behavior and intuition

High temperature ($\beta \rightarrow 0$). All configurations are nearly equally likely. Expected energy is close to the uniform average, and magnetization concentrates near zero.

Low temperature ($\beta \rightarrow \infty$). Probability concentrates on global minimizers of $E(x)$. This is the link between statistical physics and optimization: sampling from μ_β at large β approximates solving $\min_x E(x)$.

4.8 Optimization viewpoint

Define a cost function $E(x)$ for a combinatorial optimization problem. The Boltzmann distribution provides a soft relaxation: instead of searching only for the minimum, we study the full distribution μ_β . At finite β , typical configurations balance cost and multiplicity, which can help avoid poor local minima. Simulated annealing exploits this by slowly increasing β to focus on low-cost configurations.

Example: softening a constrained problem. Let $\phi_a(x_{\partial a}) \in \{0, 1\}$ be constraint indicators and define the number of violated constraints

$$E(x) = \sum_{a \in F} (1 - \phi_a(x_{\partial a})).$$

Then $E(x) = 0$ if and only if all constraints are satisfied. The Boltzmann distribution

$$\mu_\beta(x) \propto \exp(-\beta E(x))$$

assigns exponentially larger mass to configurations with fewer violations. As $\beta \rightarrow \infty$, μ_β concentrates on satisfying assignments, while at finite β it explores near-feasible solutions.

Example: quadratic unconstrained binary optimization. For $x_i \in \{\pm 1\}$ and symmetric weights Q_{ij} , consider

$$E(x) = -\sum_{i < j} Q_{ij} x_i x_j - \sum_i b_i x_i.$$

This is an Ising model with local fields and pairwise couplings. The distribution μ_β favors low-energy assignments; at large β the MAP state solves the QUBO, while at smaller β the distribution weights multiple near-optimal states, which can be useful for search and for quantifying uncertainty among competing optima.

4.9 Summary

- Boltzmann distributions formalize how local energies induce global probability measures.
- Free energy, entropy, and internal energy translate naturally into probabilistic terms and provide useful identities.
- Phase transitions correspond to sharp changes in typical behavior and often correlate with algorithmic hardness.
- Energy-based models unify statistical physics with inference and optimization on factor graphs.

4.10 Exercises

1. **Two-state Boltzmann model.** Let $\mathcal{X} = \{0, 1\}$ with energies $E(0) = 0$ and $E(1) = \Delta$ where $\Delta > 0$. For $\beta \geq 0$, define $\mu_\beta(x) \propto e^{-\beta E(x)}$.
 - (a) Compute $Z(\beta)$ and $\mu_\beta(1)$.
 - (b) Compute $U(\beta) = \mathbb{E}_{\mu_\beta}[E(X)]$ and $S(\beta) = -\sum_x \mu_\beta(x) \ln \mu_\beta(x)$.
 - (c) Verify directly that $\ln Z(\beta) = S(\beta) - \beta U(\beta)$ for this model.
2. **Derivative identities for F and U .** Let $Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}$, $\Phi(\beta) = \ln Z(\beta)$, and $F(\beta) = -(1/\beta)\Phi(\beta)$, with $U(\beta) = -\Phi'(\beta)$.
 - (a) Show that $\frac{d}{d\beta}(\beta F(\beta)) = U(\beta)$.
 - (b) Show that $\frac{dF}{d\beta} = \frac{S(\beta)}{\beta^2}$ where $S(\beta) = \beta U(\beta) + \Phi(\beta)$.
 - (c) Conclude that $F(\beta)$ is nondecreasing in β .
3. **Legendre transform / variational form of $F(\beta)$.** Define the Gibbs free energy functional for any distribution p on \mathcal{X} :

$$\mathcal{G}_\beta(p) := \sum_x p(x)E(x) + \frac{1}{\beta} \sum_x p(x) \ln p(x).$$

Show that $\min_p \mathcal{G}_\beta(p) = F(\beta)$ and identify the minimizer.

4. **Heat capacity identity.** Define $U(\beta) = \mathbb{E}_{\mu_\beta}[E]$ and recall $\Phi''(\beta) = \text{Var}_{\mu_\beta}(E)$.
 - (a) Show that $\frac{dU}{d\beta} = -\text{Var}_{\mu_\beta}(E)$.
 - (b) Let temperature be $T = 1/\beta$. Define the heat capacity $C(T) := \frac{dU}{dT}$. Show that $C(T) = \beta^2 \text{Var}_{\mu_\beta}(E)$.
5. **Entropy monotonicity.** Let $S(\beta) = \beta U(\beta) + \Phi(\beta)$.
 - (a) Show that $\frac{dS}{d\beta} = -\beta \text{Var}_{\mu_\beta}(E)$.
 - (b) This shows mathematically that $S(\beta)$ is nonincreasing in β . Provide an intuitive explanation for this fact.

6. **Transfer matrix: recover λ_{\pm} directly.** For the 1D Ising chain with uniform coupling J and field h , the transfer matrix is

$$T = \begin{bmatrix} e^{\beta(J+h)} & e^{-\beta J} \\ e^{-\beta J} & e^{\beta(J-h)} \end{bmatrix}.$$

- (a) Compute $\text{tr}(T)$ and $\det(T)$.
(b) Use the quadratic formula for eigenvalues (i.e., $\lambda^2 - (\text{tr } T)\lambda + \det T = 0$) to show

$$\lambda_{\pm} = e^{\beta J} \left(\cosh(\beta h) \pm \sqrt{\sinh^2(\beta h) + e^{-4\beta J}} \right).$$

- (c) In the zero-field case $h = 0$, simplify λ_{\pm} and the correlation length $\xi^{-1} = \ln(\lambda_+/\lambda_-)$.