

# ECE 590.17 Lecture 4: Machine Learning and Statistical Physics

Duke University, Spring 2026  
Instructor: Henry Pfister

**Last Modified:** 02/11/2026

- Energy-based models and the Boltzmann distribution; max-entropy derivation.
- Partition function  $Z(\beta)$  and  $\ln Z(\beta)$ : moments, free energy, and entropy.
- Convexity/variance links and the energy–entropy trade-off interpretation.
- Factor graphs as energies; inference motivation and Bethe/BP preview.
- Ising model example: factor graph, transfer matrix solution, correlation length.
- Optimization view: annealing, constraints, and QUBO  $\leftrightarrow$  Ising.

# Energy-based models: A probabilistic view

- For a model with configuration energy  $E(x)$ ,

$$\mu_\beta(x) = \frac{1}{Z(\beta)} \exp(-\beta E(x)),$$

$$Z(\beta) = \sum_{x \in \mathcal{X}} e^{-\beta E(x)}.$$

- This is exactly the **Boltzmann distribution**.
- **Interpretation:**  
lower energy (cost or loss)  $\Rightarrow$  higher probability.
- Inverse temperature  $\beta = 1/T$  controls width:
  - small  $\beta \Rightarrow$  diffuse distribution,
  - large  $\beta \Rightarrow$  concentrates near minimizers.

## Two useful limits

- $\beta \rightarrow 0$ :

$$\mu_\beta(x) \rightarrow \frac{1}{|\mathcal{X}|} \quad (\text{uniform}).$$

- $\beta \rightarrow \infty$ :

$$\mu_\beta \text{ concentrates on } \arg \min_{x \in \mathcal{X}} E(x).$$

## Big picture

- Inference: compute marginals /  $Z(\beta)$
- Optimization: recover minimizers of  $E$
- Sampling: explore near-optimal states

# The Boltzmann distribution

## Lemma (energy vs. entropy trade-off).

Among all distributions  $p$  on  $\mathcal{X}$  such that

$$\sum_{x \in \mathcal{X}} p(x) = 1, \quad \sum_{x \in \mathcal{X}} p(x)E(x) = U, \quad (\text{i.e., Average energy } U)$$

the distribution maximizing Shannon entropy,

$$H(p) = - \sum_x p(x) \ln p(x)$$

**Proof outline.** Write Lagrangian for max entropy under an energy constraint:

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

Using the stationary condition  $\partial \mathcal{L} / \partial p(x') = 0$ , we find that

$$-\ln p(x') - 1 + \lambda - \beta E(x') = 0 \quad \Rightarrow \quad p(x') = \exp(\lambda - 1) e^{-\beta E(x')}.$$

Normalizing gives  $\exp(\lambda - 1) = 1/Z(\beta)$ .

## Moment generating function of $E$

$$\mu_\beta(x) = \frac{e^{-\beta E(x)}}{Z(\beta)}$$

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

Cumulant (or log-moment) generating function:

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

## Derivatives at $t = 0$

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

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

## Takeaway

- $\ln Z(\beta)$  is a generating function for energy moments.
- Curvature of  $\ln Z(\beta)$  is **energy variance**.

# Free energy, internal energy, and free entropy

## Definitions

Free entropy and free energy:

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

Internal energy:

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

Entropy (nats):

$$S(\beta) = -\sum_{x \in \mathcal{X}} \mu_\beta(x) \ln \mu_\beta(x).$$

**Key identity** Using  $\mu_\beta(x) = Z(\beta)^{-1} e^{-\beta E(x)}$ ,

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

Equivalently,

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

Balances energy vs. multiplicity (entropy).

# Curvature of free entropy = variance of energy fluctuations

## Convexity of $\ln Z(\beta)$

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

- log-sum-exp:  $\ln \sum_i e^{\beta a_i}$  always convex
- Thus,  $\ln Z(\beta)$  is convex in  $\beta$ .
- And  $\beta F(\beta) = -\ln Z(\beta)$  is concave.

## Interpretation

- Curvature  $\leftrightarrow$  energy fluctuations.
- Large variance means average energy decreases rapidly with  $\beta$ .

## Useful mental model

- $U(\beta)$ : typical energy level
- $S(\beta)$ : number of typical configurations
- $F(\beta)$ : energy–entropy trade-off

# Energy-based models and factor graphs

## Factor graph form

distribution specified by positive factors  $f_a(x_{\partial a})$ :

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

Define **local energies**

$$E_a(x_{\partial a}) := -\ln f_a(x_{\partial a}), \quad E(x) = \sum_{a \in \mathcal{F}} E_a(x_{\partial a}).$$

Then, we have

$$\mu(x) = \frac{1}{Z} e^{-E(x)} \quad (\text{Boltzmann with } \beta = 1).$$

## Why this matters

- Inference often requires  $\ln Z$  (or marginals).
- Statistical physics language becomes natural:

$\ln Z \leftrightarrow$  free entropy.

- **Preview:** Free energy approximations
- BP fixed points provide a variational interpretation and approximation known as the Bethe free energy.

# Example: Ising model on a graph

## Spins and pairwise interactions.

Let  $\sigma \in \{\pm 1\}^n$  and  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  with  $\mathcal{V} = [n]$ .

## Factors:

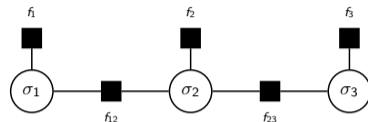
$$f_i(\sigma_i) = \exp(\beta h_i^{\text{ext}} \sigma_i), \quad f_{ij}(\sigma_i, \sigma_j) = \exp(\beta J_{ij} \sigma_i \sigma_j).$$

## Energy:

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

## Interpretation

- $J_{ij} > 0$  favors **aligned** neighbors.
- $h_i^{\text{ext}}$  is an **external field** biasing  $\sigma_i$ .
- As  $\beta$  increases: distribution moves from near-uniform to highly aligned configurations.



3-spin chain with unary external fields and pairwise couplings.

## Avg spin, Avg energy, effective field

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

$$U(\sigma) = \frac{E(\sigma)}{|\mathcal{V}|},$$

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

# Worked example: 1D Ising chain via transfer matrices

**1D uniform chain** ( $J_{i,i+1} = J$ ,  $h_i^{\text{ext}} = h$ )

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

Transfer matrix  $T \in \mathbb{R}^{2 \times 2}$  indices  $\sigma, \tau \in \{\pm 1\}$ :

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

**Partition function:**

$$Z_n(\beta) = \sum_{\sigma_1, \sigma_n} u_{\sigma_1} (T^{n-1})_{\sigma_1, \sigma_n} u_{\sigma_n}.$$

Boundary vector:

$$u_{\sigma} = \exp\left(\frac{\beta h}{2} \sigma\right).$$

**Explicit matrix**

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

**Eigenvalues (closed form)**

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

**Key idea:** for large  $n$ ,  $Z_n$  is dominated by the largest eigenvalue  $\lambda_+(\beta)$ .

## Free energy per spin (limit $n \rightarrow \infty$ )

$$F_n(\beta) := -\frac{1}{\beta n} \ln Z_n(\beta).$$

Since  $Z_n$  is dominated by  $\lambda_+^{n-1}$ ,

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

## Exactly Solvable Model

- 1D Ising has exact solution via linear algebra.
- Most models lack closed form solutions  $\Rightarrow$  approximation necessary (BP/Bethe).

## Correlation function

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

For the 1D chain,

$$C_{ij} \sim \exp(-|i-j|/\xi) \quad \text{for large } |i-j|.$$

Correlation length given by eigenvalue gap:

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

Large  $\xi \Rightarrow$  long-range dependence.

## High temperature: $\beta \rightarrow 0$

- All configurations nearly equally likely:

$$\mu_\beta(x) \approx \frac{1}{|\mathcal{X}|}.$$

- Expected energy close to the uniform average.
- Magnetization concentrates near 0 (symmetry, weak coupling).

## Low temperature: $\beta \rightarrow \infty$

- Distribution concentrates near global minimizers:

$$\mu_\beta \Rightarrow \text{mass on } \arg \min E(x).$$

- Direct link to **combinatorial optimization**.
- Sharp change with  $\beta$  hints at **phase transition** (later lecture).

## Boltzmann as a soft relaxation

Instead of only searching for the minimum of  $E(x)$ , study the full family  $\{\mu_\beta\}_{\beta \geq 0}$ :

$$\mu_\beta(x) \propto e^{-\beta E(x)}.$$

- Small  $\beta$ : explore broadly.
- Large  $\beta$ : focus on low-cost states.
- **Simulated annealing**: slowly increase  $\beta$ .

## Example: constraint violations

Constraints  $\phi_a(x_{\partial a}) \in \{0, 1\}$ , define

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

Then  $E(x) = 0$  iff all constraints satisfied.

Consider the Boltzmann distribution:

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

- finite  $\beta$ : explores near-feasible assignments,
- $\beta \rightarrow \infty$ : concentrates on satisfying assignments (if any).

## Quadratic unconstrained binary optimization

For  $x_i \in \{\pm 1\}$ , symmetric  $Q_{ij}$  and biases  $b_i$ :

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

- This is exactly an **Ising model**: pairwise couplings  $J_{ij} = Q_{ij}$  and external fields  $h_i^{\text{ext}} = b_i$ .
- MAP / ground state solves the QUBO:

$$x^* \in \arg \min_x E(x).$$

## Why keep $\beta$ finite?

- At moderate  $\beta$ ,  $\mu_\beta$  weighs many near-optimal states.
- Useful for:
  - randomized search / exploration,
  - quantifying uncertainty among competing optima,
  - avoiding poor local minima (annealing heuristics).
- Partition function  $Z(\beta)$  summarizes the entire landscape through a single scalar.

# Summary

- Boltzmann distribution formalizes how **energies/costs** induce global probability.
- $\ln Z(\beta)$  is central: derivatives give **mean energy** and **energy variance**.
- Free energy identity:

$$F(\beta) = U(\beta) - \frac{1}{\beta} S(\beta) \quad (\text{energy-entropy trade-off}).$$

- Factor graphs naturally define energies:

$$\mu(x) \propto \prod_a f_a(x_{\partial a}) \iff \mu(x) \propto e^{-E(x)}.$$

- Ising models connect inference, statistical physics, and optimization; 1D chain solved via transfer matrices; correlation length from eigenvalue gap.
- Foundation for later ideas regarding phase transitions and algorithmic hardness.