The Maximum Entropy Principle

Applications in Machine Learning

Stuart Truax, 2025-08-01

github.com/StuartTruax

Outline

A. The Concept of Entropy
B. Maximum Entropy Principle
C. Maximum Entropy vs. Maximum Likelihood
D. Multinomial Logistic Regression
E. Energy-based Models
F. Bayesian Formalism
G. Energy Function and Functionals
H. Energy Functionals in Bayesian Variational Inference
I. Summary

2024 Nobel Prize in Physics

© Nobel Prize Outreach. Photo: Nanaka Adachi

John J. Hopfield

Prize share: 1/2

© Nobel Prize Outreach. Photo: Clément Morin

Geoffrey Hinton

Prize share: 1/2

The Nobel Prize in Physics 2024 was awarded jointly to John J. Hopfield and Geoffrey Hinton "for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Hopfield Network

energy function $E = -\frac{1}{2}\sum_{i,j} w_{ij} s_i s_j - \sum_i \theta_i s_i$

Boltzmann Machine

energy function $E = -\left(\sum_{i<j} w_{ij}\, s_i\, s_j + \sum_i \theta_i\, s_i\right)$

Ising Model

Hamiltonian $H(\sigma) = -\sum_{\langle ij \rangle} J_{ij}\sigma_i\sigma_j - \mu\sum_j h_j\sigma_j,$

[1] https://www.nobelprize.org/prizes/physics/2024/summary/
[2] https://en.wikipedia.org/wiki/Hopfield_network
[3] https://en.wikipedia.org/wiki/Ising_model
[4] https://en.wikipedia.org/wiki/Boltzmann_machine

The Concept of Entropy

Historical Development of the Concept of Entropy

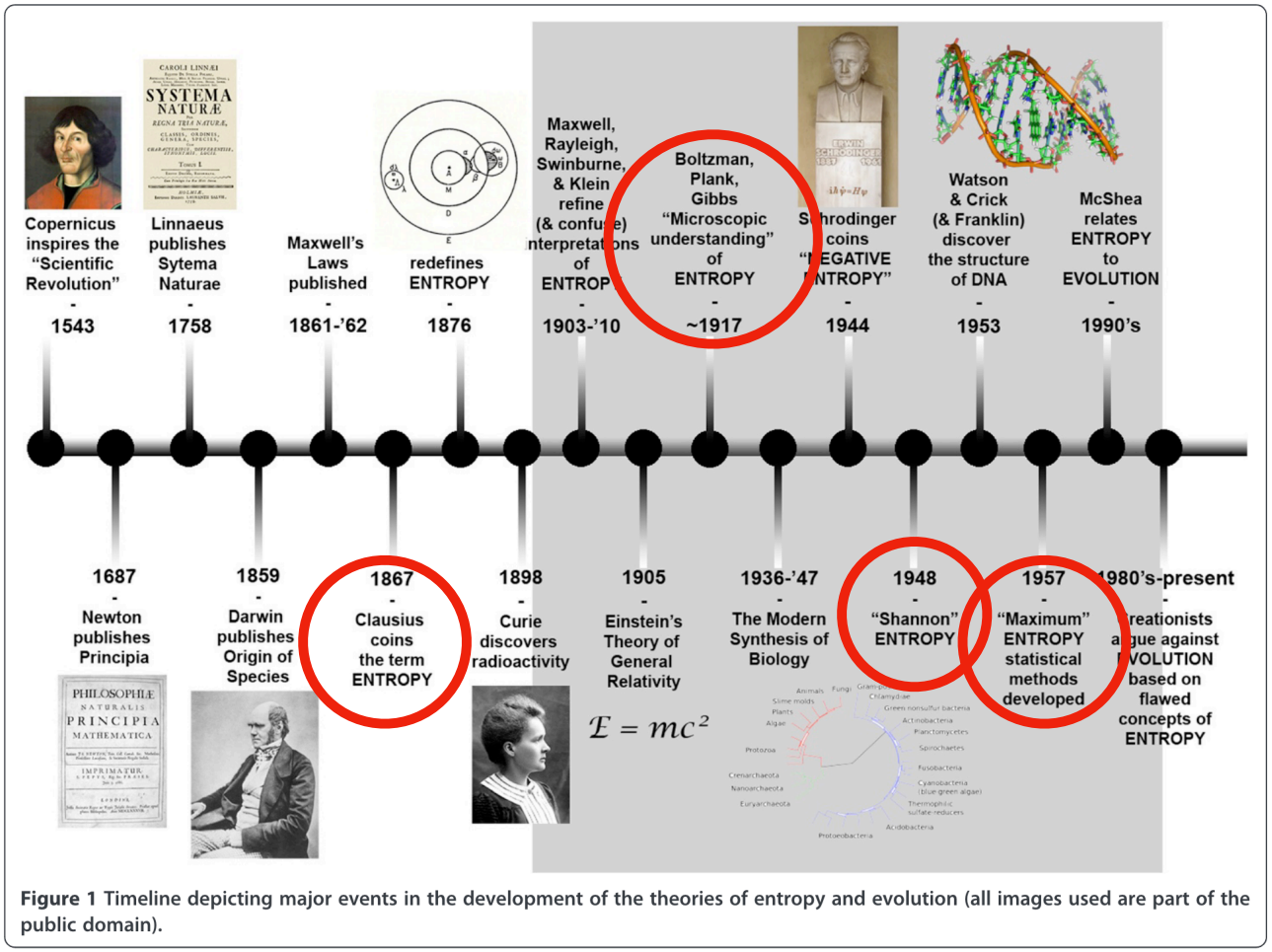


Figure 1 Timeline depicting major events in the development of the theories of entropy and evolution (all images used are part of the public domain).

[1] J. S. Martin, N. A. Smith, and C. D. Francis, "Removing the entropy from the definition of entropy: clarifying the relationship between evolution, entropy, and the second law of thermodynamics".

ONE

THERMODYNAMICS AND STATISTICAL MECHANICS

1.1 INTRODUCTION: THERMODYNAMICS AND STATISTICAL MECHANICS OF THE PERFECT GAS

Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest, carrying on the work, died similarly in 1933. Now it is our turn to study statistical mechanics.

Perhaps it will be wise to approach the subject cautiously. We will begin by considering the simplest meaningful example, the perfect gas, in order to get the central concepts sorted out. In Chap. 2 we will return to complete the solution of that problem, and the results will provide the foundation of much of the rest of the book.

The quantum mechanical solution for the energy levels of a particle in a box (with periodic boundary conditions) is

$$\frac{h^2 q^2}{\quad} \qquad (1.1.1)$$

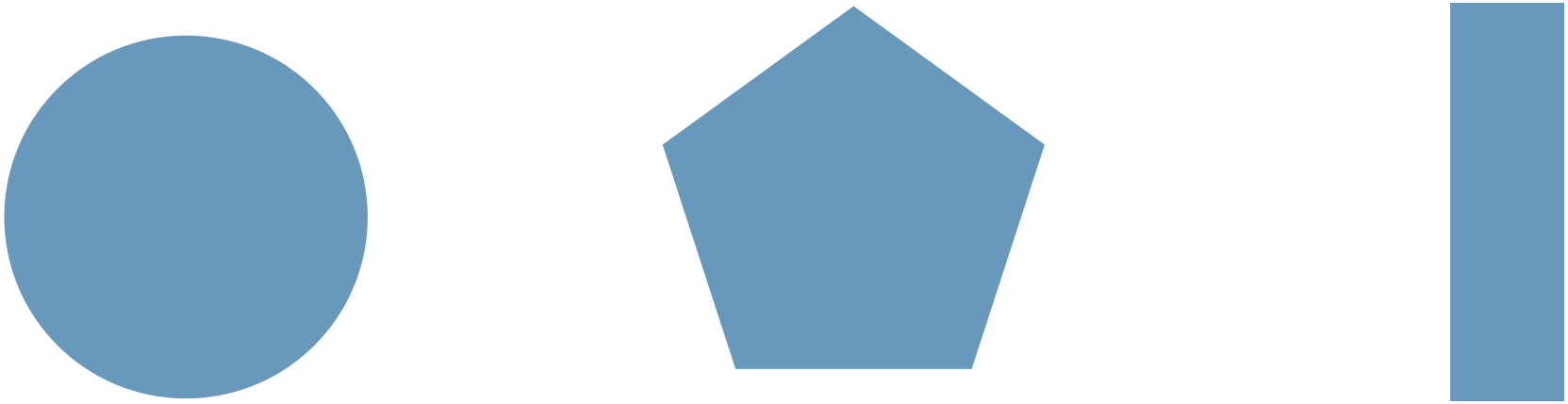
D. Goodstein, *States of Matter*, 1975


S = k. log W
LVDWIG
BOLTZMANN
1844 – 1906
HENRIETTE
BOLTZMANN
GEB. EDLE von AIGENTLER
1854 – 1938
DR PHIL. PAULA
BOLTZMANN
GEB. CHIARI
1891 – 1977
ARTHUR
BOLTZMANN
DIPL. ING. DR PHIL. HOFRAT
1881 – 1932
LVDWIG
BOLTZMANN
1923 – 1943
LETZTER MÄNNLICHER NACHKOMME.
GEFALLEN BEI SMOLENSK


S = k. log W

Intuitive Notion of Entropy

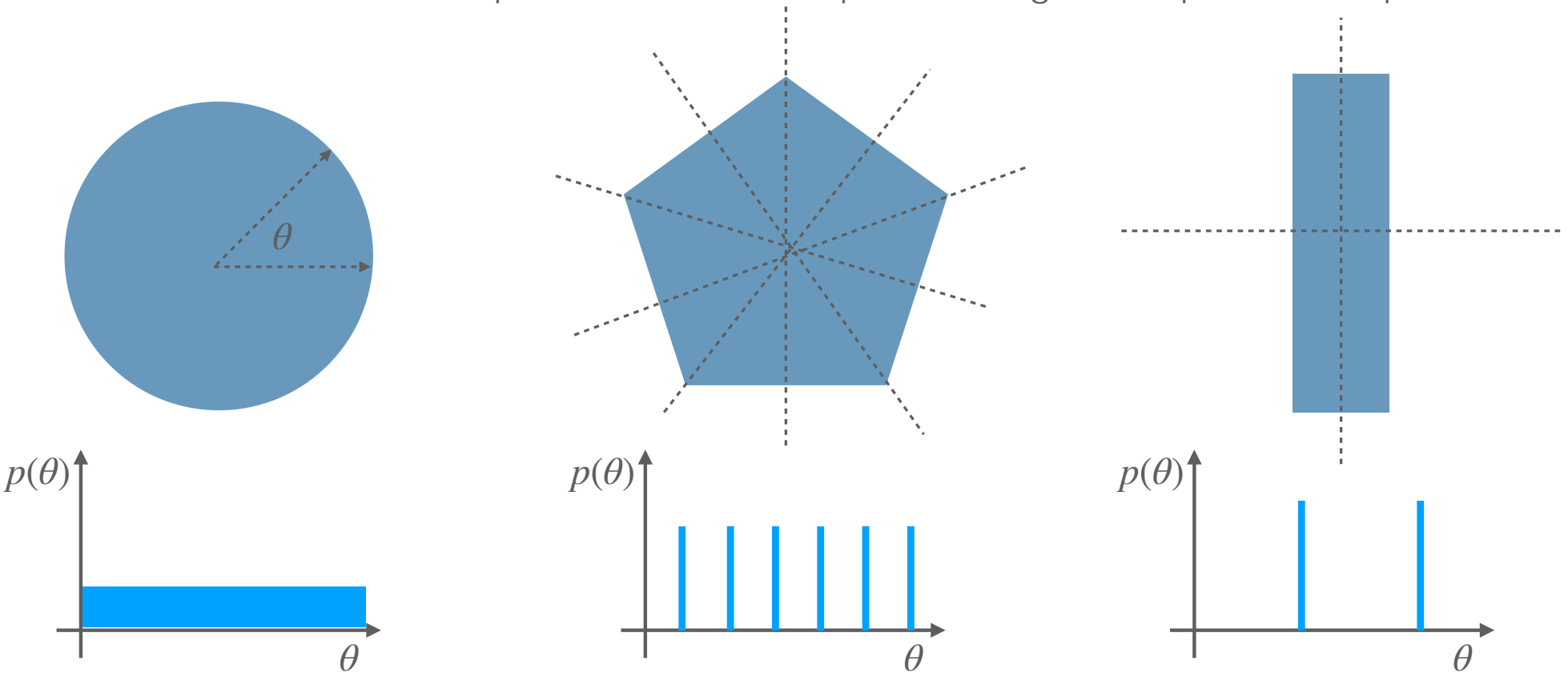
Imagine these shapes are randomly rotating so fast as to be imperceptible, yet they appear as below.



Which rotations of these respective shapes are most likely?

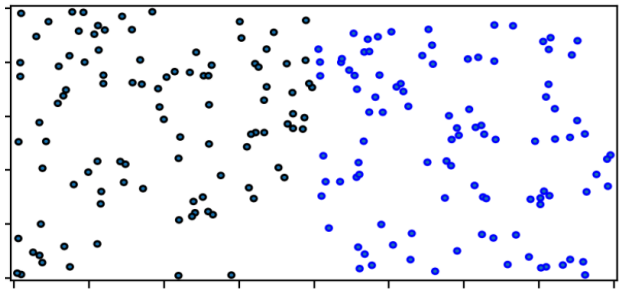
Intuitive Notion of Entropy

What kinds of rotations can I perform on each shape and still get an equivalent shape?



Statistical Mechanics

We have a system, composed of a large number of particles, each with a thermodynamic state variable X, also called an *observable.*



X:

- U - internal energy
- N - particle number
- q - charge
- x - displacement
- M - magnetization

We can only observe an average value of X for the system, and this value must satisfy an equality constraint. That is it is a *conserved quantity.*

$$\mathbb{E}_p[X] = \mu$$

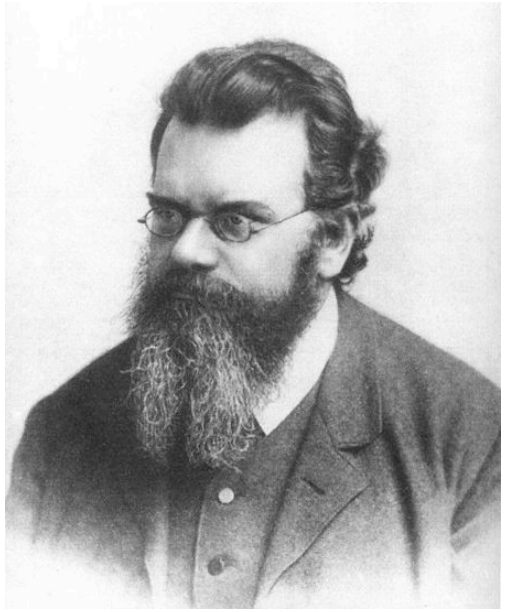
Suppose the configuration of particles in the system is random, and drawn from a distribution p.

What is the most likely p?

[1] N. Borghini, *Topics in Non-Equilibrium Physics*, U. Bielefeld, 2014

Entropy as a Count of Microstates

$$S = \mathrm{k_b} \log(\Omega)$$



where:

- Ω is the **multiplicity** (i.e. count) of possible microstates (configurations) that result in the same macrostate (observable).
- k_b is Boltzmann's constant

Macrostates with more associated microstates are more probable, and thus have higher entropy.

The observable macrostate is defined in terms of a *conserved quantity (*i.e. a *constraint on the system).*

Entropy in Expressed as Probabilities

- Entropy is a *statistic* calculated from a distribution.

Probabilisitic definition of multiplicity, where p_i is the probability of a single particle being in a given state.

N total particles
i possible states

$$\log(\Omega) = -N \sum_i p_i \log(p_i)$$

$$\rightarrow \quad S = -k_b N \sum_i p_i \log(p_i)$$

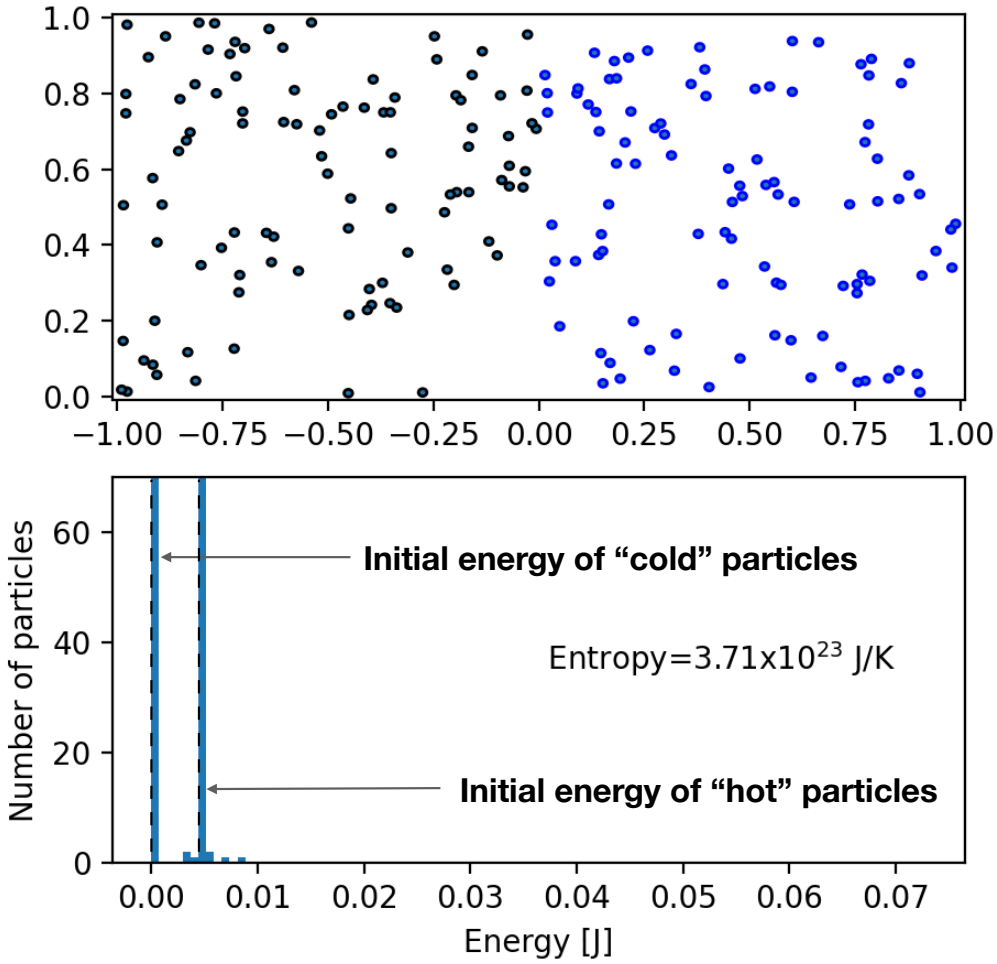
Isolated system,
Many uncorrelated particles

Gibb's Entropy:

$$S = -k_b \sum_i p_i \log(p_i)$$

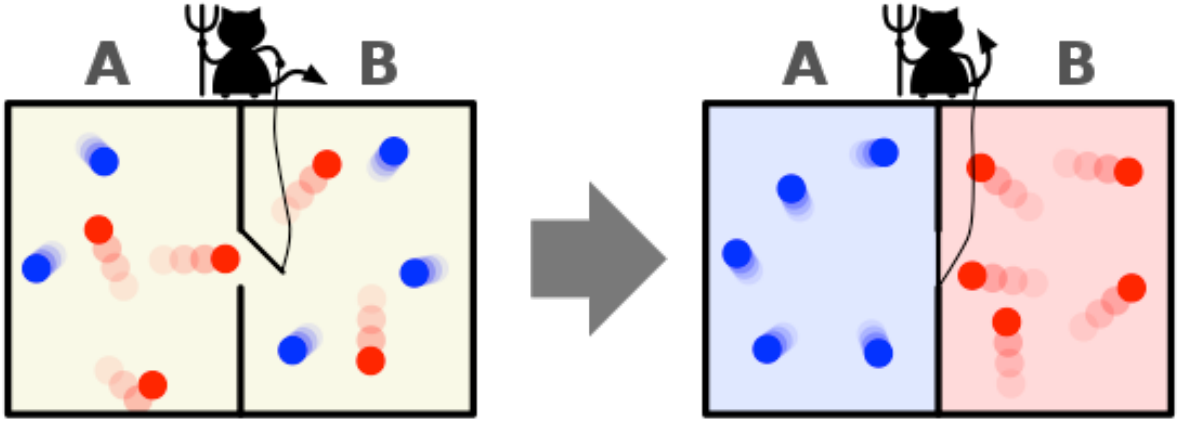
General System

Non-equilibrium: Entropy Production



Maxwell's Demon

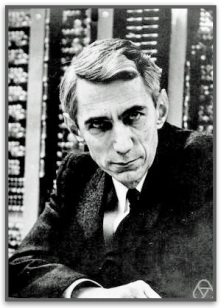
- 1929: Leo Szilard addresses the "Maxwell's demon" problem
- An intelligent being intervenes to reverse the production of entropy, particle-by-particle.
- Szilard concludes that the being must acquire **information** about the microscopic state of the system, and this acquisition would in itself (regardless of the details) **produce sufficient entropy to offset the entropy reversal** (i.e. "no free lunch").
- The act of measurement by the demon constitutes a gain in information (i.e. entropy).

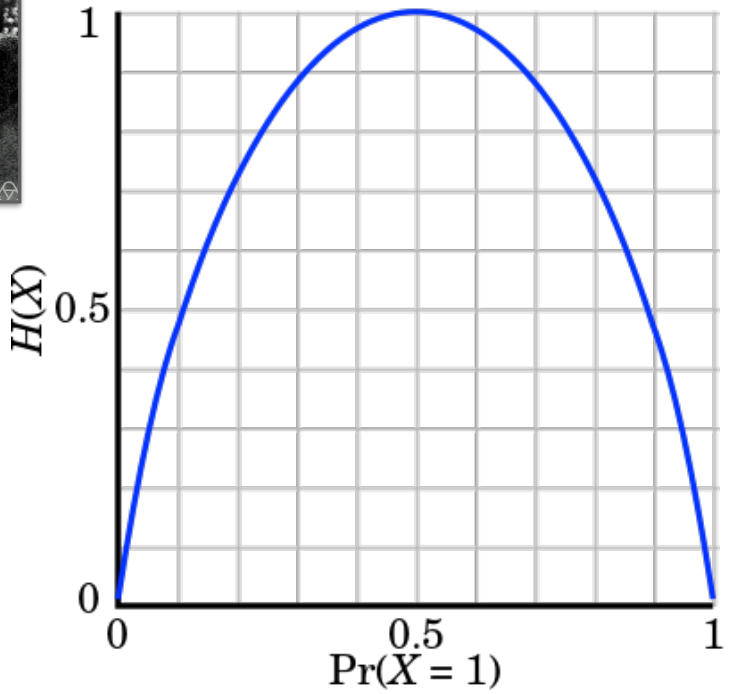


[1] L. Szilard, "On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings," *Zeitschrift fur Physik*, vol. 53, pp. 840–856, 1929
[2] https://en.wikipedia.org/wiki/Maxwell%27s_demon
[3] Landauer, 1961

Entropy as Information: Shannon and Boltzmann

- 1948: Claude Shannon was working on a mathematical theory of communication: **information theory**
- Modeled a discrete information source as a **Markov process** *X*
- Derived a metric which characterized **how much information** is "produced" by the process





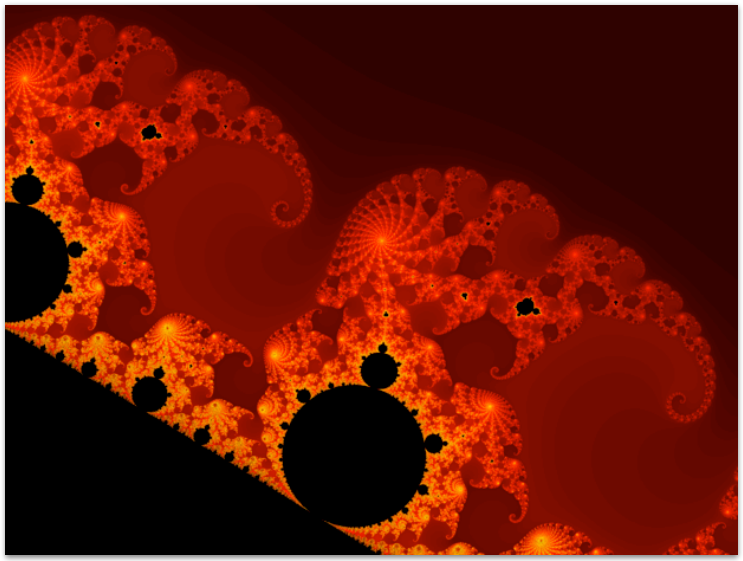
$$H(X) = -K \sum_i p(x_i) \log(p(x_i))$$

[1] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.
[2] https://en.wikipedia.org/wiki/Entropy_(information_theory)

Entropy as Computational Complexity

Kolmogorov complexity:

For a given fixed universal description language L, the Kolmogorov complexity of a string x is the shortest length program P that can produce x



Results:

- Kolmogorov complexity equivalent to the Shannon entropy for Turing-computable probability distributions.
- Kolmogorov complexity converges to the Shannon entropy for a Markovian information source.

[1] A. Teixeira, A. Matos, A. Souto, and L. Antunes, “Entropy Measures vs. Kolmogorov Complexity,” *Entropy*, vol. 13, no. 3, pp. 595–611, Mar. 2011, doi: 10.3390/e13030595.
[2] A. Kaltchenko, “Algorithms for Estimating Information Distance with Application to Bioinformatics and Linguistics”.
[3] https://en.wikipedia.org/wiki/Kolmogorov_complexity

Maximum Entropy Principle

Maximum Entropy: Derivation via the Method of Lagrange

Entropy of a continuous distribution

$$H(x) = -\int p(x)\log(p(x))dx$$

Subject to:

1. Normalization constraint (i.e. probability distribution):

$$\int p(x)dx = 1$$

2. Moment constraints ($i = 1...k$):

$$\int p(x)f_i(x)dx = \mu_i$$

$f_i(x)$ is just some function on x. These constraints are constraints on the expectation of these functions, e.g. $\mathbb{E}_p[f_i(x)] = \mu_i$

[1] https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution (Section: Distribution with Measured Constraints)
[2] warwick.ac.uk/fac/cross_fac/complexity/study/msc_and_phd/co904/co904online/lecture-3-4.pdf

Maximum Entropy: Derivation via the Method of Lagrange

Process

1. Form the Lagrangian

$$\mathcal{L}[p] = -\int p(x)\log p(x)dx + \lambda_0\left(1 - \int p(x)dx\right) + \sum_{i=1}^{k}\lambda_i\left(\mu_i - \int p(x)f_i(x)dx\right)$$

2. Take the functional derivative

$$\frac{\delta\mathcal{L}}{\delta p(x)} = -\log p(x) - 1 - \lambda_0 - \sum_{i=1}^{k}\lambda_i f_i(x) = 0$$

[1] https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution (Section: Distribution with Measured Constraints)
[2] warwick.ac.uk/fac/cross_fac/complexity/study/msc_and_phd/co904/co904online/lecture-3-4.pdf

Maximum Entropy: Derivation via the Method of Lagrange

Process

3. Solve for $p(x)$

$$\log p(x) = -1 - \lambda_0 - \sum_{i=1}^{k} \lambda_i f_i(x)$$

$$\rightarrow \quad p(x) = \exp\left(-1 - \lambda_0 - \sum_{i=1}^{k} \lambda_i f_i(x)\right)$$

4. Exponential distribution (Boltzmann)

$$p(x) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{k} \lambda_i f_i(x)\right)$$

Where $Z = \int \mathbf{exp}\left(-\sum_{i=1}^{k} \lambda_i f_i(x)\right)$

is called the **normalization** or **partition function**

[1] https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution (Section: Distribution with Measured Constraints)
[2] warwick.ac.uk/fac/cross_fac/complexity/study/msc_and_phd/co904/co904online/lecture-3-4.pdf

Maximum Entropy Distributions Under Different Constraints

Distribution		Constraint	
Uniform	$x \sim U(0,a)$	Continuous variable	
Exponential	$x \sim exp(-\alpha x)$	Positive continuous variable	$\mathbb{E}_p[x] = \frac{1}{\alpha}$
Normal	$x \sim \mathcal{N}(\mu, \sigma)$	Continuous variable	$\mathbb{E}_p[x] = \mu$ $Var[x] = \sigma^2$

[1] K. Bacalwski, *Introduction to Probability with R*, Chapter 9, CRC Press, 2008

Maximum Entropy vs. Maximum Likelihood

Maximum Entropy Distribution of a discrete RV: Uniform Distribution

Minimize $\sum_{i}^{n} p_i \log(p_i)$

Subject to:

$\mathbf{1}^T p = 1$

$p_i \geq 0$

Where:

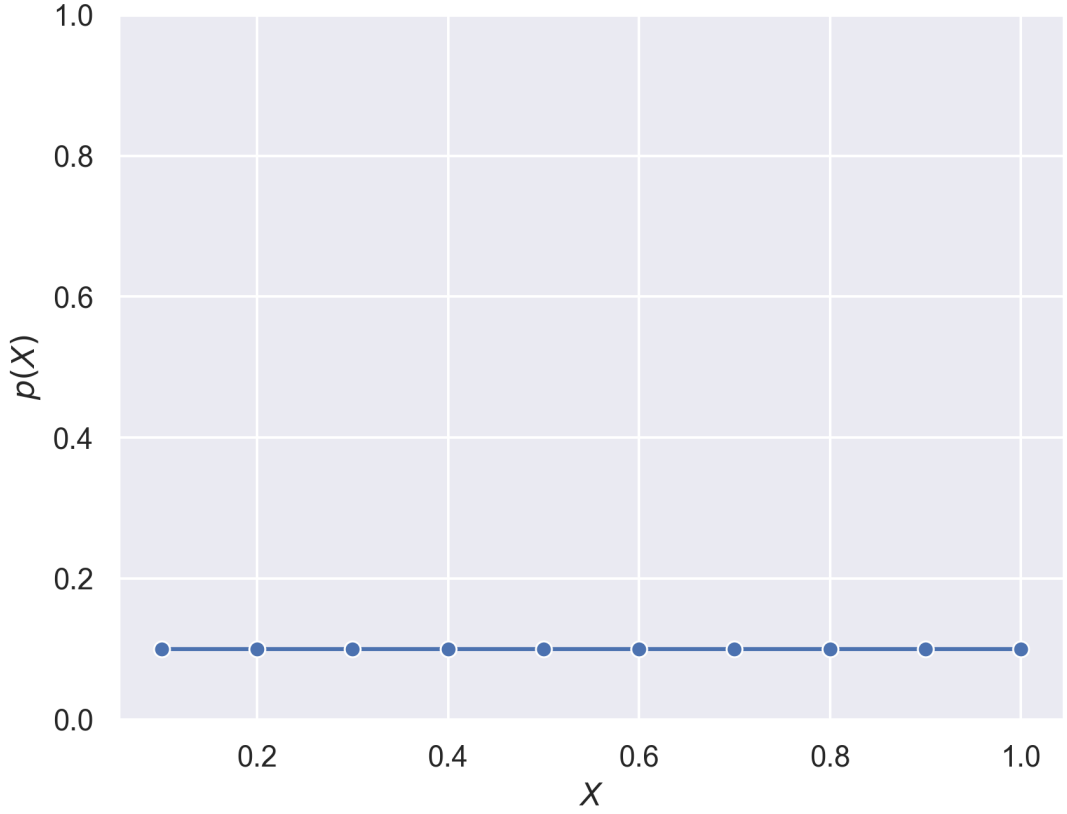
$p_i = Pr(X = i)$ with

$X \in \{1,..n\}$ and

$i = 1,...,n$

Note that the only constraints are that the probabilities behave like probabilities

Convex optimization immediately yields the uniform distribution



[2] https://en.wikipedia.org/wiki/Maximum_entropy_probability_distribution

Maximum Entropy vs Maximum Likelihood for an Empirical Distribution

Subject to:

$$\mathbf{1}^T p = 1$$

$$p_i \geq 0$$

$$E_p[X] = s_x$$

Where:

$$p_i = Pr(X = i)$$

$$N_i = count(X = i)$$

with

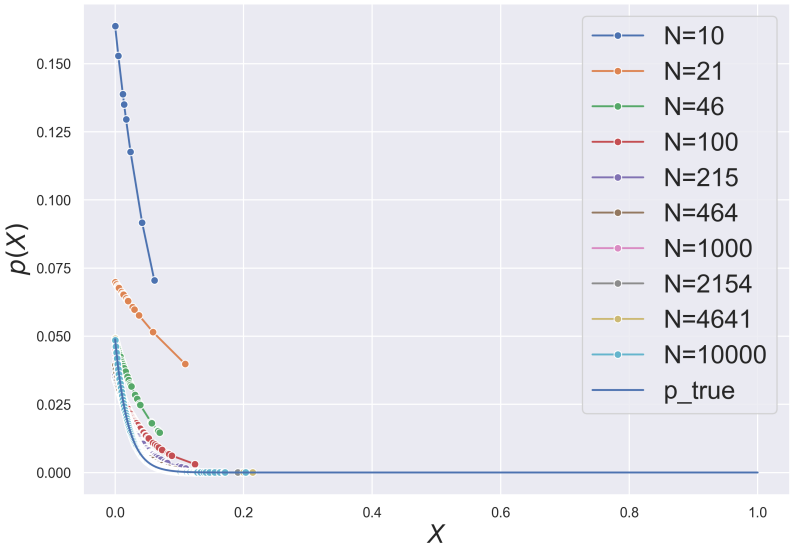
$X \in \{1,..n\}$ and

$i = 1,...,n$

Note that we add an equality constraint: the expectation must equal the sample mean

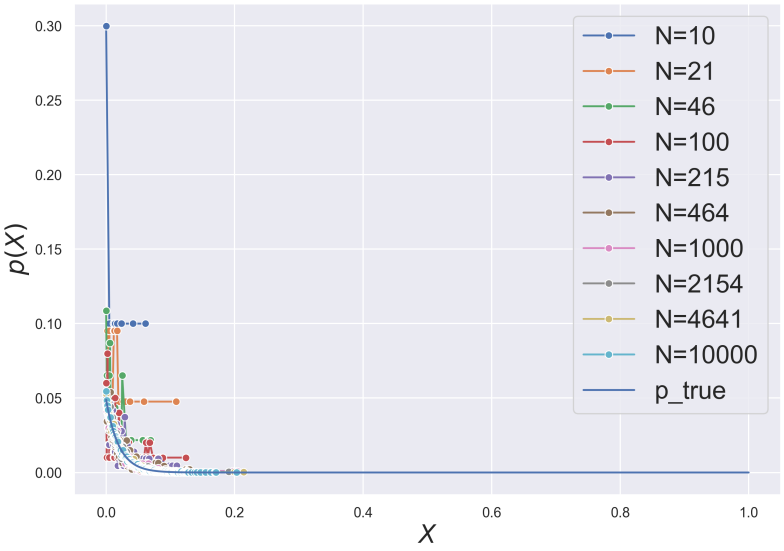
Maximum Entropy

Minimize $\sum_i^n p_i \log(p_i)$



Maximum Likelihood

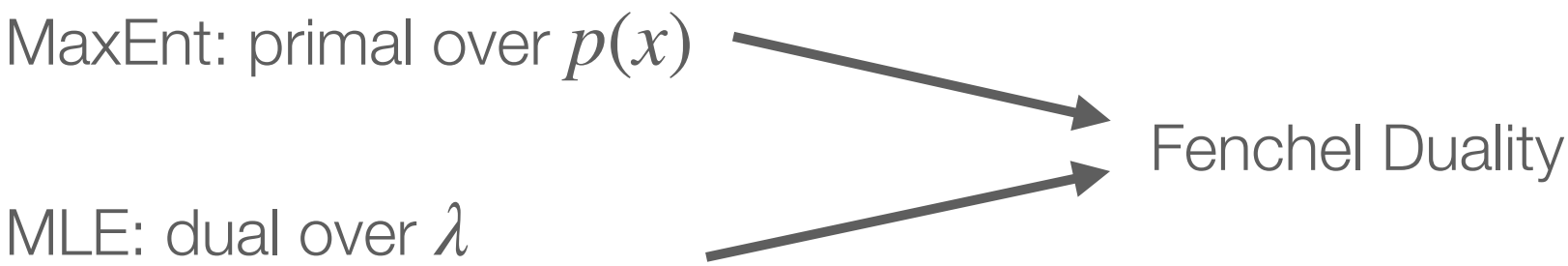
Maximize $\sum_i^n N_i \log(p_i)$



Drawing samples of increasing size from an exponential distribution.
Notice how each type of solution converges to the true distribution differently.

Maximum Entropy and Maximum Likelihood

Furthermore, the Maximum Entropy Principle and Maximum Likelihood Estimation have a primal dual relationship (within the family of exponential models).



[1] https://stats.stackexchange.com/questions/504117/is-there-a-relationship-between-maximum-likelihood-estimation-and-the-maximum-en (see answer)
[2]
Martin J. Wainwright and Michael I. Jordan (2008), "Graphical Models, Exponential Families, and Variational Inference", Foundations and Trends® in Machine Learning: Vol. 1: No. 1–2, pp 1-305. http://dx.doi.org/10.1561/2200000001

Maximum Entropy and Maximum Likelihood

MaxEnt: primal over $p(x)$

$$\max_{p(x)} -\int p(x)\log(p(x))dx$$

Subject to $$\int p(x)\phi(x)dx = \mu$$

$$\int p(x)dx = 1$$

$$p(x) \geq 1$$

Yields $$p(x) = \frac{1}{Z}\exp\left(-\sum_{i=1}^{k}\lambda_i f_i(x)\right)$$

MLE: dual over λ

For data $x_i \in \{x_1, \ldots, x_N\}$

from the primal distribution.

Plugging the primal distribution into the Lagrangian yields the dual problem:

$$\max_{\lambda} \sum_{i}^{N} p_\lambda(x_i)$$

$$= \max_{\lambda} \sum_{i}^{N} \left[\lambda^\intercal \phi(x_i) - \log(Z)\right]$$

Multinomial Logistic Regression

Exponential Models: Multinomial Logistic Regression

Exponential models are maximum entropy under equality constraints on expectations.

$$p(y \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_i \lambda_i f_i(\mathbf{x}, y) \right)$$

This is form of the **multinomial logistic regression** and **softmax.**

Multinomial Logistic Regression: MaxEnt and MLE

MaxEnt

$$\max_{p(x)} \sum_{y} p(y \mid \mathbf{x})]\log p(y \mid \mathbf{x})$$

$$= \mathbb{E}_p[-\log(p(y \mid \mathbf{x}))]$$

Subject to: $\mathbb{E}_p[\phi_k(\mathbf{x}, y)] = s_k$

Where $\phi_k(\mathbf{x}, y) = \mathbf{1}[y = c] \cdot x_i$

$$p(y \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_i \lambda_i f_i(\mathbf{x}, y)\right)$$

MLE

$$\max_{\lambda} \sum_{i}^{N} \log p(y^{(i)} \mid \mathbf{x}^{(i)})$$

Same as:

$$\min_{\lambda} \mathscr{L}_{\text{CE}} = \sum_{i}^{N} \sum_{k}^{K} \mathbf{1}[y^{(i)} = k]\log p(y = k \mid \mathbf{x}^{(i)})$$

(Minimization of cross entropy loss)

Energy-based Models

Energy-based Models

Energy-based models can be seen as a generalization of other models which use model distributions from the exponential family.

$$p(\mathbf{x}) = \frac{1}{Z(\theta)} \exp\left(-E_\theta(\mathbf{x})\right)$$

Where $Z(\theta) = \int \mathbf{exp}\left(-E_\theta(\mathbf{x})\right) d\mathbf{x}$ is called the **normalization** or **partition function**

Here, form of the energy function $E_\theta(\mathbf{x})$ determines the model type:

Logistic regression: $E_\theta(\mathbf{x}) = -\theta^\intercal \mathbf{x}$

Softmax classifier: $E_\theta(\mathbf{x}) = -\theta^\intercal \mathbf{x}$ for class y

Gaussian: $E_\theta(\mathbf{x}) = \frac{1}{2} || \mathbf{x} - \mu ||$

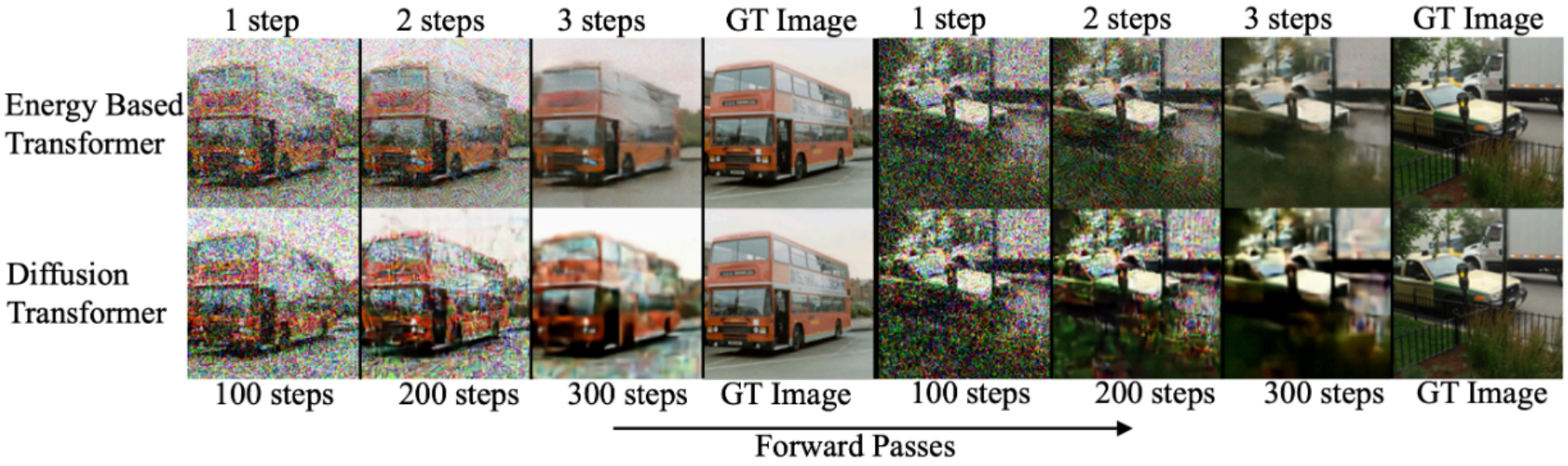
Boltzmann Machine: $E_\theta(\mathbf{x}) = -\mathbf{x}^\intercal W \mathbf{x} - b^\intercal \mathbf{x}$

Score-based models: $E_\theta(\mathbf{x}) = -\log p(\mathbf{x})$, (as recovered from $-\nabla_{\mathbf{x}} \log p(\mathbf{x})$)

[1] https://en.wikipedia.org/wiki/Energy-based_model
[2] "Your Classifier is Secretly an Energy Based Model and You Should Treat it Like One, Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, Kevin Swersky https://arxiv.org/abs/1912.03263

Energy-based Generative Models

Goal: learn a generative model $p_\theta(\mathbf{x})$, based on some real data distribution $p(\mathbf{x})$, so that $p_\theta(\mathbf{x})$ can be sampled to generate realistic synthetic data.

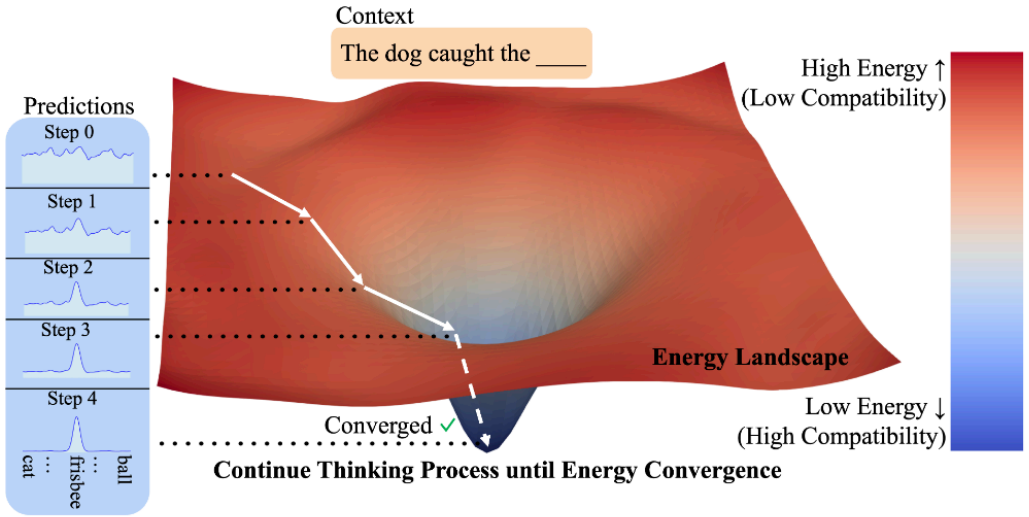


[1] https://en.wikipedia.org/wiki/Energy-based_model
[2] "Energy-based transformers are scalable learners and thinkers," A. Gladstone et al., 2025, https://arxiv.org/pdf/2507.02092

Energy-based Generative Models

One learns the generative mode $p_\theta(\mathbf{x})$ over a corpus of (complex) data via use of an energy function $E_\theta(\mathbf{x})$.

$$p_\theta(\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left(-E_\theta(\mathbf{x})\right)$$



The energy function $E_\theta(\mathbf{x})$ **is a trained neural network** that outputs a scalar value that encodes the negative log likelihood of the data $\mathbf{x}$.

High energy is assigned to unlikely samples, and low energy to likely samples.

[1] https://en.wikipedia.org/wiki/Energy-based_model
[2] "Energy-based transformers are scalable learners and thinkers," A. Gladstone et al., 2025, https://arxiv.org/pdf/2507.02092

Energy-based Generative Models

1. Sample from the real data

Here, one is sampling N times from x, each sample is $x^{(i)}$

$$\mathbf{x}^{(i)}_{\text{real}} \sim p_{\text{real}}(\mathbf{x})$$

2. Sample from the model data

For the model data, MCMC sampling is used, since the distribution is not normalized (Z cannot be computed).

$$\mathbf{x}^{(i)}_{\text{model}} \sim p_{\theta}(\mathbf{x})$$

3. Calculate the contrastive loss

$$\mathscr{L}(\theta) = E_{\theta}(\mathbf{x}_{\text{model}}) - E_{\theta}(\mathbf{x}_{\text{real}})$$

4. Update $E_{\theta}(\mathbf{x})$ via backpropagation

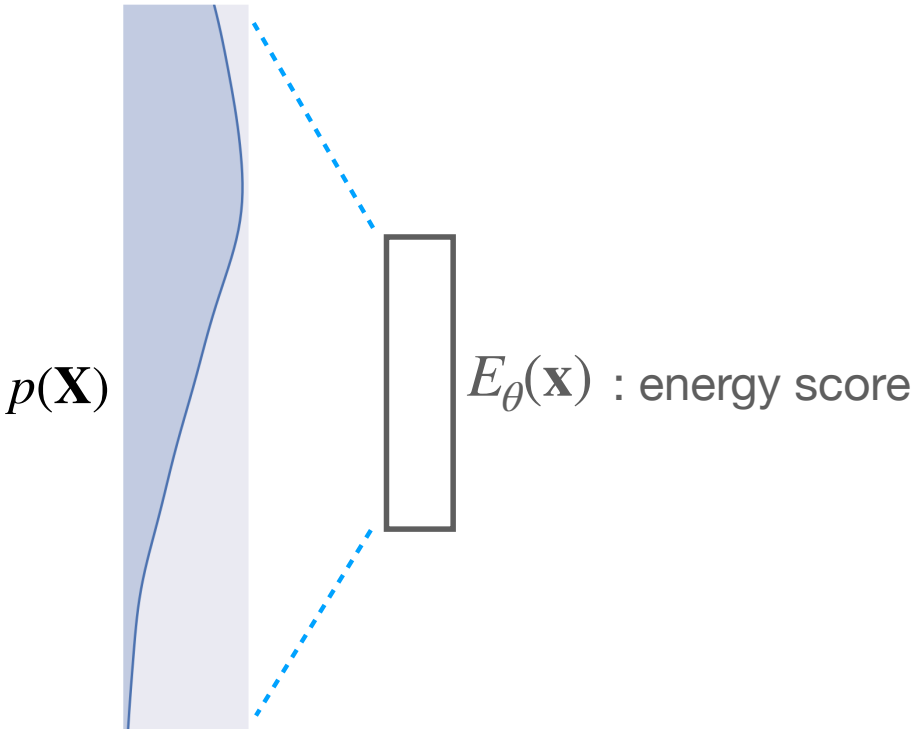
[1] https://en.wikipedia.org/wiki/Energy-based_model

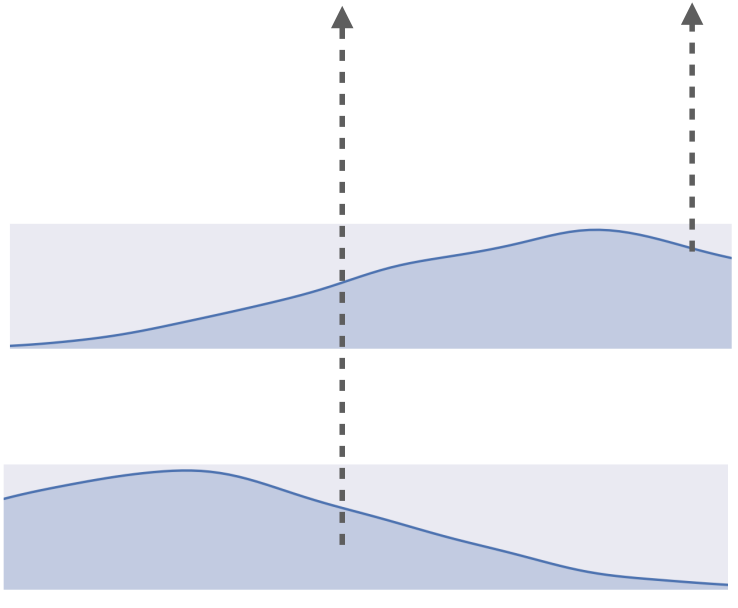
Energy-based Generative Models

$$p_\theta(\mathbf{x}) = \frac{1}{Z_\theta(\mathbf{x})} \exp\left(-E_\theta(\mathbf{x})\right)$$

Contrastive Loss function to Train the Generative Model

$$\mathscr{L}(\theta) = E_\theta(\mathbf{x}_{\text{model}}) - E_\theta(\mathbf{x}_{\text{real}})$$





[1] https://en.wikipedia.org/wiki/Energy-based_model

Energy-based Generative Models

Process for training $E_\theta(\mathbf{x})$ via Backpropagation

1. Define training objective: log likelihood

$$\mathscr{L}(\theta) = \sum_i^N \log p_\theta(x^{(i)}) = -\sum_i^N E_\theta(x^{(i)}) - N\log\left(Z(\theta)\right)$$

Here, one is sampling N times from x, each sample is $x^{(i)}$

2. Differentiate loss to get rid of partition function

$$\mathscr{L}(\theta) = \sum_i^N \log p_\theta(x^{(i)}) = -\sum_i^N E_\theta(x^{(i)}) - N\log\left(Z(\theta)\right)$$

Intractable partition function

$$\nabla_\theta \mathscr{L}(\theta) = -\sum_i^N \nabla_\theta E_\theta(x^{(i)}) + N\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right]$$

Energy-based Models in Generative Modeling

Process (simplified)

3. Approximate $\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right]$

$$\nabla_\theta \mathcal{L}(\theta) = -\sum_i^N \nabla_\theta E_\theta(x^{(i)}) + N\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right]$$

Requires expectation of $p_\theta(x)$ over x

$\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right]$ requires sampling from $p_\theta(x)$, which cannot be done directly.

Common methods used for approximation:

- MCMC/Langevin
- Replay buffer
- Contrastive divergence
- Score-based methods

Energy-based Models in Generative Modeling

Process (simplified)

4. Equivalence to Contrastive Loss

$$\nabla_\theta \mathcal{L}(\theta) = -\sum_i^N \nabla_\theta E_\theta(x^{(i)}) + N\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right]$$

$$N\mathbb{E}_{p_\theta(x)}\left[\nabla_\theta E_\theta(x)\right] \approx \nabla_\theta E_\theta(\mathbf{x}_{\text{model}})$$

$$\sum_i^N \nabla_\theta E_\theta(x^{(i)}) \approx \nabla_\theta E_\theta(\mathbf{x}_{\text{real}})$$

$$\rightarrow \quad \nabla_\theta \mathcal{L}(\theta) \approx \nabla_\theta E_\theta(\mathbf{x}_{\text{model}}) - \nabla_\theta E_\theta(\mathbf{x}_{\text{real}})$$

5. Update $E_\theta(x)$ based on new θ

Energy-based Models

- Energy-based models are a generalization of exponential models, which are maximum entropy models.

- Energy-based generative models use the same formalism in a generative context.

- Energy-based generative models relative to GAN, diffusion, and transformer-based models :
  - Can model complex, multi-modal latent spaces.
  - Better out-of-distribution detection (e.g. fewer hallucinations)
  - Several key disadvantages:
    - Computational complexity: sampling, expectation calculation, and optimization steps.

[1] https://en.wikipedia.org/wiki/Energy-based_model

Summary

Summary

- Entropy developed as a concept in physics.
- Maximization of entropy under expectation constraints produces models of the exponential family, which are foundational to machine learning and statistical inference.
- Entropy maximization has a primal-dual relationship with maximum likelihood
- Energy-based models generalize ML classifiers based on the exponential distribution family, and can be used generatively.