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

Bayesian Formalism

Bayesian vs. Frequentist Philosophies

Frequentist approach:
Probability is a *proportion* of events in an *infinite number of trials.*

Bayesian approach:
Probability is a *distribution* over events for *one trial.*

In this sense, for a given inference, the Bayesian approach quantifies the *uncertainty* in the inference.

[1] Sarkka, Section 2.1, p.178

Frequentist Formalism

In the frequentist formalism, the data (comprising both features and labels $\mathscr{D} = \{(\mathbf{X}, \mathbf{y})_i\}$,

is used to find a distribution (model) $p(\mathbf{y} \,|\, \mathbf{X}, \theta)$ by minimizing some loss $\mathscr{L}(\mathbf{X}, \mathbf{y}, \theta)$.

Here, θ are **fixed but unknown parameters found via optimization.**

Data	**Model**	**Inference**
$\mathscr{D} = \{(\mathbf{X}, \mathbf{y})_i\}$	$p(\mathbf{y} \,\vert\, \mathbf{X}, \theta)$	$\mathbb{E}_p[\mathbf{y} \,\vert\, \mathbf{X}, \theta^*]$

Trained by finding

$$\theta^* = \text{argmin}\ \mathscr{L}(\mathbf{X}, \mathbf{y}, \theta)$$

Expectation is the predicted summary statistic of the conditional distribution for the case of an MSE loss function.

Bayesian Formalism

In the Bayesian formalism, the data (comprising both features and labels $\mathscr{D} = \{(\mathbf{X}, \mathbf{y})_i\}$), is generated by an underlying **generative model,** parameterized by θ, which defines a joint distribution $p(\mathscr{D}, \theta)$. The key point is that θ is treated as a **random variable**.

The goal is to calculate the **posterior distribution** of the parameters θ: $p(\theta \,|\, \mathscr{D})$

Data	**Model**	**Inference**
$\mathscr{D} = \{(\mathbf{X}, \mathbf{y})_i\}$	$p(\mathscr{D}, \theta)$	$p(\mathscr{D}_{\textbf{new}} \,\vert\, \mathscr{D}) = \int p(\mathscr{D}_{\textbf{new}} \,\vert\, \theta) p(\theta \,\vert\, \mathscr{D}) d\theta$

Find the posterior

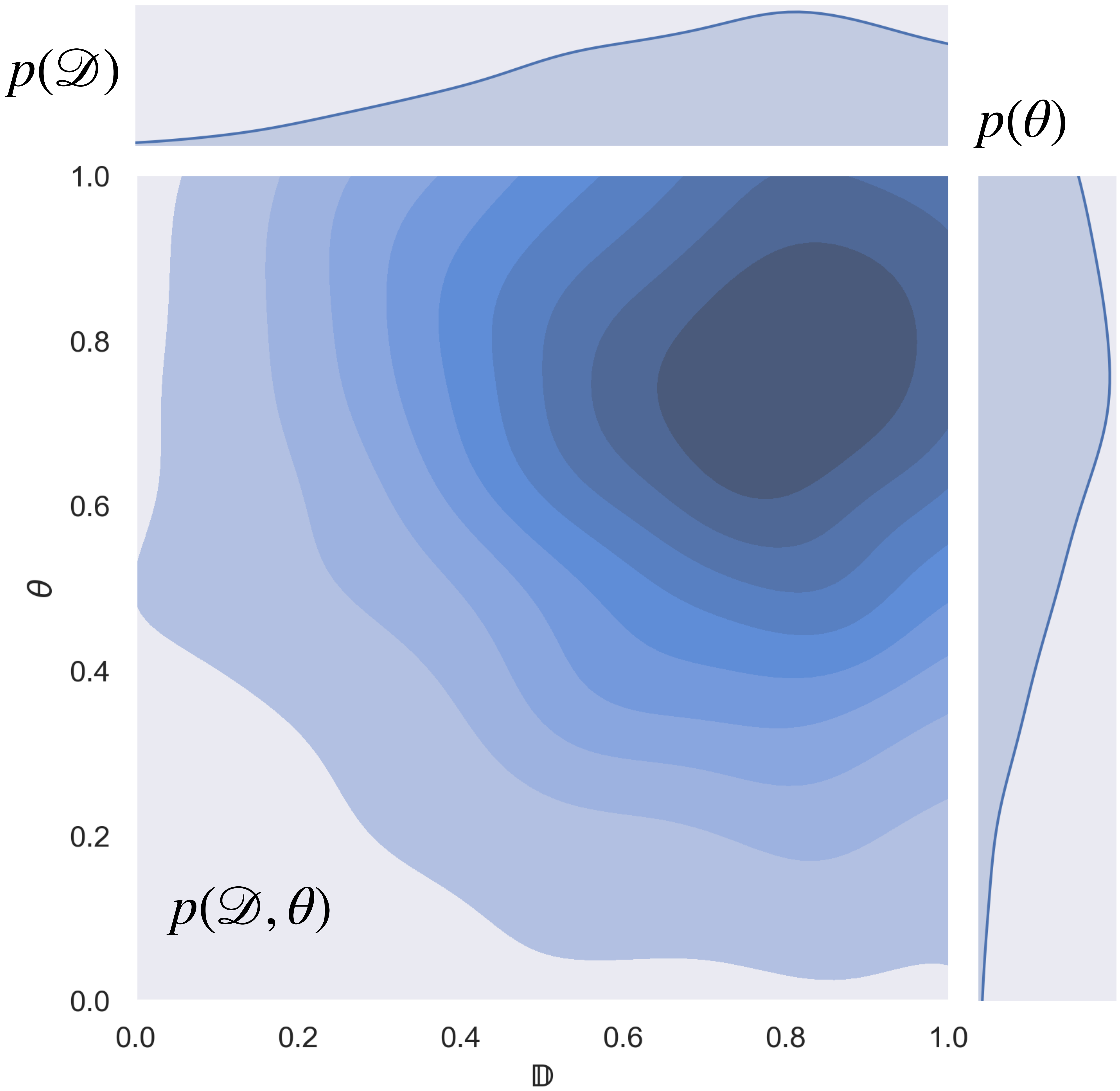
→ $p(\theta \,|\, \mathscr{D})$

Bayes Rule:

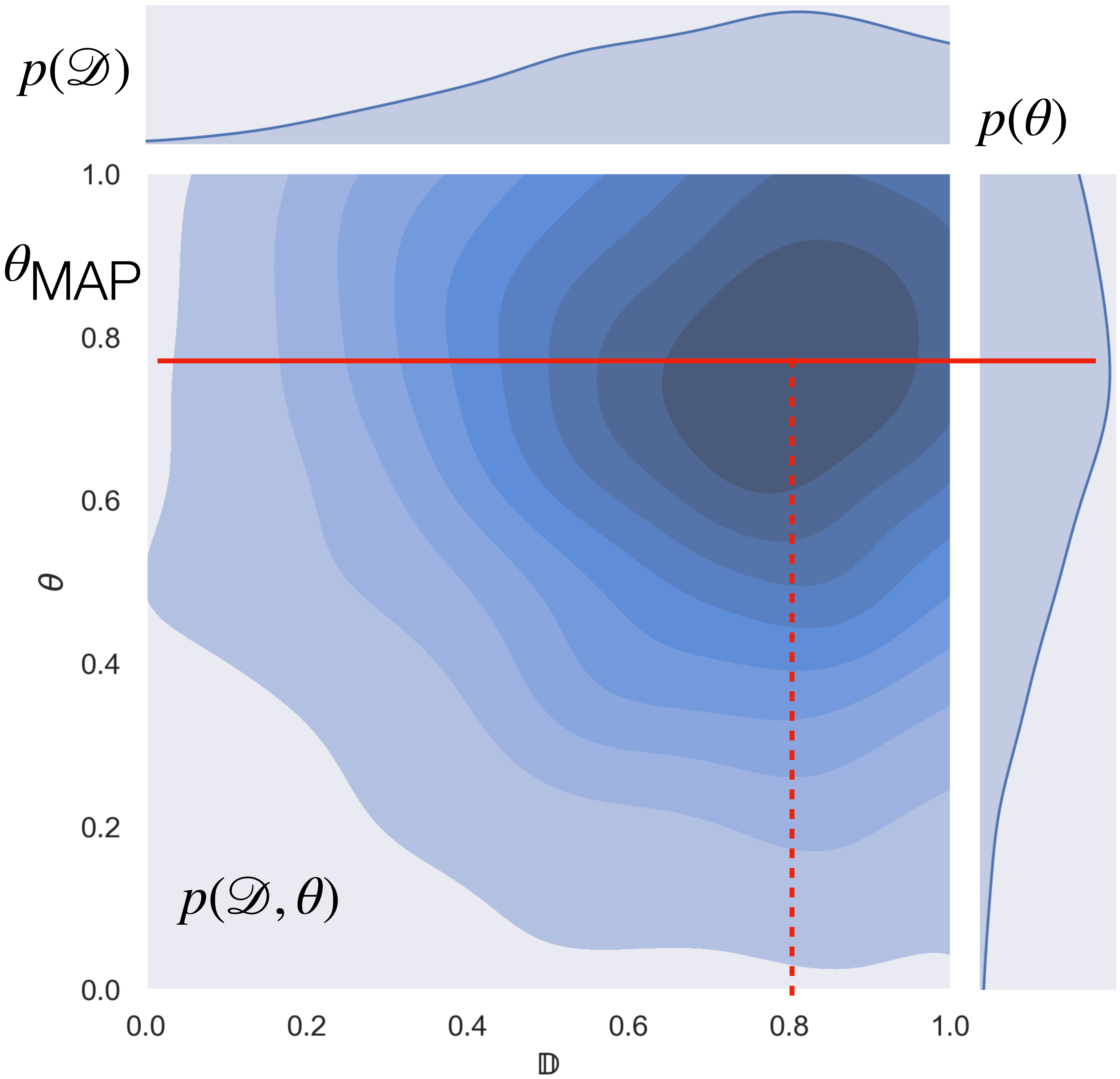
$$p(\theta \,|\, \mathscr{D}) = \frac{p(\mathscr{D} \,|\, \theta) p(\theta)}{p(\mathscr{D})}$$

[1] Murphy, Section, 7.6, p.232
[2] Sarkka , Section 3.1, p.27

Bayesian Generative Model



Bayesian Generative Model



The maximum a posteriori (MAP) estimate of θ is the highest probability estimate of θ given the data $\mathscr{D}$:

$$\theta_{\text{MAP}} = \text{argmax}\ p(\theta|\mathscr{D})$$

One is just performing an optimization of the posterior $p(\theta|\mathscr{D})$.

Why use Bayesian instead of Frequentist ML?

	Bayesian	**Frequentist**
Data	limited	large
Inference	distributions, CIs	point prediction
Implementation	• computationally intensive • requires up-front derivation and modeling	fast, scalable
Other	• prior-based regularization (prevents overfitting) • supports decision-making under uncertainty	

Energy Functions and Functionals

Energy Function in Physics

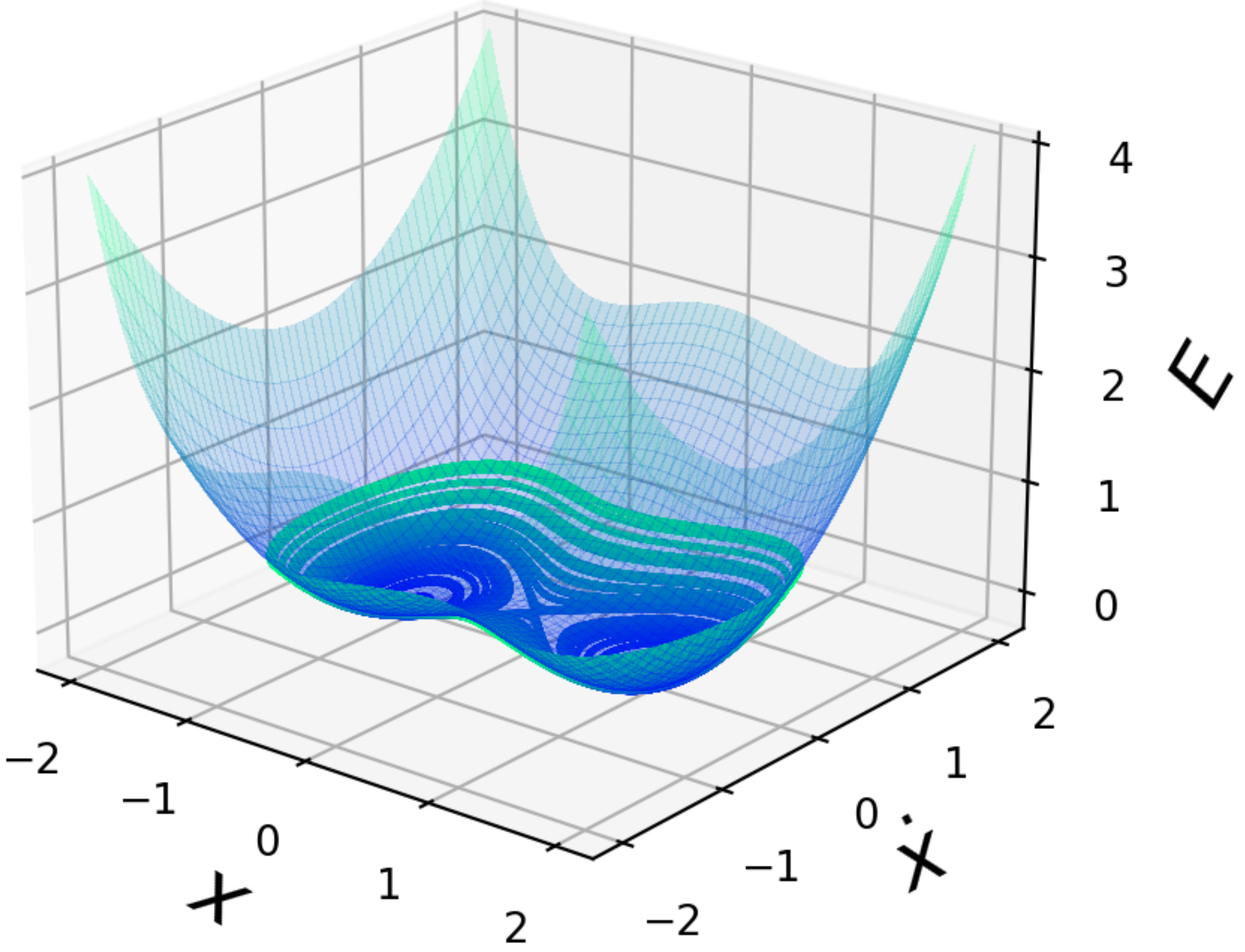
The term **energy function** is borrowed from physics, where potential functions represent the potential energy of of the system.

Given dynamics

$$\dot{\mathbf{x}} = f(\mathbf{x})$$

the potential function $V(\mathbf{x})$ is defined as

$$\nabla V(\mathbf{x}) = -f(\mathbf{x})$$



[1] https://en.wikipedia.org/wiki/Principle_of_minimum_energy

Maximum Entropy Duality to Minimum Energy

In physics and Bayesian statistics, one defines a **free energy functional** that maps a distribution to a free energy:

$$\mathscr{F}[p] = \mathbb{E}_p[E(\theta)] - \mathbb{H}[p]$$

Where $E(\theta)$ is an **energy function** and $\mathbb{H}[p]$ is an **entropy.** p is the distribution.

It follows that minimizing this functional is achieved by either:

1. **Maximizing the entropy** under an **energy constraint:** the energy constraints are moment constraints on p.
2. **Minimizing the energy** under an **entropy constraint:** an equality or inequality constraint on the entropy.

[1] https://en.wikipedia.org/wiki/Principle_of_minimum_energy

Energy Functional in Variational Inference

Variational Inference

Variational inference is a subset of Bayesian inference where inference is posed as an *optimization* problem.

The posterior distribution is to be found

$$p(\theta \,|\, \mathscr{D}) = \frac{p(\mathscr{D} \,|\, \theta)p(\theta)}{p(\mathscr{D})} \approx p(\mathscr{D} \,|\, \theta)p(\theta)$$

A problem here:

$p(\mathscr{D} \,|\, \theta)p(\theta)$ is often intractable (i.e. it is difficult to sample from) because the likelihood $p(\mathscr{D} \,|\, \theta)$ is **non-conjugate** with the prior $p(\theta)$
(i.e. they do not belong to the same distribution family).

[1] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, 2012, Section 21.2

Variational Inference

$$p(\theta \,|\, \mathscr{D}) = \frac{p(\mathscr{D} \,|\, \theta)p(\theta)}{p(\mathscr{D})} \approx p(\mathscr{D} \,|\, \theta)p(\theta)$$

Solution:

Set the prior $p(\theta)$.
Then choose an approximative **variational distribution** $q(\theta \,|\, \mathscr{D})$ (conjugate to the prior) to approximate the true posterior $p(\theta \,|\, \mathscr{D})$**,** and **optimize a functional** that closes the distance between $q(\theta \,|\, \mathscr{D})$ and $p(\theta \,|\, \mathscr{D})$

[1] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, 2012, Section 21.2

Energy Functional and Entropy Maximization

A functional is an object that maps functions to scalars.

Define the **energy functional** as the distance between $q(\theta|\mathscr{D})$ and $p(\theta|\mathscr{D})$**:**

$$\mathscr{L}(q) = \mathbb{KL}(q(\theta|\mathscr{D}), p(\theta|\mathscr{D})Z)$$

$$= \mathbb{KL}(q(\theta|\mathscr{D}), p(\theta|\mathscr{D})) - \log Z$$

Here $Z = p(\mathscr{D})$, and is called the normalization, or evidence. It is generally treated as a proportionality constant.

Let $E(\theta) = -\log p(\theta|\mathscr{D})$ ***.** Then:

$$\mathscr{L}(q) = \underbrace{\mathbb{E}_{q(\theta|\mathscr{D})}\left[-\log p(\theta|\mathscr{D})\right]}_{\text{Energy (accuracy)}} - \underbrace{H(q(\theta|\mathscr{D}))}_{\text{Entropy (dispersion)}} + C$$

The functional $\mathscr{L}(q)$ is thus minimized by **maximizing the entropy** $H(q(\theta|\mathscr{D}))$ **under an energy constraint**

[1] K.P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge, 2012, Section 21.2 p.732

[2] Koller and Friedman, *Probabilistic Graphical Models: Principles and Techniques,* 2009

* Can use negative log joint or negative log posterior forms to derive corresponding forms of the energy functional

Variational Inference for Bayesian Logistic Regression

Setup

Data:

$\mathscr{D} = \{(x_i, y_i)\}_{i=1}^{n}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{0,1\}$

Likelihood:

$$p(y_i \mid x_i, \theta) = \sigma\left(x_i^\intercal, \theta\right)^{y_i}\left(1 - \sigma(x_i^\intercal, \theta)\right)^{1-y_i}$$

Prior:

$$p(\theta) = \mathcal{N}(0, \lambda^{-1} I)$$

Posterior (to be found):

$$p(\theta \mid X, y) \propto p(y \mid X, \theta)p(\theta) = p(\theta)\prod_{i=1}^{n} p(y_i \mid x_i, \theta)$$

Bayes Rule:

$$p(\theta \mid X, y) = \frac{p(y \mid X, \theta)p(\theta)}{p(X, y)}$$

Where:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

*This distribution is intractable due to the **non-conjugacy** between the logistic likelihood and the Gaussian prior.*

(Log posterior is not quadratic in θ due to the logistic function, which means the posterior is not Gaussian and therefore not conjugate with the prior (i.e. posterior and prior do not belong to the same distribution family.)

[2] Murphy, Chapter 21

Variational Inference for Bayesian Logistic Regression

Process

1. Approximate the Posterior with a Variational Distribution

$$p(\theta|X,y) \;\Longrightarrow\; q(\theta|X,y) = \mathcal{N}(\theta|\mu,\Sigma)$$

2. Define the Energy Functional/ELBO

$$\mathscr{L}(q) = \mathbb{E}_{q(\theta|X,y)}\left[\underbrace{\log p(y|X,\theta)}_{\textit{Intractable log likelihood}} - \mathbb{KL}(q(\theta|X,y), p(\theta|X,y))\right]$$

$$\Longrightarrow\; \mathscr{L}(\mu,\Sigma) = \sum_{i=1}^{n} \mathbb{E}_{q(\theta|X,y)}\left[\underbrace{\log \sigma\left(x_i^{\intercal},\theta\right)^{y_i}\left(1-\sigma(x_i^{\intercal},\theta)\right)^{1-y_i}}_{\textit{Intractable log likelihood}} - \mathbb{KL}(q(\theta|X,y), p(\theta|X,y))\right]$$

[2] Murphy, Chapter 21

Variational Inference for Bayesian Logistic Regression

Process

3. Estimate the Expected Log Likelihood

The expected log-likelihood is intractable:

$$\mathbb{E}_{q(\theta|X,y)}\left[\log p(y\,|\,X,\theta)\right]$$

It is approximated using
Monte Carlo sampling (i.e. the black-box variational inference approach):

$$\mathbb{E}_{q(\theta|X,y)}\left[\log p(y\,|\,X,\theta)\right] \approx \frac{1}{S}\sum_{s=1}^{S}\log p(y\,|\,X,\theta^{(s)})$$

[2] Murphy, Chapter 21

Variational Inference for Bayesian Logistic Regression

Process

4. Find the KL Divergence between Gaussians

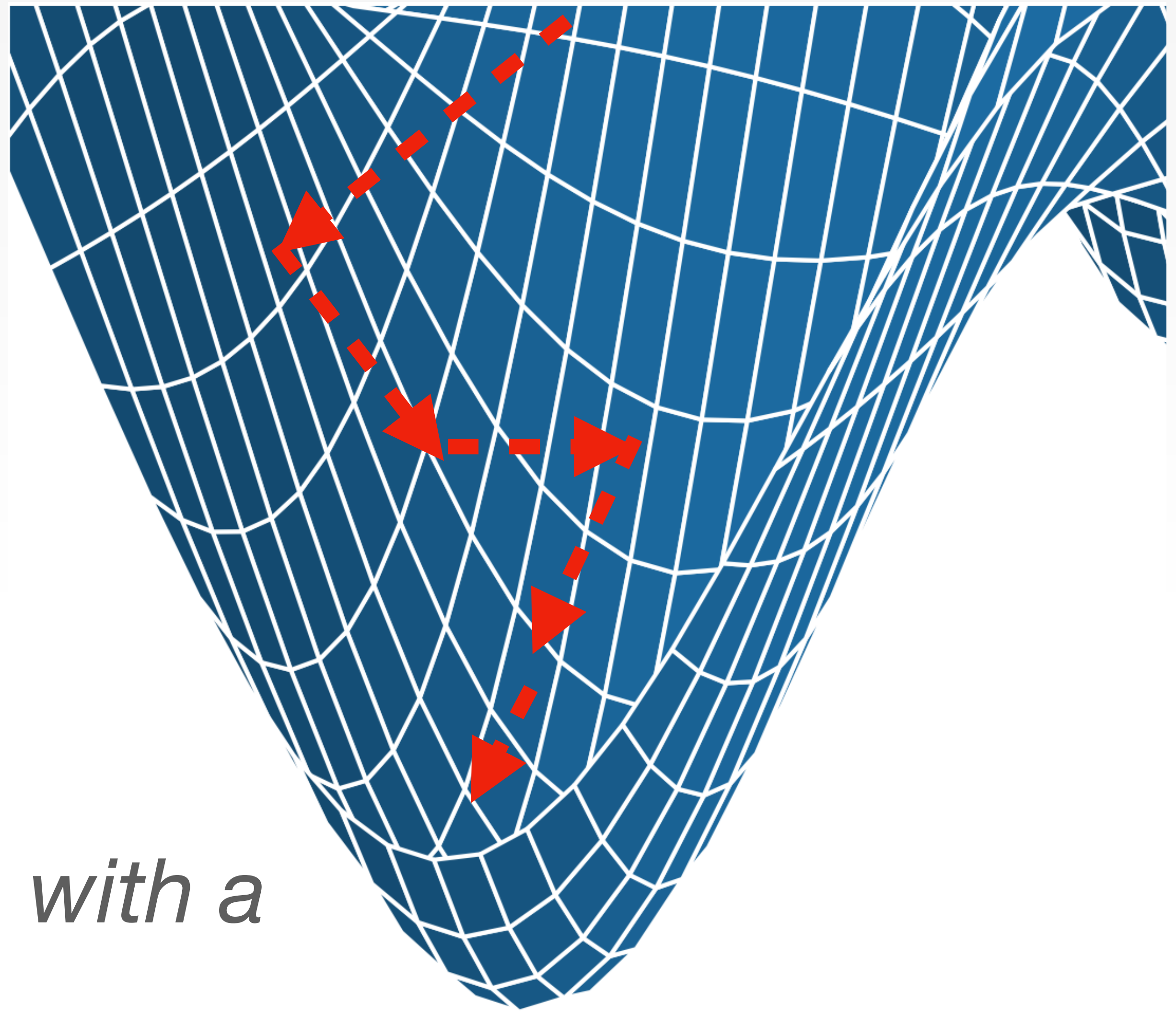
$$\mathscr{L}(\mu, \Sigma) = \sum_{i=1}^{n} \mathbb{E}_{q(\theta|X,y)} \left[\underline{\log p(y|X,\theta)} - \underline{\mathbb{KL}(q(\theta|X,y), p(\theta|X,y))}\right]$$

$$\boxed{\mathbb{E}_{q(\theta|X,y)} \left[\log p(y|X,\theta)\right] \approx \frac{1}{S} \sum_{s=1}^{S} \log p(y|X,\theta^{(s)})}$$

$$\boxed{\mathbb{KL}(q(\theta|X,y), p(\theta|X,y)) = \frac{1}{2} \left[\text{Tr}\,(\lambda\Sigma) + \lambda\mu^{\intercal}\mu - d + \log\frac{\lambda^{-1}I}{|\Sigma|}\right]}$$

5. Optimize the ELBO via Gradient Methods

The ELBO/energy functional $\mathscr{L}(\mu, \Sigma)$ is iteratively optimized to find $q(\theta|X,y)$ via $(\mu^{opt}, \Sigma^{opt})$



Methods:

- Stochastic gradient descent
- ADAM

*Here one obtains the **global optimum** with a higher quality posterior (c.f. the Laplace approximation)*

[2] Murphy, Chapter 21

Variational Inference for Bayesian Logistic Regression

Inference

If one has the posterior distribution, one has a distribution over the model parameters. To perform inference, one must form the **posterior predictive distribution,** which is a distribution over y for some new x using $q(\theta \mid X, y) = \mathcal{N}(\theta \mid \mu^{opt}, \Sigma^{opt})$.

posterior predictive distribution:

$$p(y^* = 1 \mid x^*) = \int \sigma(\theta^\intercal x^*) \mathcal{N}(\theta \mid \mu^{opt}, \Sigma^{opt}) d\theta$$

[2] Murphy, Section 8.41, p.255

Variational Inference for Bayesian Logistic Regression

Some notes:

- The variational inference approach is advantageous when one is working with limited data and one seeks to model uncertainty.

- It produces globally optimal estimates of the model parameters with higher-quality posterior distributions.

- Several key disadvantages:
    - Computational complexity: sampling, expectation calculation, and optimization steps.
    - Modeling complexity: Have to choose variational distribution family, prior distribution

Summary

Summary

- Entropy developed as a concept in physics and was foundational to information theory.
- Maximization of entropy under expectation constraints produces models of the exponential family, which are foundational to machine learning and statistical inference.
- The energy function is a generalized function on the model likelihood, and is used in Bayesian models, Bayesian variational inference, and generative models (e.g. energy-based models).
    - Bayesian variational inference is useful in situations with limited data and modeling uncertainty
    - Energy-based methods are useful in generative modeling of complex latent spaces.
    - Both methods have high computational complexity due to sampling during training.