Lecture 1:

Entropy and Information

Markus Hohle

University California, Berkeley

Bayesian Data Analysis and Machine Learning for Physical Sciences

Course Map

Module 1	Maximum Entropy and Information, Bayes Theorem
Module 2	Naive Bayes, Bayesian Parameter Estimation, MAP
Module 3	Model selection: Comparing Distributions vs Frequentist Methods
Module 4	Model Selection: Bayesian Signal Detection
Module 5	Variational Bayes, Expectation Maximization
Module 6	Stochastic Processes
Module 7	Monte Carlo Methods
Module 8	Markov Models, Graphs
Module 9	Machine Learning Overview, Supervised Methods
Module 10	Unsupervised Methods
Module 11	ANN: Perceptron, Backpropagation
Module 12	ANN: Basic Architecture, Regression vs Classification, Backpropagation again
Module 13	Convolution and Image Classification and Segmentation
Module 14	TBD (GNNs)
Module 15	TBD (RNNs and LSTMs)
Module 16	TBD (Transformer and LLMs)

Course Map

Module 1	**Maximum Entropy and Information, Bayes Theorem**
Module 2	Naive Bayes, Bayesian Parameter Estimation, MAP
Module 3	Model selection: Comparing Distributions vs Frequentist Methods
Module 4	Model Selection: Bayesian Signal Detection
Module 5	Variational Bayes, Expectation Maximization
Module 6	Stochastic Processes
Module 7	Monte Carlo Methods
Module 8	Markov Models, Graphs
Module 9	Machine Learning Overview, Supervised Methods
Module 10	Unsupervised Methods
Module 11	ANN: Perceptron, Backpropagation
Module 12	ANN: Basic Architecture, Regression vs Classification, Backpropagation again
Module 13	Convolution and Image Classification and Segmentation
Module 14	TBD (GNNs)
Module 15	TBD (RNNs and LSTMs)
Module 16	TBD (Transformer and LLMs)

Maximum Entropy and Information, Bayes Theorem

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

Maximum Entropy and Information, Bayes Theorem

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

Maximum Entropy and Information, Bayes Theorem:

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$p(x_i)$: probability of event x_i

↓

$h(x_i)$: function that measures "information"

$$\log[p(x_i)\, p(x_j)] = \log[p(x_i)] + \log[p(x_j)]$$

1) $h(x_i)$ should be additive

2) $h(x_i)$ should be monotonic

3) if x_i and x_j are independent, then

$$h(x_i, x_j) = h(x_i) + h(x_j)$$

$$p(x_i, x_j) = p(x_i)p(x_j)$$

Maximum Entropy and Information, Bayes Theorem:

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$p(x_i)$: probability of event x_i

$h(x_i)$: function that measures "information"

1) $h(x_i)$ should be additive $\quad h(x_i, x_j) = h(x_i) + h(x_j)$

2) $h(x_i)$ should be monotonic

3) if x_i and x_j are independent, then $\quad p(x_i, x_j) = p(x_i)p(x_j)$

$$\log[p(x_i)\,p(x_j)] = \log[p(x_i)] + \log[p(x_j)] \qquad p(x_i) \leq 1$$

$h(x_i) = -\log[p(x_i)]$ information is ***positive***
low $p(x_i)$ → "great surprise"

definition
conditional Entropy
KL divergence
connection to TD

idea: gain of information = **"degree of surprise"**
if something happens that we expected already → no surprise
→ no information

$$h(x_i) = -\log[p(x_i)]$$

$\boldsymbol{p(x_i)}$: **probability of event $\boldsymbol{x_i}$**

The event x_i is randomly drawn from $p(x_i)$ → ***average*** amount of information

Entropy S

$$S = -\sum_{i=1}^{I} p(x_i)\log[p(x_i)] \qquad \text{(discrete)}$$

$$S = -\int p(x)\log[p(x)]\,dx \qquad \text{(continuous or } \textit{differential}\text{)}$$

note:
- the **base** of log is **arbitrary** (often 2 or e)
- $\lim_{p\to 0}(p\log p) = 0$
- S is large → no information
- S is zero → all information
- continuous entropy **can** be negative
- continuous entropy is **not** exactly equivalent to discrete entropy (see LDDP)

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

Maximum Entropy and Information, Bayes Theorem:

Entropy S

$$S = -\sum_{i=1}^{I} p(x_i)\log[p(x_i)] \qquad \text{(discrete)}$$

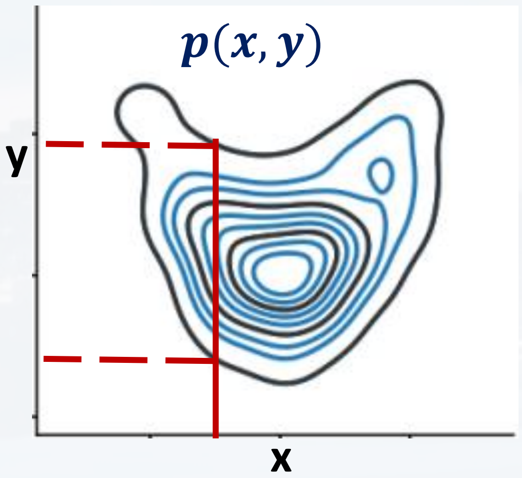
$$S = -\int p(x)\log[p(x)]\,dx \qquad \text{(continuous)}$$

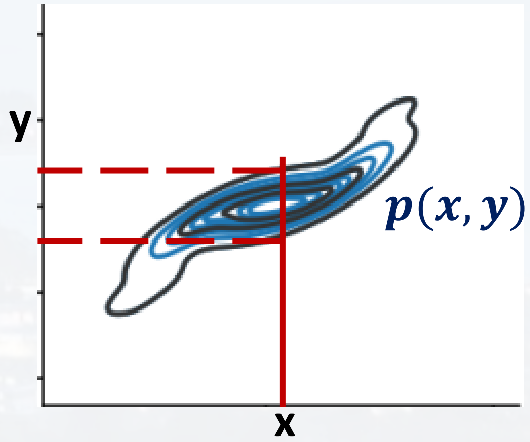
$\boldsymbol{p(x_i)}$: **probability of event $\boldsymbol{x_i}$**

joint distribution $p(x, y)$, what is

$S(y|x)$: entropy of *y*, **"given"** *x* (we know *x* already)

"given"





Maximum Entropy and Information, Bayes Theorem:

Entropy S

$$\boxed{S = -\int p(x)\log[p(x)]\,dx}$$

$p(x_i)$: probability of event x_i

joint distribution $p(x,y)$, what is

$S(y|x)$: entropy of *y*, **"given"** *x* (we know *x* already)

$$S[p(x,y)] = -\iint p(x,y)\, log[p(x,y)]\, dxdy$$

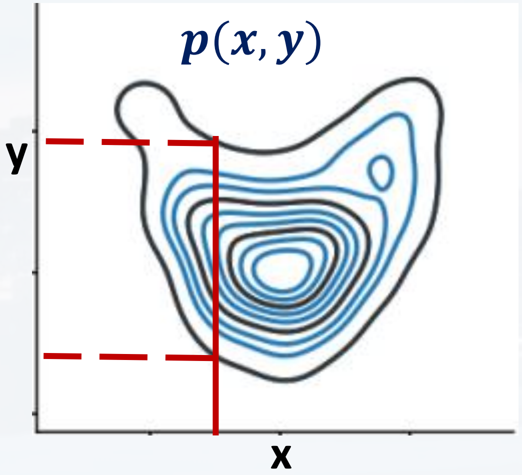
$$= -\iint p(x,y)\, log[\boldsymbol{p(y|x)p(x)}]\, dxdy$$

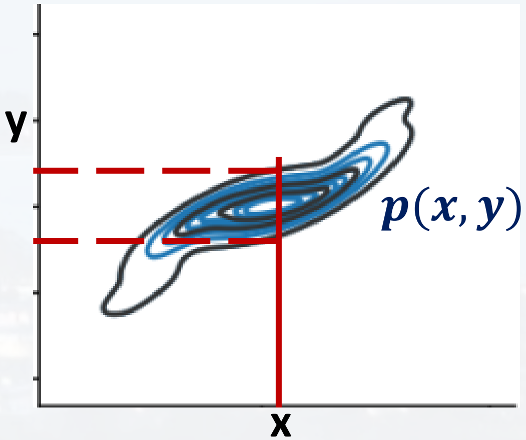
conditional probabilities (see later)

$$= -\iint p(x,y)\, \{log[\boldsymbol{p(y|x)}] + log[\boldsymbol{p(x)}]\}\, dxdy$$

$$= -\iint p(x,y)\, log[\boldsymbol{p(y|x)}]\, dx\, dy - \iint p(x,y)\, log[\boldsymbol{p(x)}]\, dxdy$$

we still sample from $p(x,y)$





Entropy S

$$\boxed{S = -\int p(x)\log[p(x)]\,dx}$$

$\boldsymbol{p(x_i)}$: **probability of event x_i**

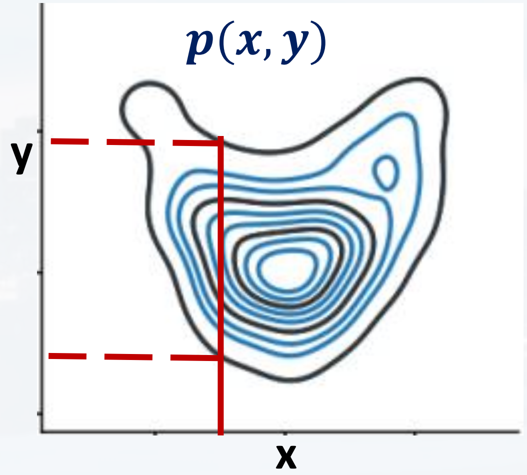
joint distribution $p(x, y)$, what is

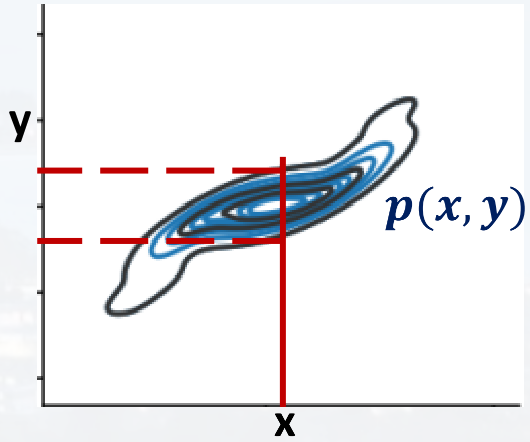
$S(y|x)$: entropy of *y*, **"given"** *x* (we know *x* already)

$$S[p(x,y)] = -\iint p(x,y)\, log[\mathbf{p(y|x)}]\, dx\, dy \; - \iint p(x,y)\, log[\mathbf{p(x)}]\; dxdy$$

$$S[p(x,y)] = S[p(y|x)] + S[p(x)]$$

differential entropy of $\boldsymbol{p(x,y)}$ | **conditional entropy** | **differential entropy of $\boldsymbol{p(x)}$**





We can learn about $p(x, y)$ when we have information about *x* first and then *y*, given *x*

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

often, we (need to) approximate $\boldsymbol{p}(\boldsymbol{x})$ by some other distribution $\boldsymbol{q}(\boldsymbol{x})$

How much is our information "off" if we work with the approximation $\boldsymbol{q}(\boldsymbol{x})$?

$$-\int p(x)\log[\boldsymbol{q}(\boldsymbol{x})]\,dx - \left[-\int p(x)\log[\boldsymbol{p}(\boldsymbol{x})]\,dx\right] = -\int p(x)\log\left[\frac{\boldsymbol{q}(\boldsymbol{x})}{\boldsymbol{p}(\boldsymbol{x})}\right]dx = KL(p||q)$$

KL or ***Kullback-Leiber divergence***

It is **not** a distance! $KL(p||q) \neq KL(q||p)$

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

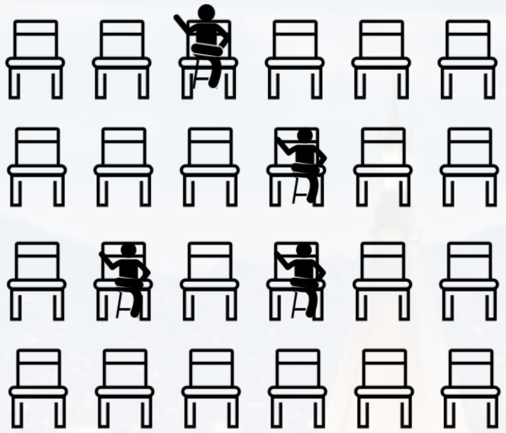
- Lagrangian Multiplier
- examples

Bayes Theorem

definition
conditional Entropy
KL divergence
connection to TD

"Shannon" Entropy S

$$S = -\sum_{i=1}^{I} p(x_i)\log[p(x_i)]$$



I = 2 states

N: number of indistinguishable particles
n_i: number of particles in micro state i
I: number of states

multiplicity Ω : number of macro states

$$\Omega = \frac{N!}{n_1!\,(N-n_1)!} = \frac{N!}{n_1!\,n_2!}$$

for I states $\quad \Omega = \dfrac{N!}{\prod_{i=1}^{I} n_i}$

for large N: $\lim_{N\to\infty}\left(\frac{n_i}{N}\right) = p_i$

p_i: probability of a particle being in micro state i

Maximum Entropy and Information, Bayes Theorem:

definition
conditional Entropy
KL divergence
connection to TD

"Shannon" Entropy S

$$S = -\sum_{i=1}^{I} p(x_i)\log[p(x_i)]$$

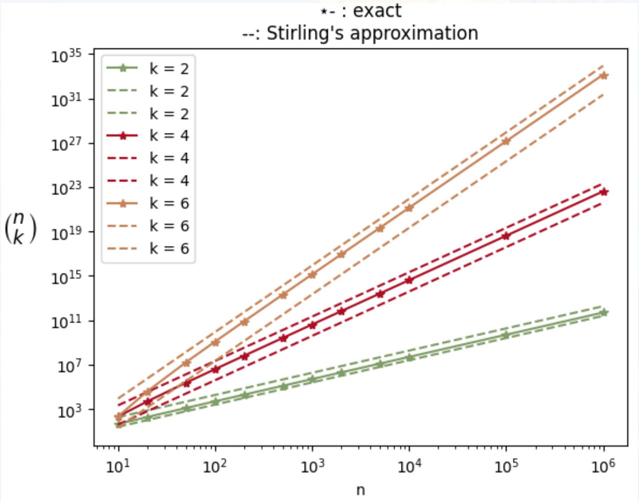
$$\Omega = \frac{N!}{\prod_{i=1}^{I} n_i}$$

for **large *N***: $\lim_{N\to\infty}\left(\frac{n_i}{N}\right) = p_i$

N:	number of indistinguishable particles
n_i:	number of particles in micro state *i*
I:	number of states
p_i:	probability of a particle being in micro state *i*

Stirling's approximation $\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k$

$$N! \approx \left(\frac{N}{e}\right)^N$$



$$\Omega = \frac{N!}{n_1!\,n_2!\ldots n_I!} \approx \frac{N^N}{{n_1}^{n_1}{n_2}^{n_2}\ldots{n_I}^{n_I}}\underbrace{\frac{e^{n_1}e^{n_2}\ldots e^{n_I}}{e^N}}_{\frac{e^{\sum_i^I n_i}}{e^N} = \frac{e^N}{e^N} = 1}$$

$$\Omega = \frac{N^N}{(Np_1)^{n_1}(Np_2)^{n_2}\ldots(Np_I)^{n_I}} = \frac{N^N}{N^{\sum_i^I n_i}}\,\frac{1}{{p_1}^{n_1}{p_2}^{n_2}\ldots{p_I}^{n_I}}$$

"Shannon" Entropy S

$$S = -\sum_{i=1}^{I} p(x_i) \log[p(x_i)]$$

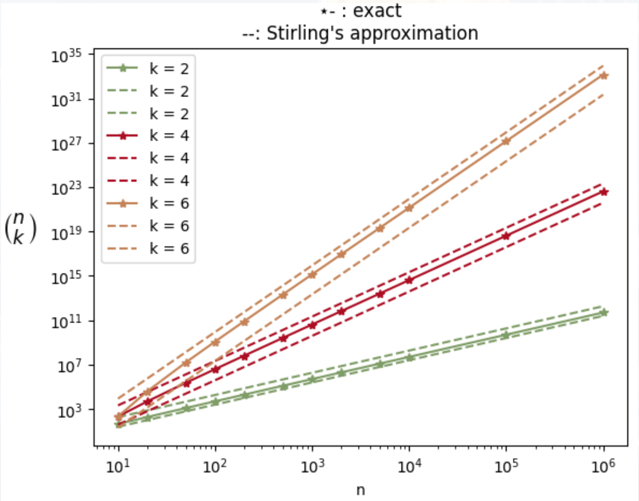
$$\Omega = \frac{N!}{\prod_{i=1}^{I} n_i}$$

for **large *N***: $\lim_{N \to \infty} \left(\frac{n_i}{N}\right) = p_i$

***N*:**	**number of indistinguishable particles**
n_i:	**number of particles in micro state *i***
I:	**number of states**
p_i:	**probability of a particle being in micro state *i***

Stirling's approximation $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{en}{k}\right)^k$

$$N! \approx \left(\frac{N}{e}\right)^N$$



$$\Omega = \frac{N^N}{(Np_1)^{n_1}(Np_2)^{n_2} \dots (Np_I)^{n_I}} = \frac{N^N}{N^{\Sigma_i^I n_i}} \frac{1}{{p_1}^{n_1}{p_2}^{n_2} \dots {p_I}^{n_I}}$$

$$log\Omega = -\sum_i^I n_i \, log(p_i) \qquad \frac{log\Omega}{N} = -\sum_i^I p_i \, log(p_i)$$

Maximum Entropy and Information, Bayes Theorem:

"Shannon" Entropy S

$$S = -\sum_{i=1}^{I} p(x_i)\log[p(x_i)]$$

$$\frac{\log\Omega}{N} = -\sum_{i}^{I} p_i\, log(p_i)$$ entropy per particle

$$S = -N\sum_{i}^{I} p_i\, log(p_i)$$ total entropy

N:	number of indistinguishable particles
n_i:	number of particles in micro state *i*
I:	number of states
p_i:	probability of a particle being in micro state *i*

note:

- using $\log\Omega$ vs Ω is **arbitrary**, but more **convenient** when calculating TD potentials

- We used $\lim_{N\to\infty}\left(\frac{n_i}{N}\right) = p_i$ and Stirling's approximation for large *N*. Therefore, $\lim_{t\to\infty} S(t) = S_{max}$ does **not** hold for small *N* in closed systems! **see HW assignment**

- interpreting *S* as order/disorder is **not** a good concept

Often people explain entropy with an ordered vs messy office…

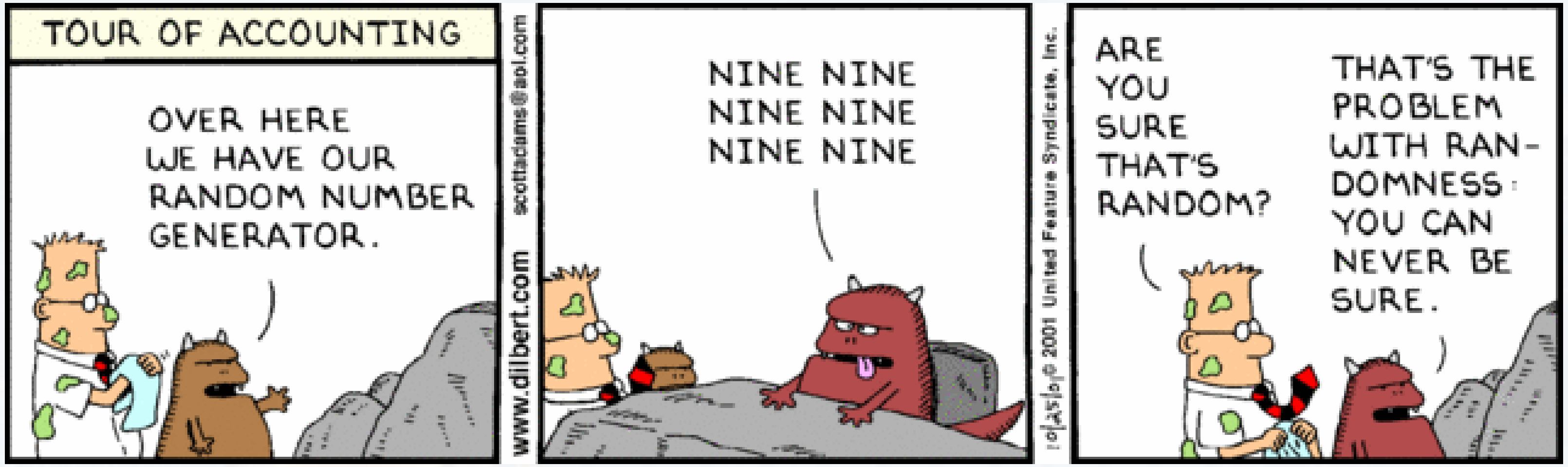
…and then say, that entropy (disorder) grows with time (in closed systems).

- But how is it possible, that an office can do that, just by itself?
- What if my office just *looks* messy, but I can still pull any file you are asking me for?

order/disorder is not a physical quantity!

Those examples have nothing to do with entropy conceptionally!

actually, the idea of entropy is more like that:



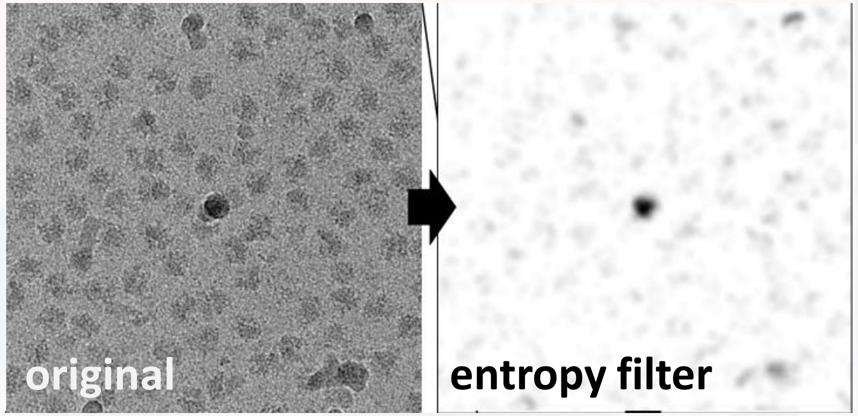
Maximum Entropy and Information, Bayes Theorem:

definition
conditional Entropy
KL divergence
connection to TD

data analysis:

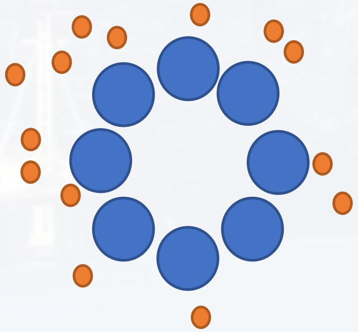
- image processing
- noise reduction
- feature detection

Cryo-EM image of ribosomes



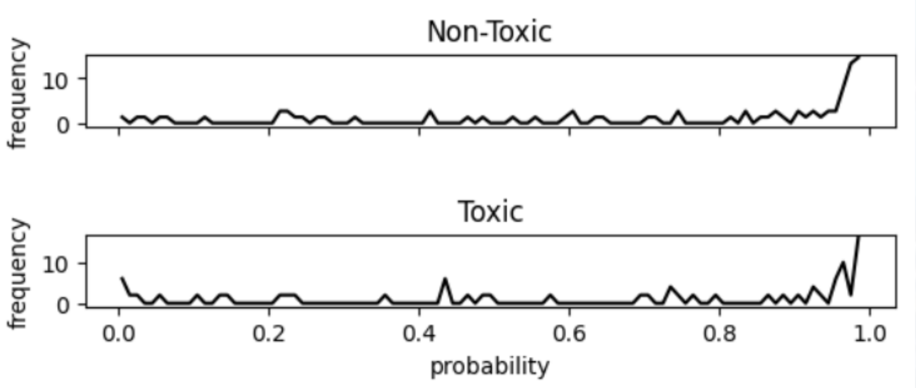
biophysics:

- molecular driving forces
- formation of macromolecules
- “***ordering forces***”



AI:

- optimization
- cross entropy



entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	↑↓↓	↓↓↓
	↑↓↑	↓↑↓	
	↓↑↑	↓↓↑	

eight possible states

no idea (before the experiment)

$$S = -\sum_{i=1}^{8} \frac{1}{8} \, ln\left(\frac{1}{8}\right) = \sum_{i=1}^{8} \frac{1}{8} \, ln(8) = ln(8) \approx \mathbf{2.08}$$

entropy S is a **measure of information** we have **about a system**

definition
conditional Entropy
KL divergence
connection to TD

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	↑↓↓	~~↓↓↓~~
	↑↓↑	↓↑↓	
	↓↑↑	↓↓↑	

one measurement
→ at least one arrow up

seven possible states

$$S = \sum_{i=1}^{7} \frac{1}{7} \, ln(7) + 0 \, \ln(0) \; = ln(7) \approx \mathbf{1.95}$$

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
↑↑↑	↑↑↓	~~↑↓↓~~	~~↓↓↓~~
	↑↓↑	~~↓↑↓~~	
	↓↑↑	~~↓↓↑~~	

second measurement
→ another arrow up

four possible states

$$S = \sum_{i=1}^{4} \frac{1}{4} \, ln(4) \; = ln(4) \approx \mathbf{1.39}$$

Maximum Entropy and Information, Bayes Theorem:

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

possible states of the system

all three up	two up	one up	all three down
~~↑↑↑~~	↑↑↓	~~↑↓↓~~	~~↓↓↓~~
	↑↓↑	~~↓↑↓~~	
	↓↑↑	~~↓↓↑~~	

third measurement
→ one arrow down

three possible states

$$S = ln(3) \approx \mathbf{1.10}$$

The lower the entropy, the more information!

Maximum Entropy and Information, Bayes Theorem:

entropy S is a **measure of information** we have **about a system**

$$S = -\sum_{i=1}^{I} p_i \, ln(p_i)$$

two states ↑ or ↓

and three entities → system

note: - since the base is **arbitrary**, we use **base two** if there are only **two** micro states

- unit is one **bit** (**bi**nary digi**t**)

- $S = -\sum_{i=1}^{8} \frac{1}{8} \, lg\left(\frac{1}{8}\right) = \lg(8) = 3$

Maximum Entropy and Information, Bayes Theorem

Outline

Entropy and Information

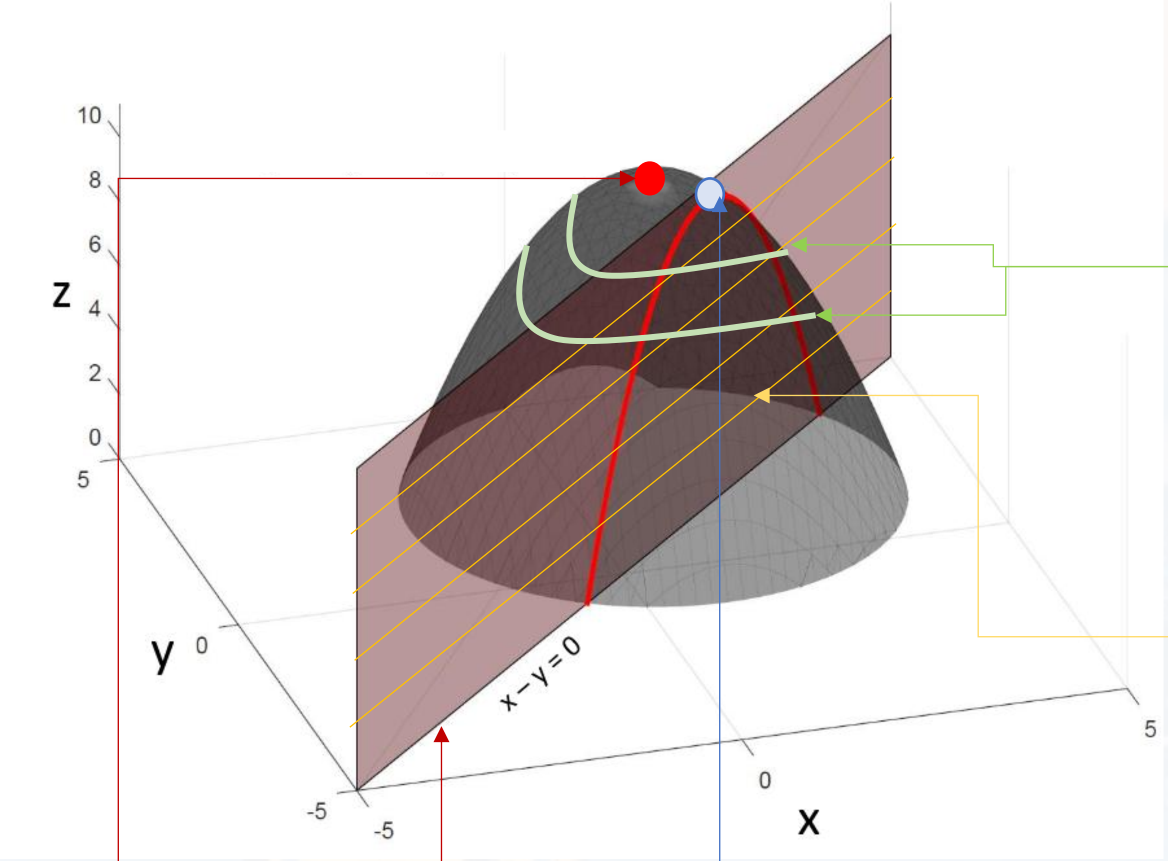
- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

Lagrangian Multiplier
Examples



maximum of the function

$$f(x,y) = z = -(x-2)^2 - (y-1)^2 + 10$$

level lines $f(x,y) = const$

$$df(x,y) = \frac{\partial f(x,y)}{\partial x}dx + \frac{\partial f(x,y)}{\partial y}dy = 0$$

$$= gradf\ d\vec{r} = 0$$

level lines $g(x,y) = const$

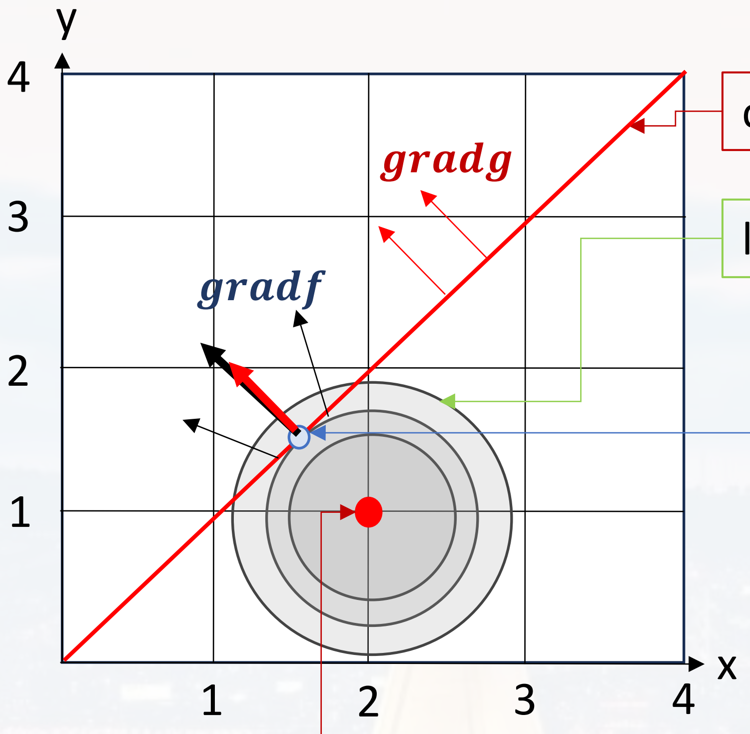
$$dg(x,y) = \frac{\partial g(x,y)}{\partial x}dx + \frac{\partial g(x,y)}{\partial y}dy = 0$$

$$= gradg\ d\vec{r} = 0$$

$z = 10$ at
$x = 2$
$y = 1$

constrain $g(x,y) = x - y = 0$

maximum of the function, subject to $g(x,y)$



constrain $g(x,y) = x - y = 0$

level lines $f(x,y) = const$

$z = 10$ at
$x = 2$
$y = 1$

maximum of the function, subject to $g(x,y)$

the maximum of $f(x,y)$ subject to $g(x,y)$ located where:

$$df(x,y) = dg(x,y)$$

$$gradf\, d\vec{r} = gradg\, d\vec{r}$$

$$gradf = gradg$$

Both gradients need to point in the same direction
(hence, can be multiplied with a constant, say λ)!

$$\boxed{gradf = \lambda\, gradg}$$

λ **Lagrangian Multiplier**

the maximum of $f(x, y)$ subject to $g(x, y)$

$$grad f = \lambda \, grad g$$

$$f(x, y) - \lambda g(x, y) = const$$

the Lagrangian
$\boldsymbol{L(x, y, \lambda)}$

more general:

$$L(x_1, x_2, \dots, x_i, x_N, \lambda_1, \lambda_2, \dots, \lambda_k, \lambda_K) = f(x_1, x_2, \dots, x_i, x_N) - \sum_{k=1}^{K} \lambda_k \, g_k(x_1, x_2, \dots, x_i, x_N)$$

note: - ***N*** **dimensions and** $\boldsymbol{K \leq N}$ **constrains**

- we need to solve *N* (from the gradient) + *K* equations by using the constrains

- optimization: more robust results (most common L1 and L2 regularization, see later)

- machine learning: including constrains in loss function (see later)

Lagrangian Multiplier
Examples

the maximum of $f(x,y)$ subject to $g(x,y)$

maximum of the function

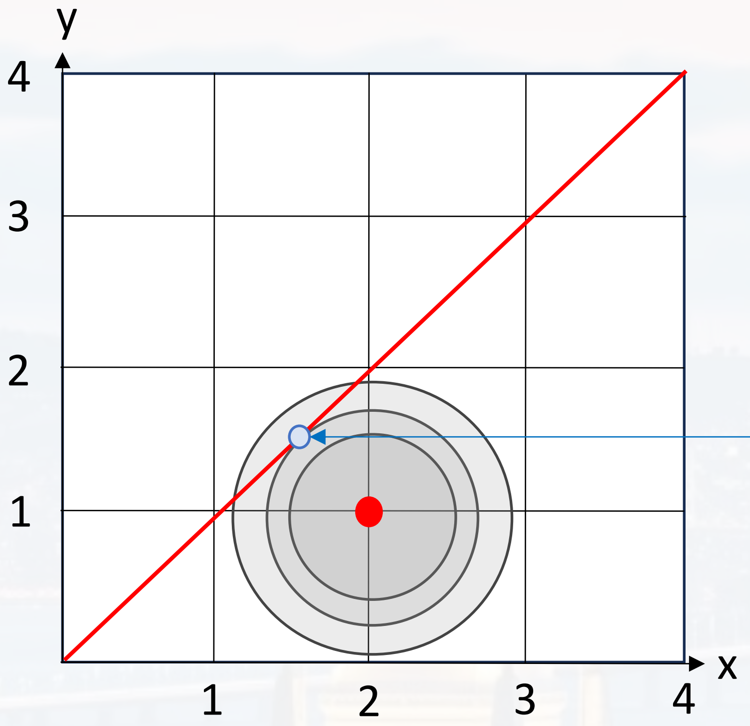
$$f(x,y) = z = -(x-2)^2 - (y-1)^2 + 10$$

constrain $g(x,y) = x - y = 0$

constrain $x = y$

$x = 1.5$
$y = 1.5$

$f(1.5, 1.5) = 9.5$



$z = 9.5$ at
$x = 1.5$
$y = 1.5$

$$grad f = \lambda \, grad g$$

$$\frac{\partial f(x,y)}{\partial x} = \lambda \frac{\partial g(x,y)}{\partial x} \qquad -2(x-2) = \lambda$$

$$\frac{\partial f(x,y)}{\partial y} = \lambda \frac{\partial g(x,y)}{\partial y} \qquad -2(y-1) = -\lambda$$

$$y = -x + 3$$

<u>Outline</u>

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

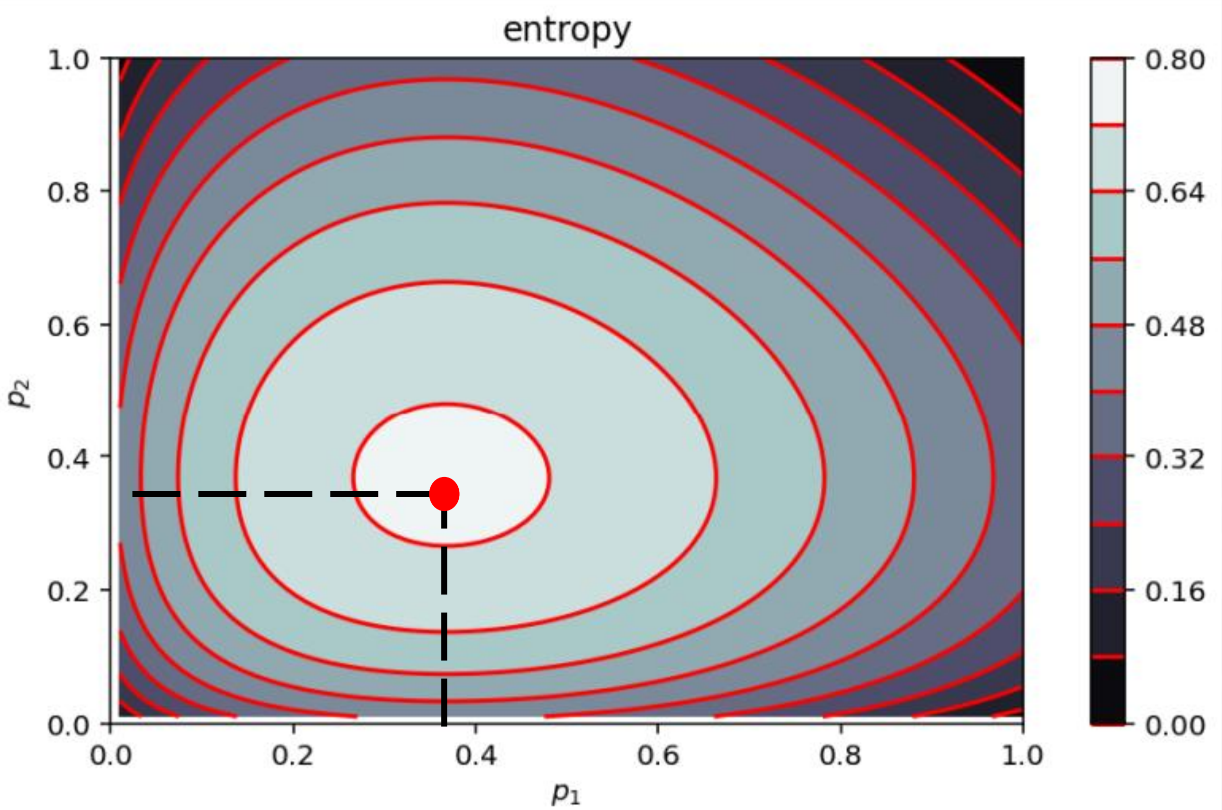
Bayes Theorem

maximum entropy of flipping a coin:

$$f(p_1, p_2) = -p_1 \ln p_1 - p_2 \ln p_2$$

subject to

$$g(p_1, p_2) = p_1 + p_2 = 1$$



absolute maximum:

$$\frac{\partial f(p_1, p_2)}{\partial p_1} = 0 \qquad \frac{\partial f(p_1, p_2)}{\partial p_2} = 0$$

$$-\ln p_1 - 1 = 0 \qquad -\ln p_2 - 1 = 0$$

$$p_1 = p_2 = \frac{1}{e}$$

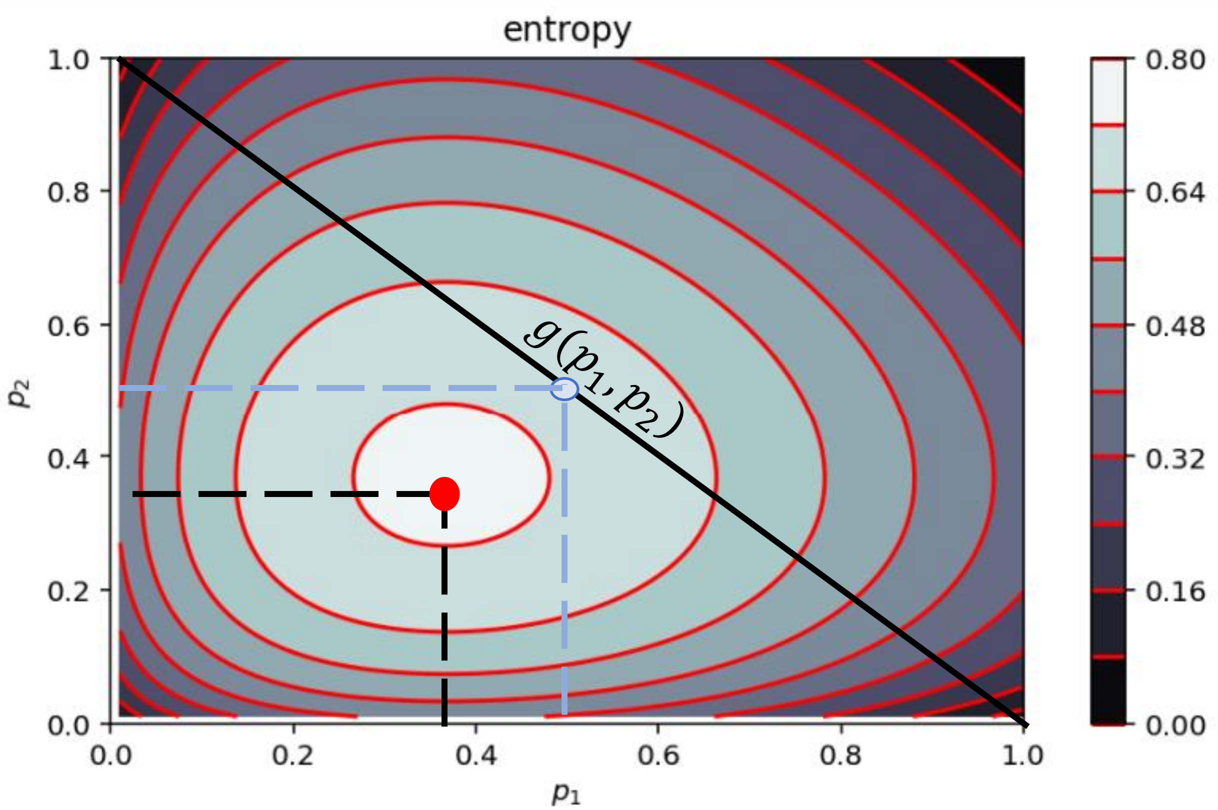
$$f\left(\frac{1}{e}, \frac{1}{e}\right) = \frac{2}{e} \approx 0.74$$

maximum entropy of flipping a coin:

$$f(p_1, p_2) = -p_1 \ln p_1 - p_2 \ln p_2$$

subject to

$$g(p_1, p_2) = p_1 + p_2 = 1$$



maximum subject to $g(p_1, p_2)$:

$$\frac{\partial f(p_1, p_2)}{\partial p_1} = \lambda \frac{\partial g(p_1, p_2)}{\partial p_1}$$

$$\frac{\partial f(p_1, p_2)}{\partial p_2} = \lambda \frac{\partial g(p_1, p_2)}{\partial p_2}$$

$$-\ln p_1 - 1 = \lambda \qquad -\ln p_2 - 1 = \lambda$$

$$p_1 = p_2$$

constrain: $\quad p_1 + p_2 = 1$

$$p_1 = p_2 = \frac{1}{2}$$

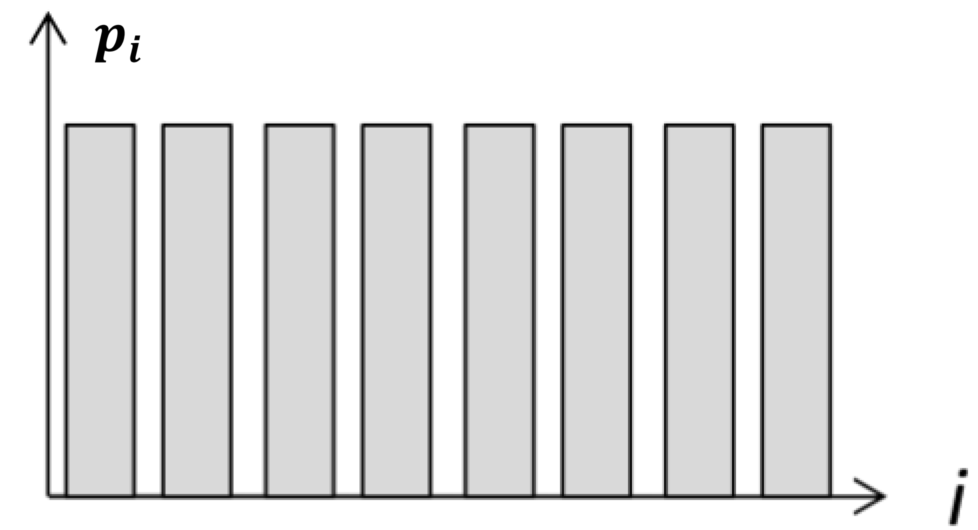
$$f\left(\frac{1}{2}, \frac{1}{2}\right) = ln2 \approx 0.69$$

maximum entropy for I states:

$$f(p_1,..,p_i,...p_I) = -\sum_{i=1}^{I} p_i \, ln p_i$$

subject to

$$g(p_1,..,p_i,...p_I) = \sum_{i=1}^{I} p_i \; = 1$$



maximum subject to $g(p_1,..,p_i,...p_I)$:

$$\frac{\partial f(p_1,..,p_i,...p_I)}{\partial p_i} = \lambda \frac{\partial g(p_1,..,p_i,...p_I)}{\partial p_i}$$

$$-\ln p_i - 1 = \lambda \qquad p_i = e^{-(1+\lambda)}$$

constrain: $$\sum_{i=1}^{I} e^{-(1+\lambda)} = 1$$

$$I e^{-(1+\lambda)} = 1$$

$$e^{-(1+\lambda)} = \frac{1}{I}$$

probabilities are constant!
→ flat distribution!

$$\boxed{p_i = \frac{1}{I}}$$

Lagrangian Multiplier
Examples

N:	**number of indistinguishable particles**
n_i:	**number of particles in micro state *i***
I:	**number of states**
p_i:	**probability of a particle being in micro state *I***
ε_i:	**energy in state *i***

maximum entropy for I states:

$$f(p_1,..,p_i,\dots p_I) = -\sum_{i=1}^{I} p_i \, ln p_i$$

subject to

$$g_1(p_1,..,p_i,\dots p_I) = \sum_{i=1}^{I} p_i \; = 1$$

if *N* and ***total energy*** is conserved

$$g_2(p_1,..,p_i,\dots p_I) = \sum_{i=1}^{I} p_i \varepsilon_\mathrm{i} = \frac{E_{tot}}{N} = \frac{1}{N}\sum_{i=1}^{I} n_i \varepsilon_\mathrm{i}$$

$$\frac{\partial f(p_1,..,p_i,\dots p_I)}{\partial p_i} = \lambda_1 \frac{\partial g_1(p_1,..,p_i,\dots p_I)}{\partial p_i} + \lambda_2 \frac{\partial g_2(p_1,..,p_i,\dots p_I)}{\partial p_i}$$

$$-\ln p_i - 1 = \lambda_1 + \lambda_2 \frac{\partial \sum_{j=1}^{I} p_j \varepsilon_\mathrm{j}}{\partial p_i}$$

maximum entropy for I states:

$$f(p_1, .., p_i, \dots p_I) = -\sum_{i=1}^{I} p_i \, ln p_i$$

subject to

$$g_1(p_1, .., p_i, \dots p_I) = \sum_{i=1}^{I} p_i \; = 1$$

if *N* and ***total energy*** is conserved

$$g_2(p_1, .., p_i, \dots p_I) = \sum_{i=1}^{I} p_i \varepsilon_\mathrm{i}$$

$$-\ln p_i - 1 = \lambda_1 + \lambda_2 \frac{\partial \sum_{j=1}^{I} p_j \varepsilon_\mathrm{j}}{\partial p_i}$$

$$-\ln p_i - 1 = \lambda_1 + \lambda_2 \varepsilon_\mathrm{i}$$

$$p_i = e^{-(1+\lambda_1)} \, e^{-\lambda_2 \varepsilon_\mathrm{i}}$$

***N*:**	**number of indistinguishable particles**
$\boldsymbol{n_i}$**:**	**number of particles in micro state *i***
I:	**number of states**
$\boldsymbol{p_i}$**:**	**probability of a particle being in micro state *I***
$\boldsymbol{\varepsilon_i}$:	**energy in state *i***

from g_1:

$$p_\mathrm{i} = \frac{1}{\sum_{i=1}^{I} e^{-\lambda_2 \varepsilon_\mathrm{i}}} \, e^{-\lambda_2 \varepsilon_\mathrm{i}}$$

partition function Z

$$Z = \sum_{i=1}^{I} e^{-\lambda_2 \varepsilon_\mathrm{i}}$$

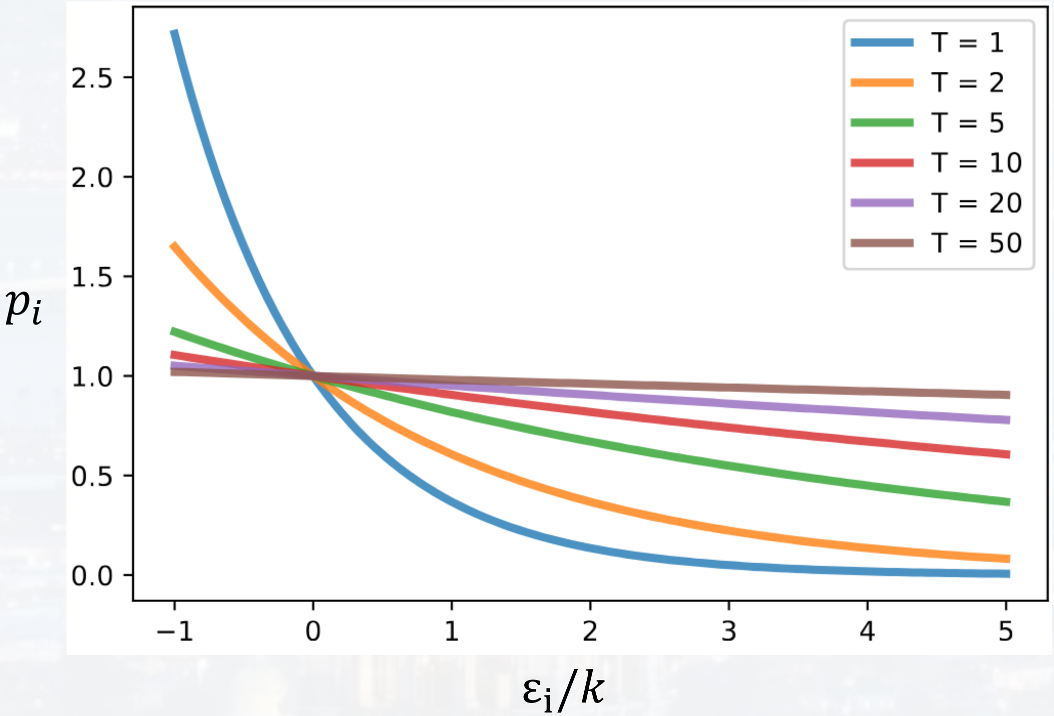
Boltzmann distribution $\boxed{p_i = \frac{1}{Z} \, e^{-\lambda_2 \varepsilon_\mathrm{i}}}$

partition function Z $$Z = \sum_{i=1}^{I} e^{-\lambda_2 \varepsilon_i}$$

Boltzmann distribution $$p_i = \frac{1}{Z} e^{-\lambda_2 \varepsilon_i}$$

one can show: $$\lambda_2 = \frac{1}{kT}$$

N:	number of indistinguishable particles
n_i:	number of particles in micro state *i*
I:	number of states
p_i:	probability of a particle being in micro state *I*
ε_i:	energy in state *i*



note:

- for $T \to \infty$, $Z \to I$, i. e. higher states become more accessible and $p_i \to \frac{1}{I}$
- we used maximum entropy: **equilibrium** state for large N
- N **and** E_{tot} are constant
- ANN: **softmax layer** for classification probabilities (see later)

$$f(x) = \int_{-\infty}^{+\infty} p(x) \ln[p(x)]\, dx$$

continuous distribution with support $(-\infty, +\infty)$

$$g_1(x) = \int_{-\infty}^{+\infty} p(x) dx = 1$$

$$g_2(x) = \int_{-\infty}^{+\infty} x\, p(x) dx = \mu$$

mean μ

$$g_3(x) = \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x) dx = \sigma^2$$

standard deviation σ^2

$$\int_{-\infty}^{+\infty} p(x) \ln[p(x)]\, dx = \lambda_1 \int_{-\infty}^{+\infty} p(x) dx + \lambda_2 \int_{-\infty}^{+\infty} x\, p(x) dx + \lambda_3 \int_{-\infty}^{+\infty} (x-\mu)^2\, p(x) dx$$

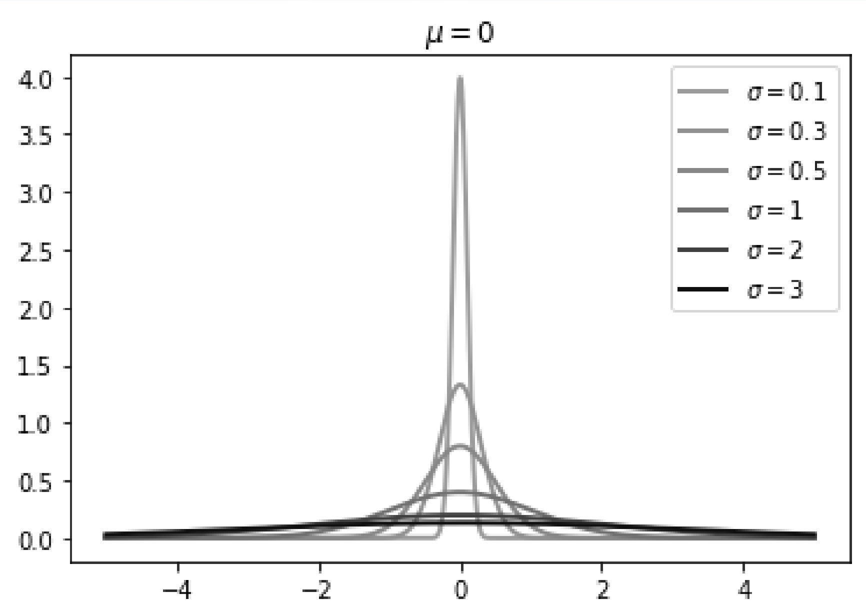
$$p(x) = \exp[-\lambda_1 - \lambda_2 x - \lambda_3 (x-\mu)^2 - 1]$$

Normal (Gauss) distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Normal (Gauss) distribution

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



note:

- $S[p(x)] = \ln(2\pi\sigma) + 1/2$ **(which can be negative)**
- constrains where very generic → normal distribution is **often a good approximation**
- often: $N(\mu, \sigma^2)$

maximum entropy distributions

Distribution name	Probability density / mass function	Maximum entropy constraint	Support
Uniform (discrete)	$f(k)=\frac{1}{b-a+1}$	None	$\{a,a+1,\ldots,b-1,b\}$
Uniform (continuous)	$f(x)=\frac{1}{b-a}$	None	$[a,b]$
Bernoulli	$f(k)=p^k(1-p)^{1-k}$	$\mathrm{E}[K]=p$	$\{0,1\}$
Normal	$f(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\mathrm{E}[X]=\mu\,,$ $\mathrm{E}[X^2]=\sigma^2+\mu^2$	$\mathbb{R}$
Gamma	$f(x)=\frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}$	$\mathrm{E}[X]=k\theta\,,$ $\mathrm{E}[\ln X]=\psi(k)+\ln\theta$	$[0,\infty)$
Binomial	$f(k)=\binom{n}{k}p^k(1-p)^{n-k}$	$\mathrm{E}[X]=\mu\,,$ $f\in$ n-generalized binomial distribution[11]	$\{0,\ldots,n\}$
Poisson	$f(k)=\frac{\lambda^k e^{-\lambda}}{k!}$	$\mathrm{E}[X]=\lambda,$ $f\in\infty$-generalized binomial distribution[11]	$\mathbb{N}=\{0,1,\ldots\}$
Logistic	$f(x)=\frac{e^{-x}}{(1+e^{-x})^2}=\frac{e^{+x}}{(e^{+x}+1)^2}$	$\mathrm{E}[X]=0,$ $\mathrm{E}[\ln(1+e^{-X})]=1$	$\{-\infty,\infty\}$

maximum entropy distributions

Distribution name	Probability density / mass function	Maximum entropy constraint	Support
Uniform (discrete)	$f(k)=\frac{1}{b-a+1}$	None	$\{a,a+1,\ldots,b-1,b\}$
Uniform (continuous)	$f(x)=\frac{1}{b-a}$	None	$[a,b]$
Bernoulli	$f(k)=p^k(1-p)^{1-k}$	$\mathrm{E}[K]=p$	$\{0,1\}$
Multivariate normal	$f_X(\mathbf{x})=\frac{\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}{\sqrt{(2\pi)^N\lvert\Sigma\rvert}}$	$\mathrm{E}[\mathbf{x}]=\boldsymbol{\mu},$ $\mathrm{E}\left[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\right]=\Sigma$	$\mathbb{R}^n$
Gamma	$f(x)=\frac{x^{k-1}e^{-x/\theta}}{\theta^k\Gamma(k)}$	$\mathrm{E}[X]=k\theta,$ $\mathrm{E}[\ln X]=\psi(k)+\ln\theta$	$[0,\infty)$
Binomial	$f(k)=\binom{n}{k}p^k(1-p)^{n-k}$	$\mathrm{E}[X]=\mu,$ $f\in$ n-generalized binomial distribution[11]	$\{0,\ldots,n\}$
Poisson	$f(k)=\frac{\lambda^k e^{-\lambda}}{k!}$	$\mathrm{E}[X]=\lambda,$ $f\in\infty$-generalized binomial distribution[11]	$\mathbb{N}=\{0,1,\ldots\}$
Logistic	$f(x)=\frac{e^{-x}}{(1+e^{-x})^2}=\frac{e^{+x}}{(e^{+x}+1)^2}$	$\mathrm{E}[X]=0,$ $\mathrm{E}\left[\ln\left(1+e^{-X}\right)\right]=1$	$\{-\infty,\infty\}$

Outline

Entropy and Information

- definition
- conditional Entropy
- KL divergence
- connection to TD

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem

Maximum Entropy and Information, Bayes Theorem:

Thomas Bayes (1701 - 1761)

$P(A \cap B)$ probability ***P*** that the events ***A* and *B*** occur

so far: A and B were independent $P(A \cap B) = P(A)P(B) = P(B)P(A)$

now: **conditional probabilities** | "given" or "under the condition"

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\longrightarrow \boxed{P(A|B)P(B) = P(B|A)P(A)}$$

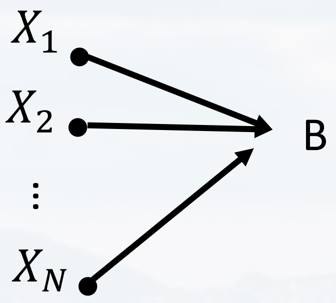
Bayes Theorem

$$\text{posterior} \quad P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad \text{prior}$$

Maximum Entropy and Information, Bayes Theorem:

Bayes Theorem

posterior $\boldsymbol{P(A|B)} = \dfrac{P(B|A)\boldsymbol{P(A)}}{\boldsymbol{P(B)}}$ prior



$$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

$$P(B) = \int P(B|X)P(X)\, dX$$

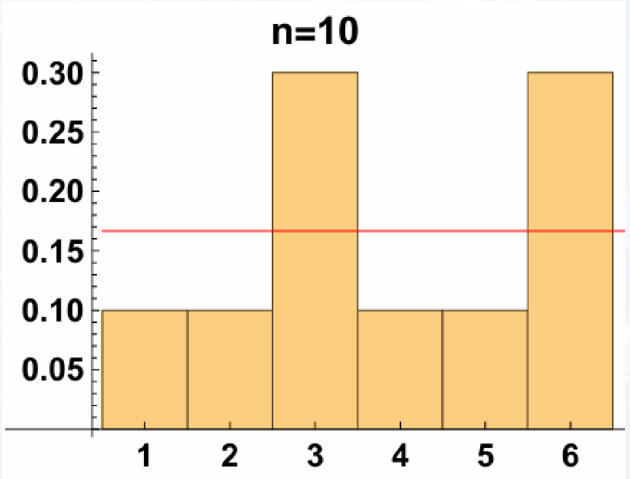
marginalization

Thomas Bayes (1701 - 1761)

for a normal distribution M $= \mathcal{N}(\mu,\sigma)$

$$P(D|\mathcal{N}) = \int P(D|\mathcal{N}(\mu,\sigma))\, P(\mu,\sigma|\mathcal{N}(\mu,\sigma))\, d\Omega_{\mu,\sigma}$$

model: M
data: D



σ = 2, μ = 3.5
σ = 2, μ = 5.0
σ = 1.5, μ = 3.5
σ = 7.0, μ = 1.0
….and so on

example: cancer diagnosis from blood test

+ : positive test result
D : diseased
H : health

Bayes Theorem $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Marginalization $$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

Maximum Entropy and Information, Bayes Theorem:

example: cancer diagnosis from blood test

Bayes Theorem $$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Marginalization $$P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$$

+ : positive test result
D : diseased
H : health

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**

statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .

statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

A person takes the test and gets a positive test result. **What is the probability that the person is sick**?

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} \underset{\text{statement 1}}{=} \frac{0.95\, P(D)}{P(+)} \underset{\text{statement 2}}{=} \frac{0.95 \cdot 0.00001}{P(+)} \underset{\text{marginalization}}{=} \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)P(H)}$$

Maximum Entropy and Information, Bayes Theorem:

example: cancer diagnosis from blood test

Bayes Theorem $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Marginalization $P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$

+ : **positive test result**
D : **diseased**
H : **health**

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**
statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .
statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{0.95\,P(D)}{P(+)} = \frac{0.95 \cdot 0.00001}{P(+)} = \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)P(H)}$$

statement 1 statement 2 marginalization

$$= \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)[1 - P(D)]}$$

complement probability

example: cancer diagnosis from blood test

Bayes Theorem $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$

Marginalization $P(B) = \sum_{n=1}^{N} P(B|X_n)P(X_n)$

+ : **positive test result**
D : **diseased**
H : **health**

statement 1: If a person is **diseased**, there is a **95% probability** that the test is **positive.**
statement 2: The **prevalence** for the disease in the average **population** is **0.001%** .
statement 3: **5% of healthy** patients have **a positive result** (aka p-value).

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|H)[1 - P(D)]}$$

$$\boxed{= \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}} = \frac{1}{1 + \frac{0.05\,[1 - 0.00001]}{0.95 \cdot 0.00001}} = 1/5000$$

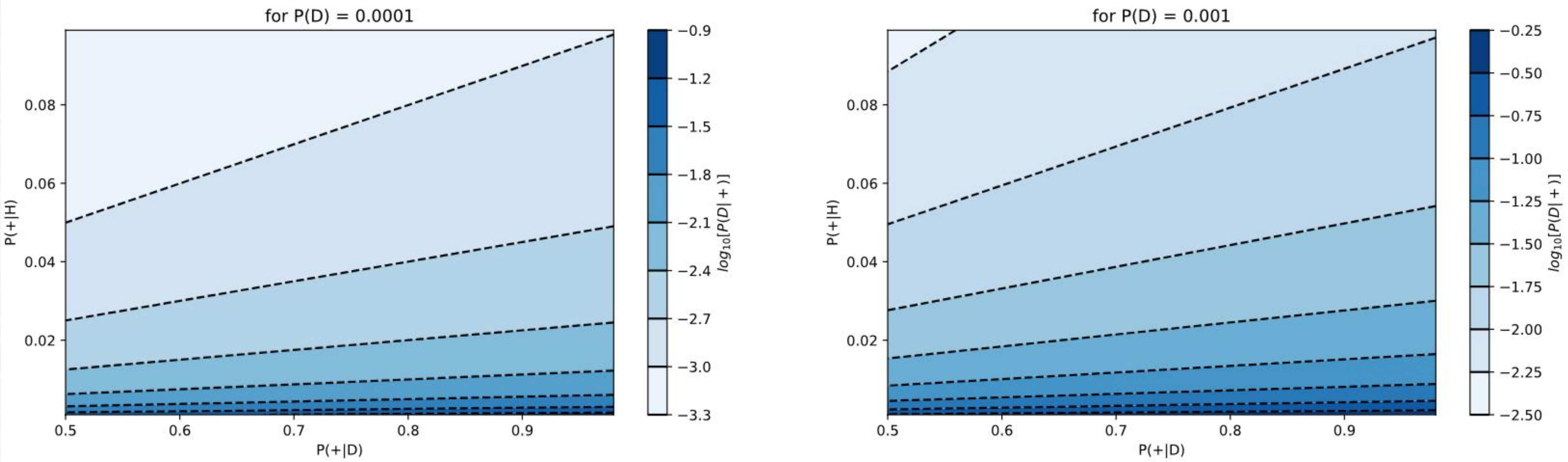
example: cancer diagnosis from blood test

+ : **positive test result**
D : **diseased**
H : **health**

statement 1:	sensitivity	P(D|+) = 95%
statement 2:	prior	P(D) = 0.001%
statement 3:	p-value or false positive rate	P(+|H) = 5%

$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}$$

check: `PlotPD_Plus.py`



Maximum Entropy and Information, Bayes Theorem:

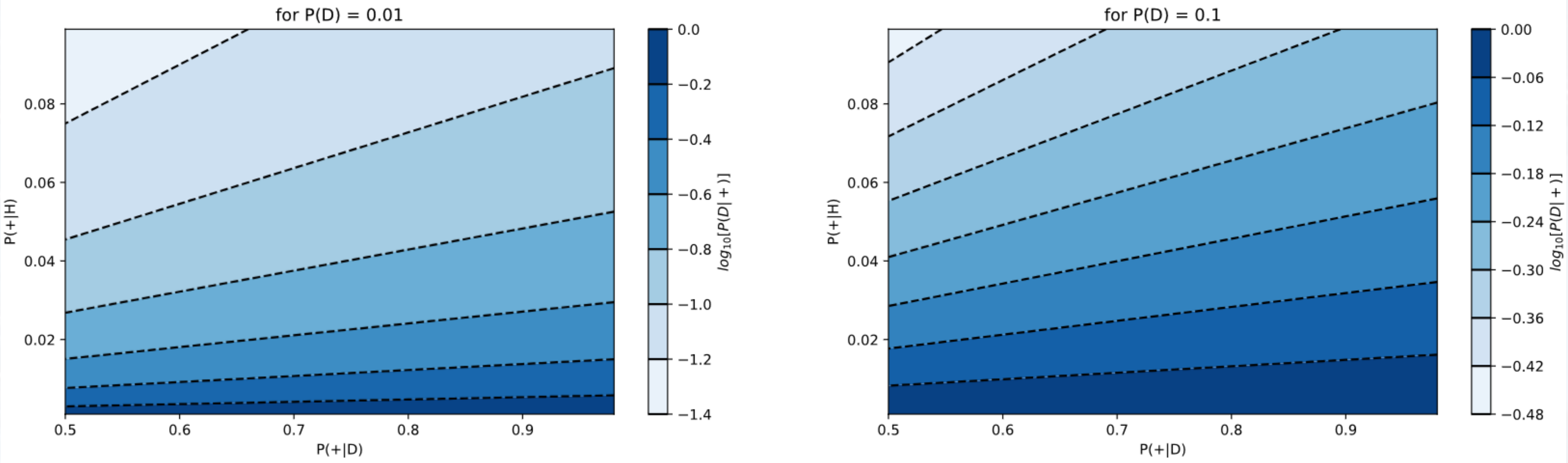
example: cancer diagnosis from blood test

+ : positive test result
D : diseased
H : health

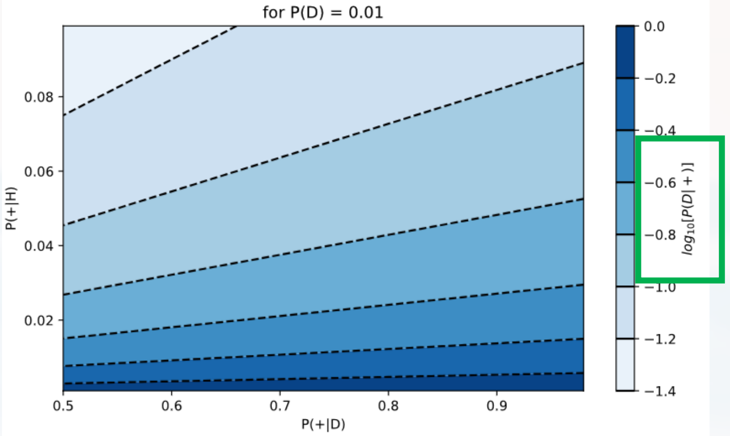
statement 1:	sensitivity	P(D|+) = 95%
statement 2:	prior	P(D) = 0.001%
statement 3:	p-value or false positive rate	P(+|H) = 5%

$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}$$

check: `PlotPD_Plus.py`



Maximum Entropy and Information, Bayes Theorem:



statement 1:	sensitivity	P(D|+) = 95%
statement 2:	prior	P(D) = 0.001%
statement 3:	p-value or false positive rate	P(+|H) = 5%

odds ratios:

$$\rho_1 = \frac{P(+|H)}{P(+|D)}$$

$$\rho_2 = \frac{1 - P(D)}{P(D)}$$

$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1 - P(D)]}{P(+|D)P(D)}}$$

log odds ratios:

$$r_1 = \log\left[\frac{P(+|H)}{P(+|D)}\right]$$

$$r_2 = \log\left[\frac{1 - P(D)}{P(D)}\right]$$

$$P(D|+) = \frac{1}{1 + e^{r_1} e^{r_2}}$$

log odds ratios: $r_1 = \log\left[\frac{P(+|H)}{P(+|D)}\right]$

$$r_2 = \log\left[\frac{1-P(D)}{P(D)}\right]$$

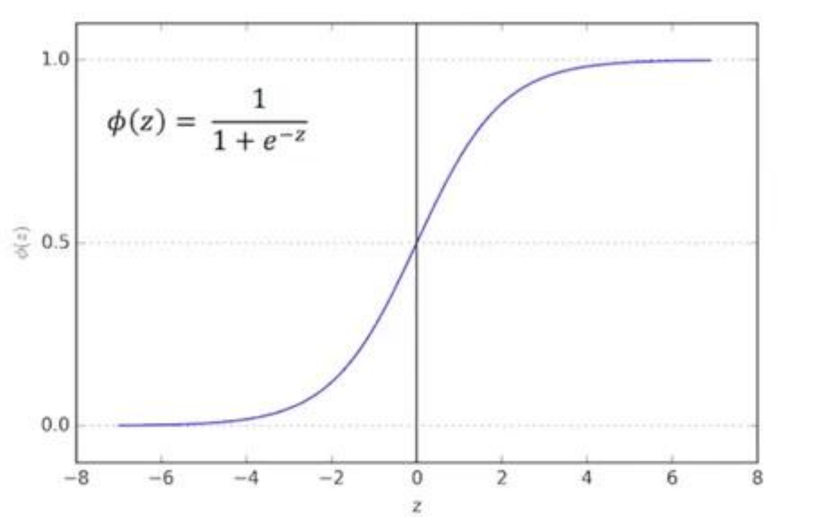
$$P(D|+) = \frac{1}{1 + \frac{P(+|H)[1-P(D)]}{P(+|D)P(D)}}$$

$$\boxed{P(D|+) = \frac{1}{1+e^{r_1}e^{r_2}}}$$

note: - **logistic (or logit or sigmoid) function**

- transfer function ANN (see later)

- bound growth (Verhulst equation)



Thank you very much for your attention!