

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)



Outline

Entropy and Information

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

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem



Outline

Entropy and Information

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

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem



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

definition
conditional Entropy
KL divergence
connection to TD

$p(x_i)$: 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

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

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



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

definition

conditional Entropy

KL divergence

connection to TD

$p(x_i)$: probability of event x_i

$h(x_i)$: function that measures “information”

1) $h(x_i)$ should be additive

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

2) $h(x_i)$ should be monotonic

3) if x_i and x_j are independent, then

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

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

$$p(x_i) \leq 1$$

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

information is **positive**

low $p(x_i) \rightarrow$ “great surprise”



idea: gain of information = “**degree of surprise**”

if something happens that we expected already → no surprise
→ no information

definition

conditional Entropy

KL divergence

connection to TD

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

$p(x_i)$: probability of event x_i

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

Entropy S

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

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

note: - the **base** of log is **arbitrary** (often 2 or e)

$$- \lim_{p \rightarrow 0} (p \log p) = 0$$

- 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



Entropy S

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

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

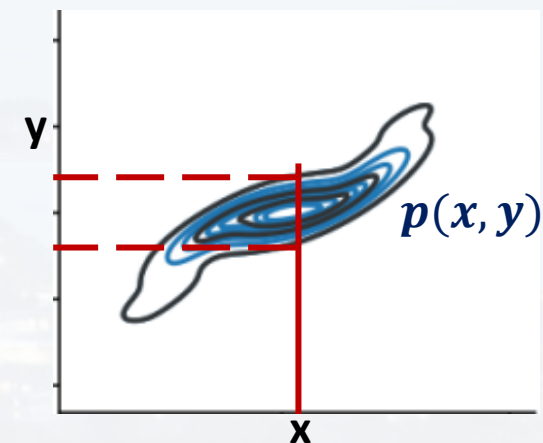
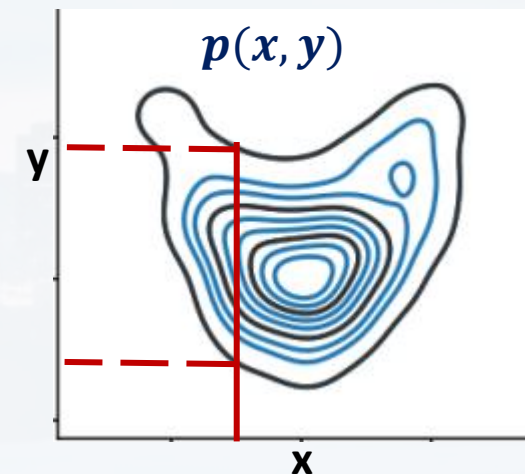
definition
conditional Entropy
KL divergence
connection to TD

$p(x_i)$: probability of event x_i

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

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

“given”





Entropy S

$$S = - \int p(x) \log[p(x)] dx$$

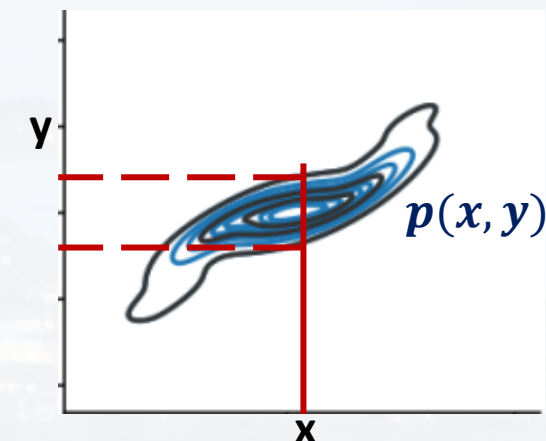
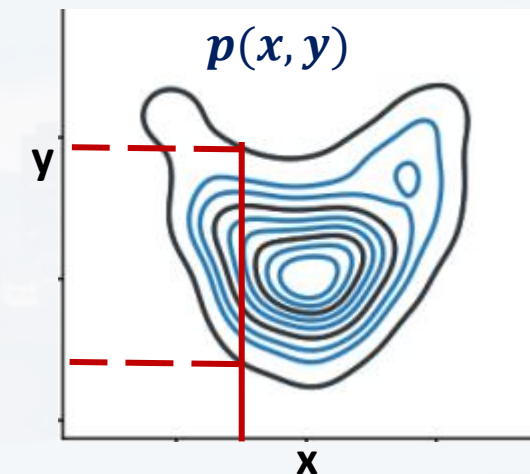
definition
conditional Entropy
KL divergence
connection to TD

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

$S(y|x)$: entropy of y , “given” x (we know x already)

$$\begin{aligned} S[p(x, y)] &= - \iint p(x, y) \log[p(x, y)] dx dy \\ &= - \iint p(x, y) \log[p(y|x)p(x)] dx dy && \text{conditional probabilities (see later)} \\ &= - \iint p(x, y) \{ \log[p(y|x)] + \log[p(x)] \} dx dy \\ &= - \iint p(x, y) \log[p(y|x)] dx dy - \iint p(x, y) \log[p(x)] dx dy \\ &&& \text{we still sample from } p(x, y) \end{aligned}$$

$p(x_i)$: probability of event x_i





Entropy S

$$S = - \int p(x) \log[p(x)] dx$$

definition
conditional Entropy
KL divergence
connection to TD

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

$S(y|x)$: entropy of y , “given” x (we know x already)

$$S[p(x, y)] = - \iint p(x, y) \log[p(y|x)] dx dy - \iint p(x, y) \log[p(x)] dx dy$$

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

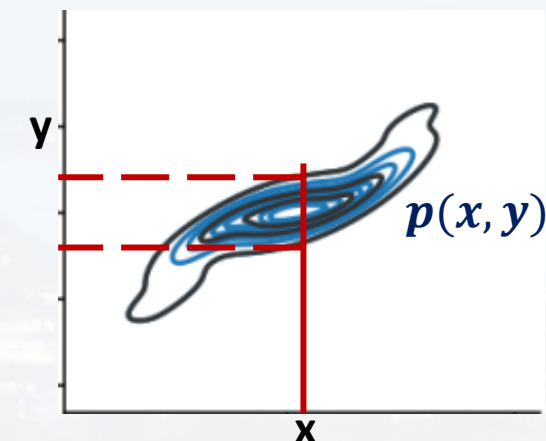
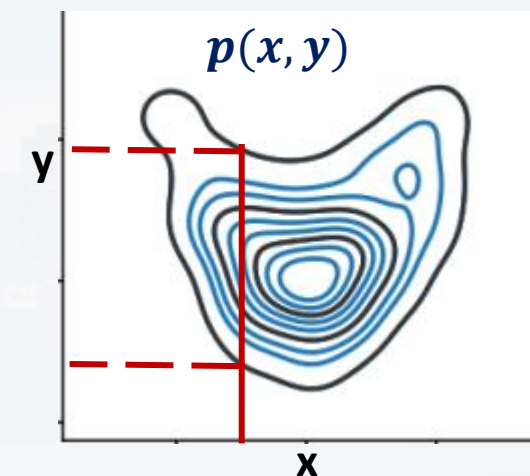
differential
entropy of
 $p(x, y)$

conditional
entropy

differential
entropy of
 $p(x)$

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

$p(x_i)$: probability of event x_i





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 $p(x)$ by some other distribution $q(x)$



definition
conditional Entropy
KL divergence
connection to TD

How much is our information “off” if we work with the approximation $q(x)$?

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

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

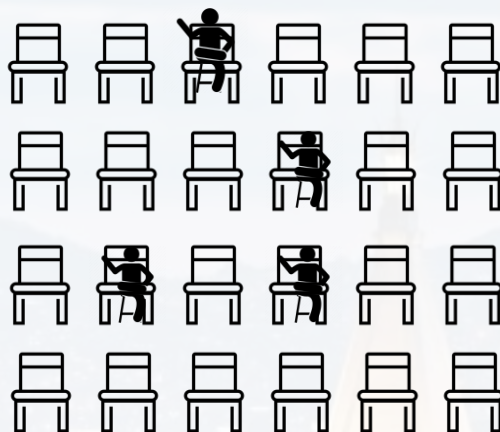
Bayes Theorem



“Shannon”
Entropy S

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

definition
conditional Entropy
KL divergence
connection to TD



$I = 2$ 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

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

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

p_i : probability of a particle being in micro state i



“Shannon”
Entropy S

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

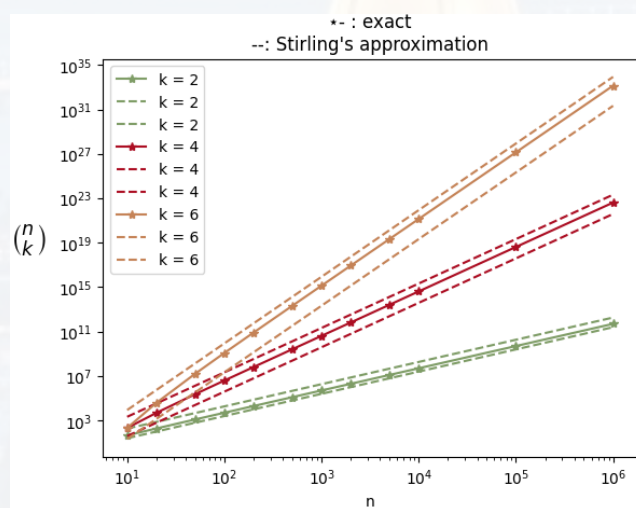
definition
conditional Entropy
KL divergence
connection to TD

$$\Omega = \frac{N!}{\prod_{i=1}^I n_i!} \quad \text{for large } N: \lim_{N \rightarrow \infty} \left(\frac{n_i}{N} \right) = p_i$$

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

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

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



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

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



“Shannon”
Entropy S

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

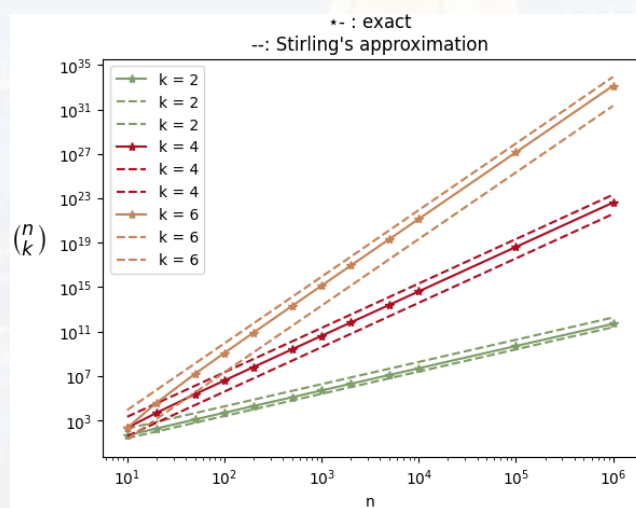
definition
conditional Entropy
KL divergence
connection to TD

$$\Omega = \frac{N!}{\prod_{i=1}^I n_i!} \quad \text{for large } N: \lim_{N \rightarrow \infty} \left(\frac{n_i}{N} \right) = p_i$$

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

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

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



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

$$\log \Omega = - \sum_i n_i \log(p_i) \quad \frac{\log \Omega}{N} = - \sum_i p_i \log(p_i)$$



“Shannon”
Entropy S

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

definition
conditional Entropy
KL divergence
connection to TD

$$\frac{\log \Omega}{N} = - \sum_i p_i \log(p_i) \quad \text{entropy per particle}$$

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

$$S = -N \sum_i p_i \log(p_i) \quad \text{total entropy}$$

note:

- using $\log \Omega$ vs Ω is **arbitrary**, but more **convenient** when calculating TD potentials
- We used $\lim_{N \rightarrow \infty} \left(\frac{n_i}{N} \right) = p_i$ and Stirling's approximation for large N . Therefore, $\lim_{t \rightarrow \infty} S(t) = S_{max}$
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).



definition
conditional Entropy
KL divergence
connection to TD

- 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:

definition
conditional Entropy
KL divergence
connection to TD

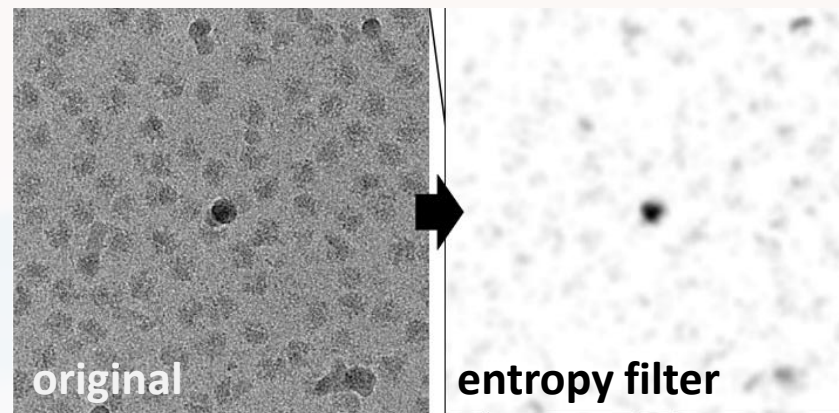




data analysis:

- image processing
- noise reduction
- feature detection

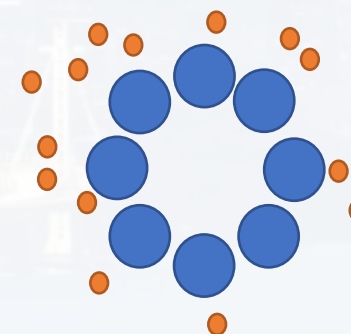
Cryo-EM image of ribosomes



definition
conditional Entropy
KL divergence
connection to TD

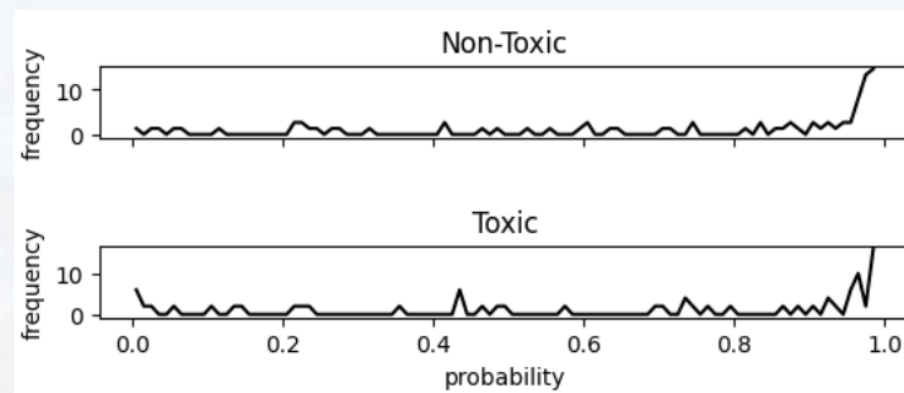
biophysics:

- 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 \uparrow or \downarrow

and three entities \rightarrow system

definition
conditional Entropy
KL divergence
connection to TD

possible states of the system

all three up

two up

one up

all three down

$\uparrow\uparrow\uparrow$

$\uparrow\uparrow\downarrow$

$\uparrow\downarrow\downarrow$

$\downarrow\downarrow\downarrow$

$\uparrow\downarrow\uparrow$

$\downarrow\uparrow\downarrow$

$\downarrow\uparrow\uparrow$

$\downarrow\downarrow\uparrow$

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**

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

two states \uparrow or \downarrow

and three entities \rightarrow system

definition
conditional Entropy
KL divergence
connection to TD

possible states of the system

all three up

two up

one up

all three down

$\uparrow\uparrow\uparrow$

$\uparrow\uparrow\downarrow$

$\uparrow\downarrow\downarrow$

~~$\downarrow\downarrow\downarrow$~~

$\uparrow\downarrow\uparrow$

$\downarrow\uparrow\downarrow$

$\downarrow\uparrow\uparrow$

$\downarrow\downarrow\uparrow$

one measurement
 \rightarrow 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 \uparrow or \downarrow

and three entities \rightarrow system

definition
conditional Entropy
KL divergence
connection to TD

possible states of the system

all three up

two up

one up

all three down

$\uparrow\uparrow\uparrow$	$\uparrow\uparrow\downarrow$	$\uparrow\downarrow\downarrow$	$\downarrow\downarrow\downarrow$
	$\uparrow\downarrow\uparrow$	$\downarrow\uparrow\downarrow$	
	$\downarrow\uparrow\uparrow$	$\downarrow\downarrow\uparrow$	

second measurement
 \rightarrow another arrow up

four possible states

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



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

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

two states \uparrow or \downarrow

and three entities \rightarrow system

definition
conditional Entropy
KL divergence
connection to TD

possible states of the system

all three up

two up

one up

all three down

↑↑↑	↑↑↓	↓↓↓	↓↓↓
	↑↓↑	↓↑↑	
	↓↑↑	↓↑↓	

third measurement
 \rightarrow one arrow down

three possible states

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

The lower the entropy, the more information!



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

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

two states \uparrow or \downarrow

and three entities \rightarrow system

definition
conditional Entropy
KL divergence
connection to TD

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

- unit is one **bit** (**b**inary **d**igit)

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



Outline

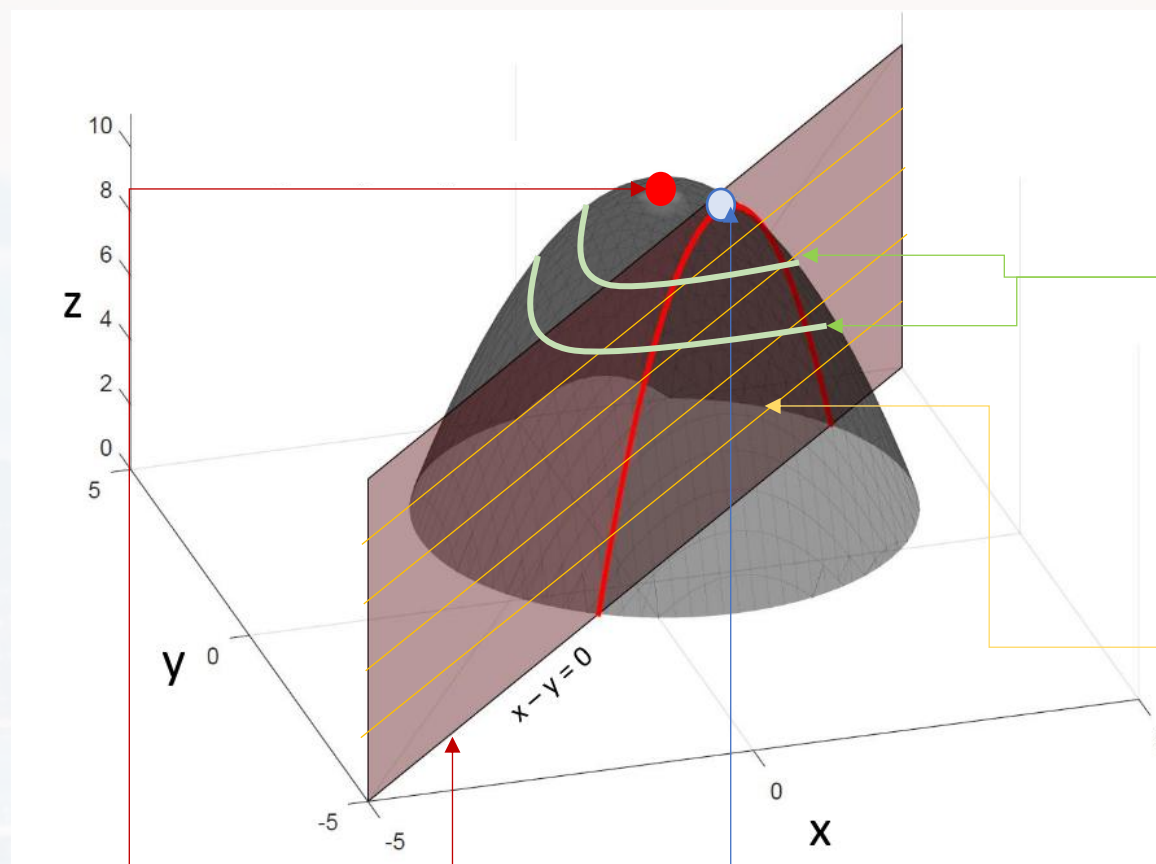
Entropy and Information

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

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem



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

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

maximum of the function

Lagrangian Multiplier
Examples

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

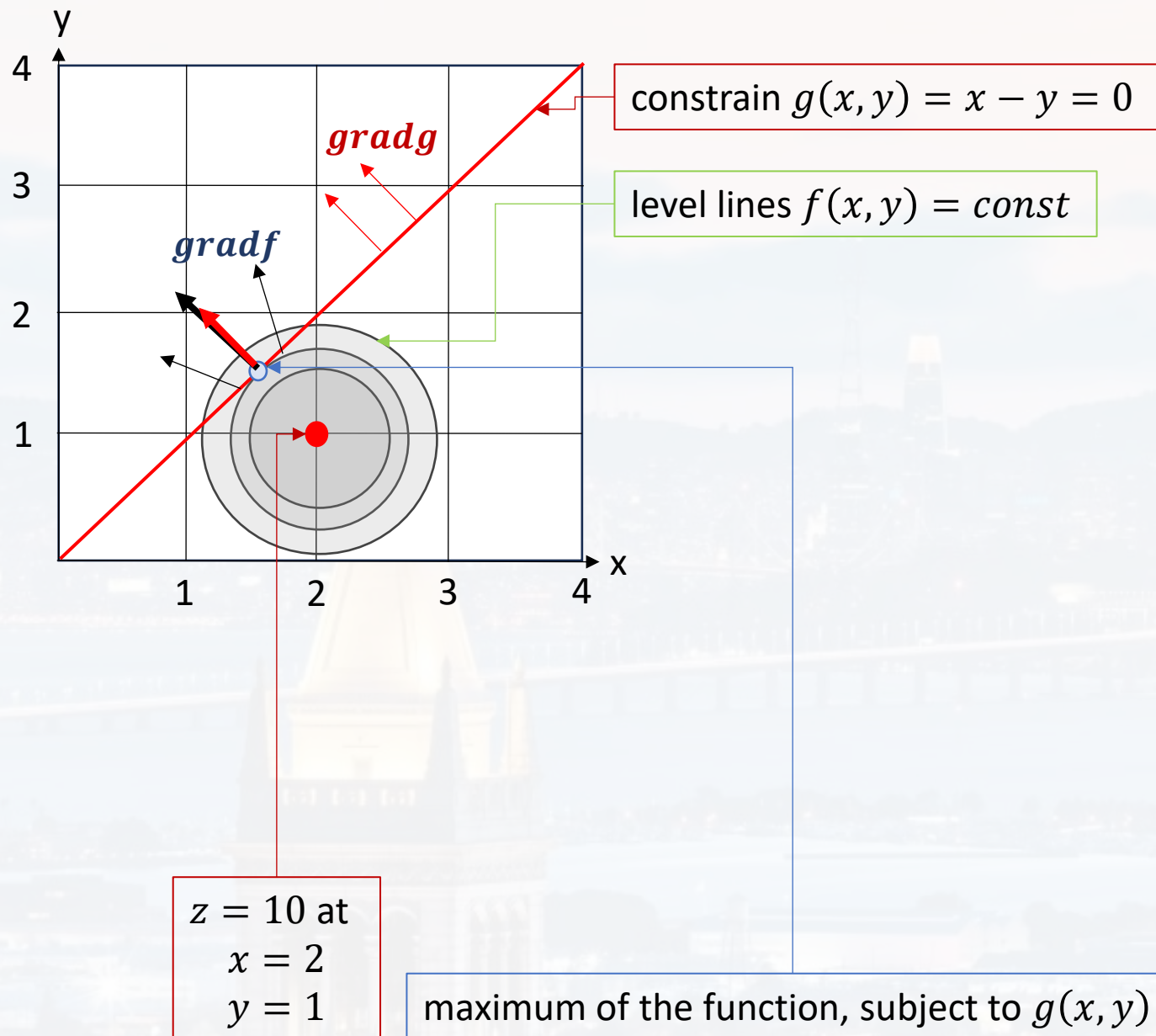
level lines $f(x, y) = \text{const}$

$$\begin{aligned} df(x, y) &= \frac{\partial f(x, y)}{\partial x} dx + \frac{\partial f(x, y)}{\partial y} dy = 0 \\ &= \text{grad} f \, d\vec{r} = 0 \end{aligned}$$

level lines $g(x, y) = \text{const}$

$$\begin{aligned} dg(x, y) &= \frac{\partial g(x, y)}{\partial x} dx + \frac{\partial g(x, y)}{\partial y} dy = 0 \\ &= \text{grad} g \, d\vec{r} = 0 \end{aligned}$$

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



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

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

$$\text{grad} f \, d\vec{r} = \text{grad} g \, d\vec{r}$$

$$\text{grad} f = \text{grad} g$$

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

$$\text{grad} f = \lambda \text{grad} g$$

λ **Lagrangian Multiplier**



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

Lagrangian Multiplier
Examples

$$\text{grad} f = \lambda \text{grad} g$$

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

the Lagrangian
 $L(x, y, \lambda)$

more general:

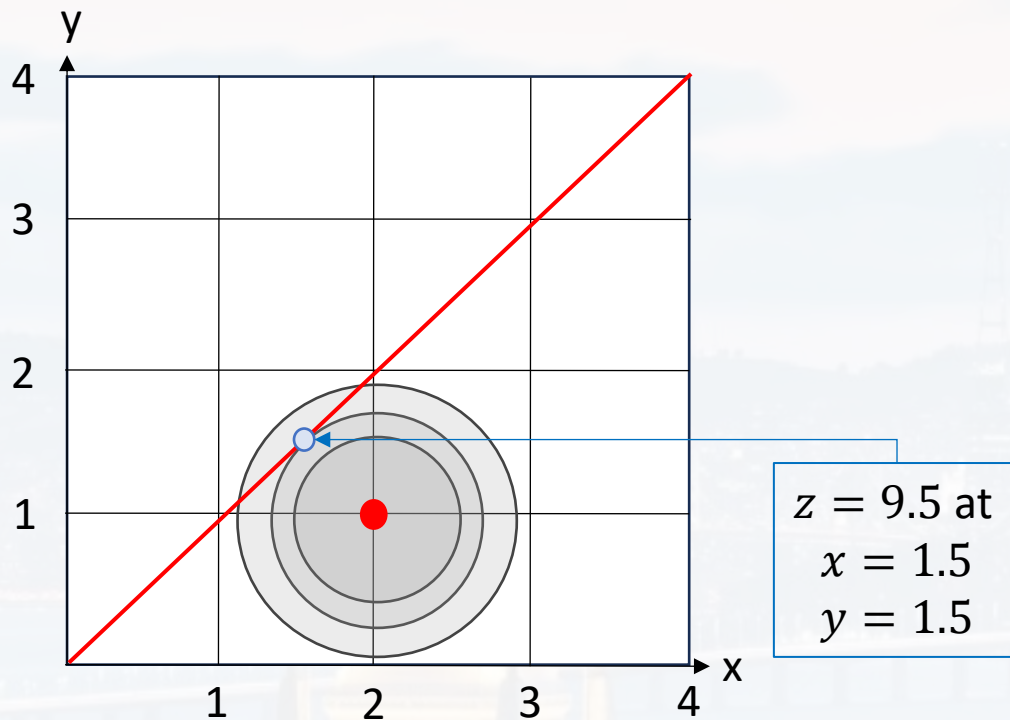
$$L(x_1, x_2, \dots, x_i, x_N, \lambda_1, \lambda_2, \dots, \lambda_k, \lambda_K) = f(x_1, x_2, \dots, x_i, x_N) - \sum_{k=1}^K \lambda_k g_k(x_1, x_2, \dots, x_i, x_N)$$

note:

- **N dimensions and $K \leq N$ 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)



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



maximum of the function **Lagrangian Multiplier**
Examples

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

$$\text{constrain } g(x, y) = x - y = 0$$

$$\text{constrain } x = y \quad \begin{aligned} x &= 1.5 \\ y &= 1.5 \end{aligned}$$

$$f(1.5, 1.5) = 9.5$$

$$\text{grad} f = \lambda \text{grad} g \quad \frac{\partial f(x, y)}{\partial x} = \lambda \frac{\partial g(x, y)}{\partial x}$$

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

$$-2(x - 2) = \lambda$$

$$-2(y - 1) = -\lambda$$

$$y = -x + 3$$



Outline

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:

Lagrangian Multiplier
Examples

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

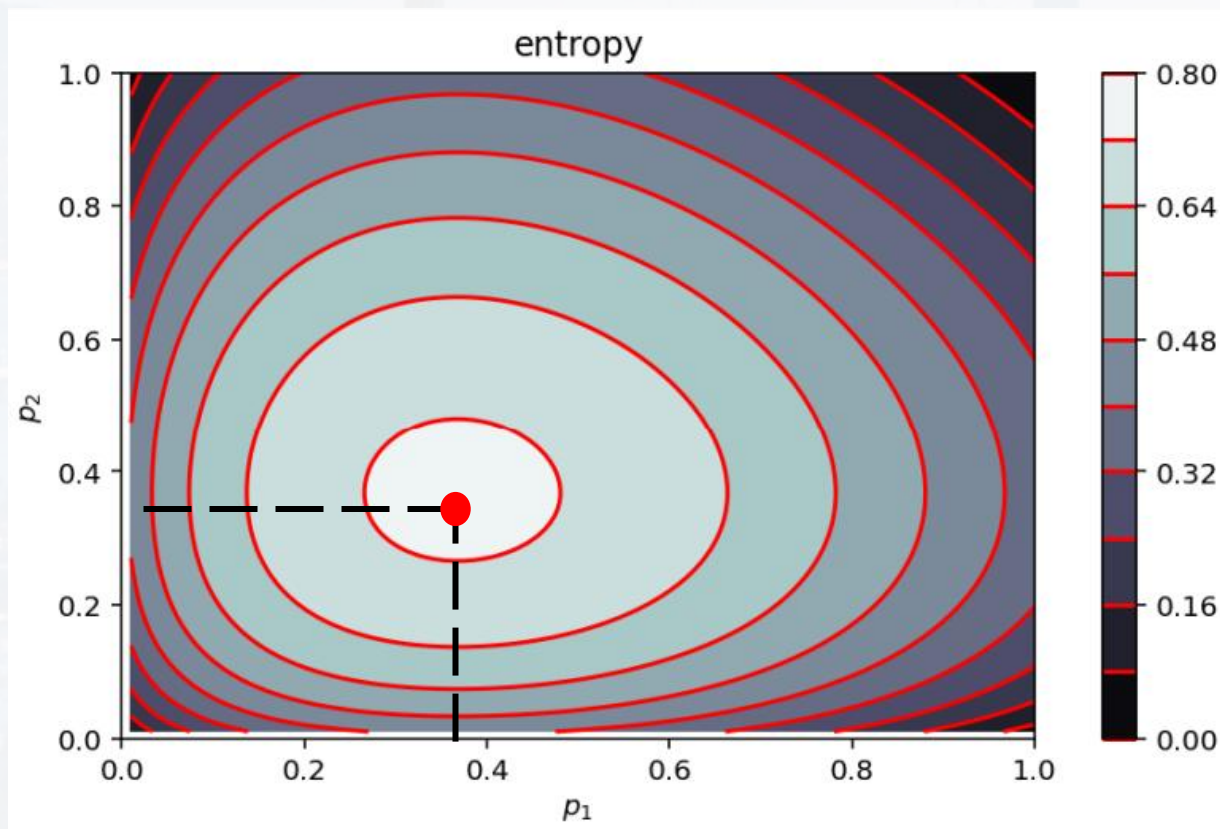
$$\frac{\partial f(p_1, p_2)}{\partial p_2} = 0$$

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

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

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

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





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 $g(p_1, p_2)$:

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

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

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

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

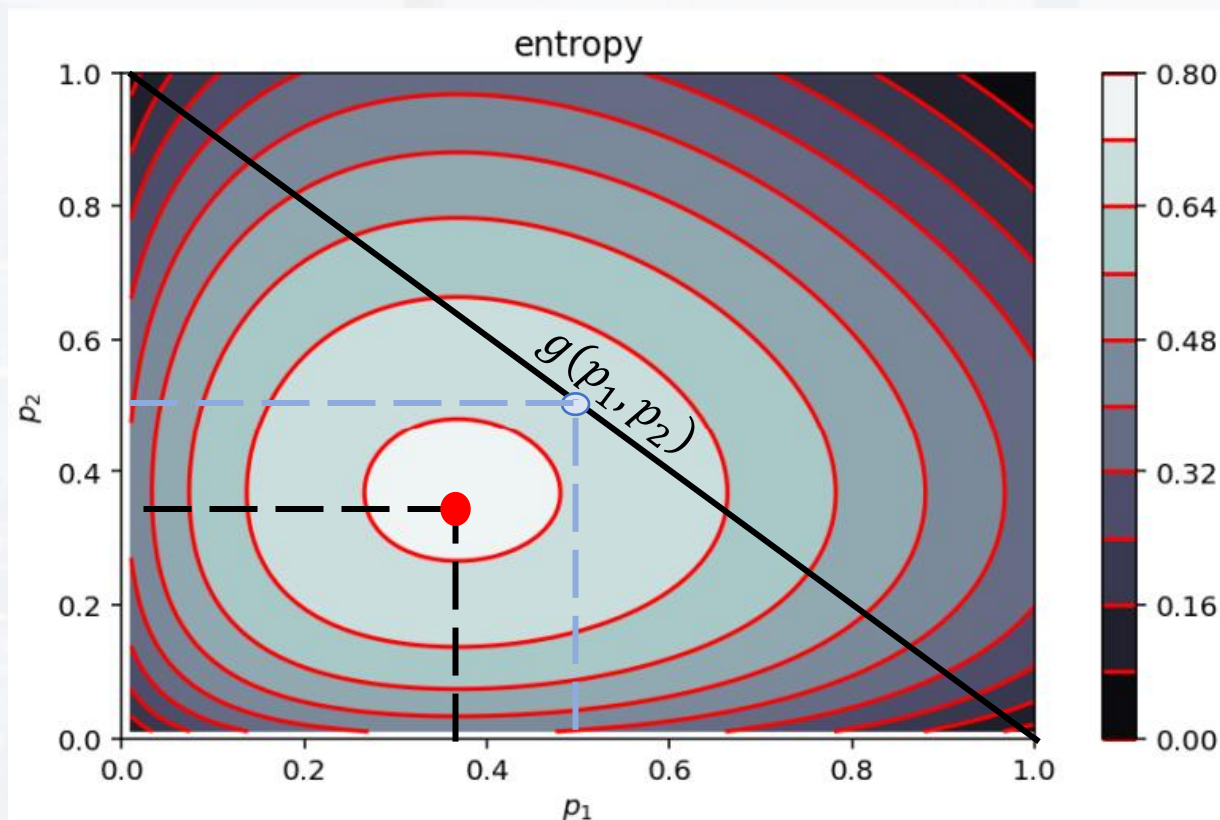
$$p_1 = p_2$$

constrain:

$$p_1 + p_2 = 1$$

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

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





Lagrangian Multiplier Examples

maximum entropy for I states:

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

subject to

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

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

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

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

constrain:

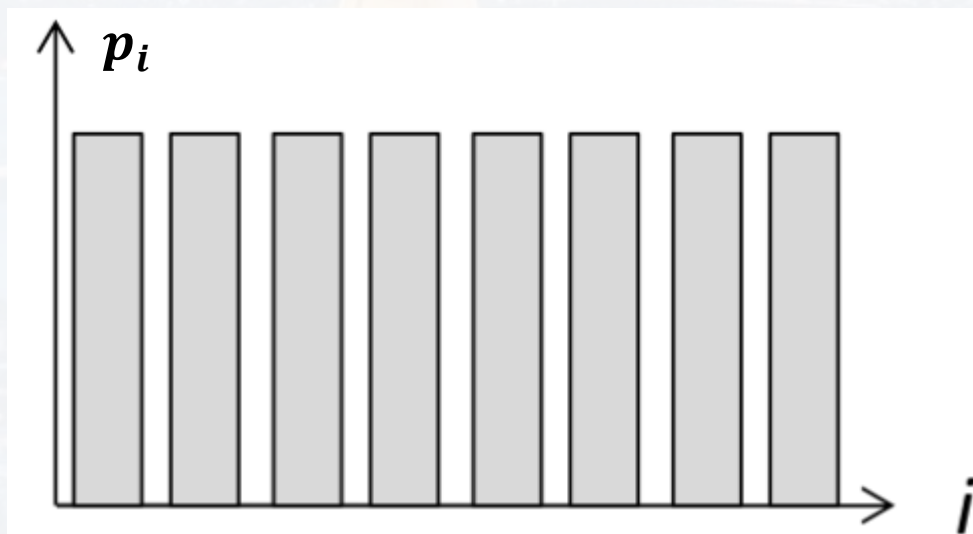
$$\sum_{i=1}^I e^{-(1+\lambda)} = 1$$

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

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

probabilities are constant!
→ flat distribution!

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





Lagrangian Multiplier Examples

maximum entropy for I states:

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

subject to

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

if N and **total energy** is conserved

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

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

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

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
ε_i :	energy in state 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:

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

subject to

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

if N and **total energy** is conserved

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

from g_1 :

$$p_i = \frac{1}{\sum_{i=1}^I e^{-\lambda_2 \varepsilon_i}} e^{-\lambda_2 \varepsilon_i}$$

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

partition function Z

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

$$-\ln p_i - 1 = \lambda_1 + \lambda_2 \varepsilon_i$$

$$p_i = e^{-(1+\lambda_1)} e^{-\lambda_2 \varepsilon_i}$$

Boltzmann distribution

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



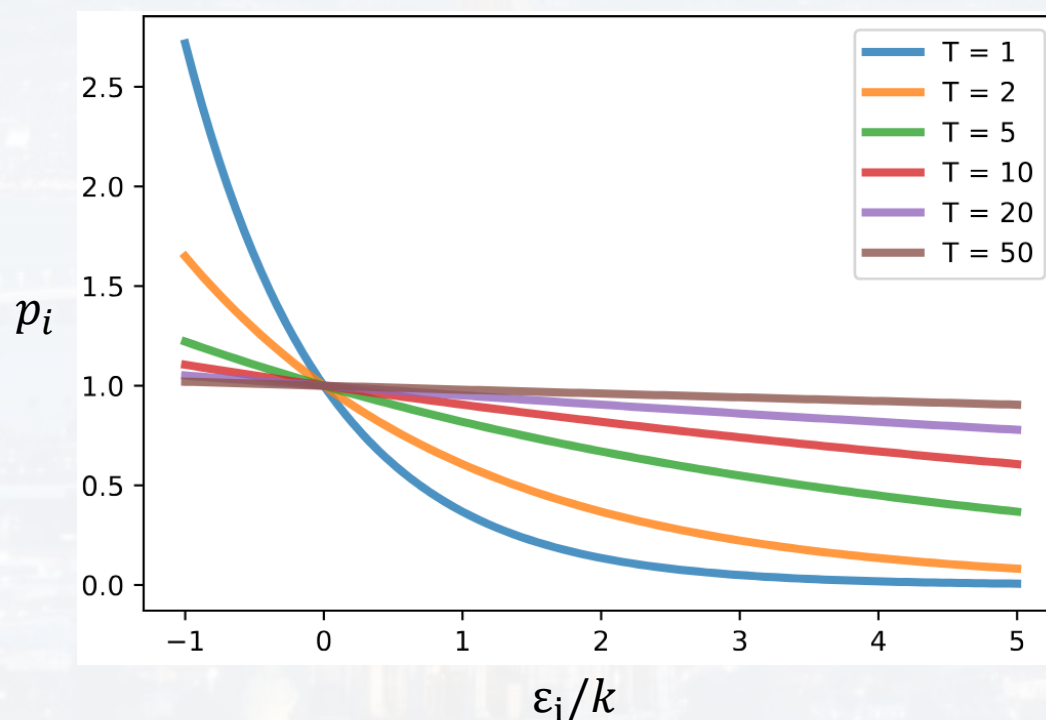
partition function Z

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

Boltzmann distribution

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

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



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

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



$$f(x) = - \int_{-\infty}^{+\infty} p(x) \ln[p(x)] dx \quad \text{continuous distribution with support } (-\infty, +\infty)$$

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

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

$$g_3 = \int_{-\infty}^{+\infty} (x - \mu)^2 p(x) dx = \sigma^2 \quad \text{standard deviation } \sigma^2$$

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

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

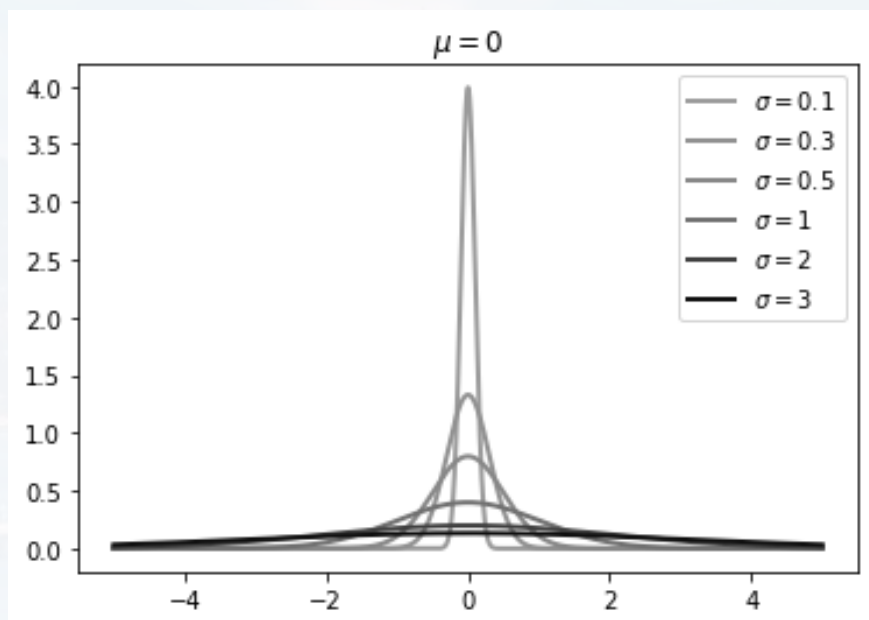
Normal (Gauss)
distribution

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



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 \rightarrow normal distribution is **often a good approximation**
 - often: $N(\mu, \sigma^2)$



maximum entropy distributions

Lagrangian Multiplier Examples

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



maximum entropy distributions

Lagrangian Multiplier Examples

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



Outline

Entropy and Information

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

Maximum Entropy Distributions

- Lagrangian Multiplier
- examples

Bayes Theorem



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

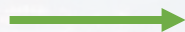
so far: A and B were independent $P(A \cap B) = P(A)P(B) = P(B)P(A)$

now: **conditional probabilities** | “given” or “under the condition”



Thomas Bayes
(1701 - 1761)

$$\begin{aligned} P(A \cap B) &= P(A|B)P(B) \\ &= P(B|A)P(A) \end{aligned}$$



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

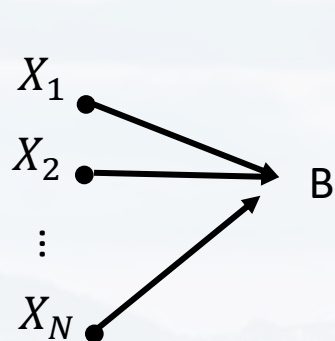
Bayes Theorem

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



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

Bayes Theorem



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

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

marginalization

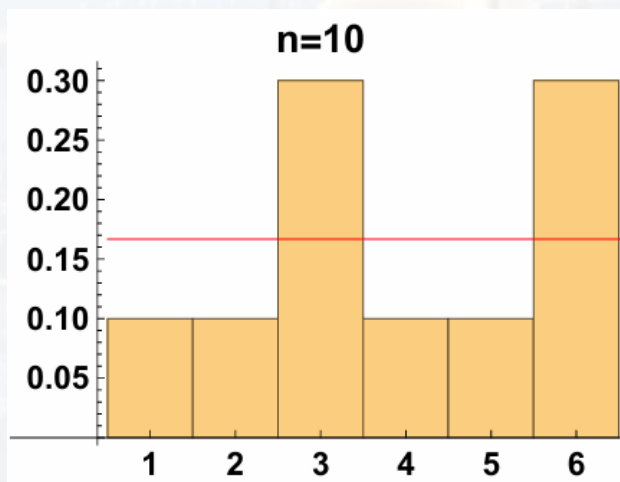


Thomas Bayes
(1701 - 1761)

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

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

model: M
data: D



$\sigma = 2, \mu = 3.5$
 $\sigma = 2, \mu = 5.0$
 $\sigma = 1.5, \mu = 3.5$
 $\sigma = 7.0, \mu = 1.0$
....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

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



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

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

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?**

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



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

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

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).

$$\begin{aligned}
 P(D|+) &= \frac{P(+|D)P(D)}{P(+)} = \underset{\substack{\uparrow \\ \text{statement 1}}}{0.95} \frac{P(D)}{P(+)} = \underset{\substack{\uparrow \\ \text{statement 2}}}{0.95 \cdot 0.00001} \frac{1}{P(+)} = \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)P(H)} \\
 &\qquad\qquad\qquad \text{marginalization} \\
 &= \frac{0.95 \cdot 0.00001}{P(+|D)P(D) + P(+|H)[1 - P(D)]} \qquad \text{complement probability}
 \end{aligned}$$



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

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

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).

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

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



example: cancer diagnosis from blood test

+ : positive test result
D : diseased
H : health

statement 1:

sensitivity

$P(D|+) = 95\%$

statement 2:

prior

$P(D) = 0.001\%$

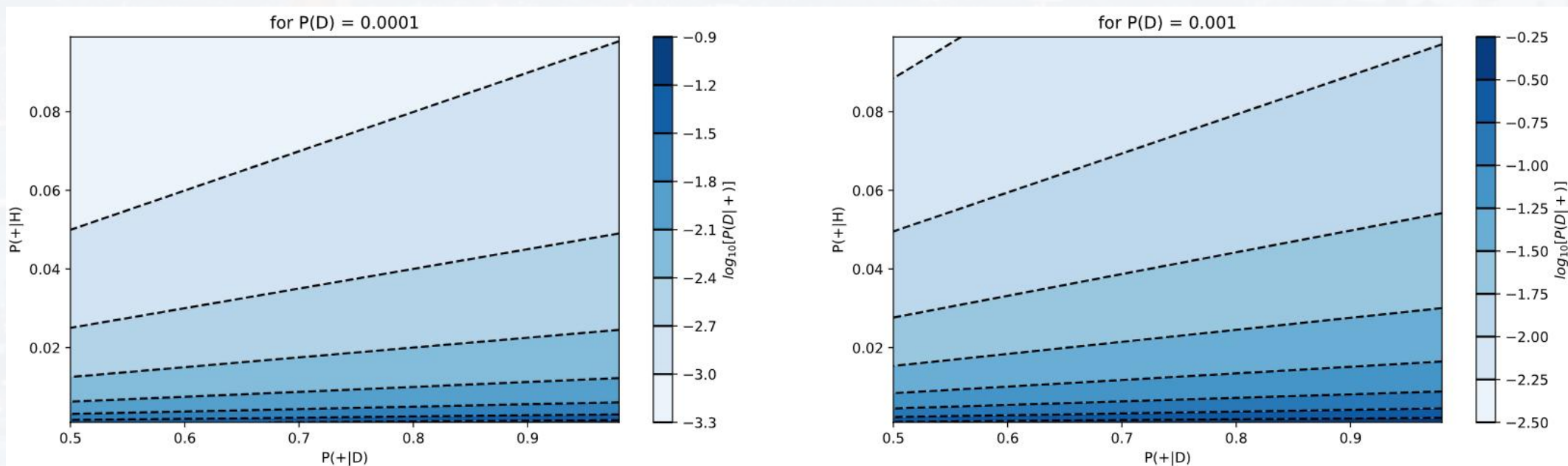
statement 3:

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`





example: cancer diagnosis from blood test

+ : positive test result
D : diseased
H : health

statement 1:

sensitivity

$P(D|+) = 95\%$

statement 2:

prior

$P(D) = 0.001\%$

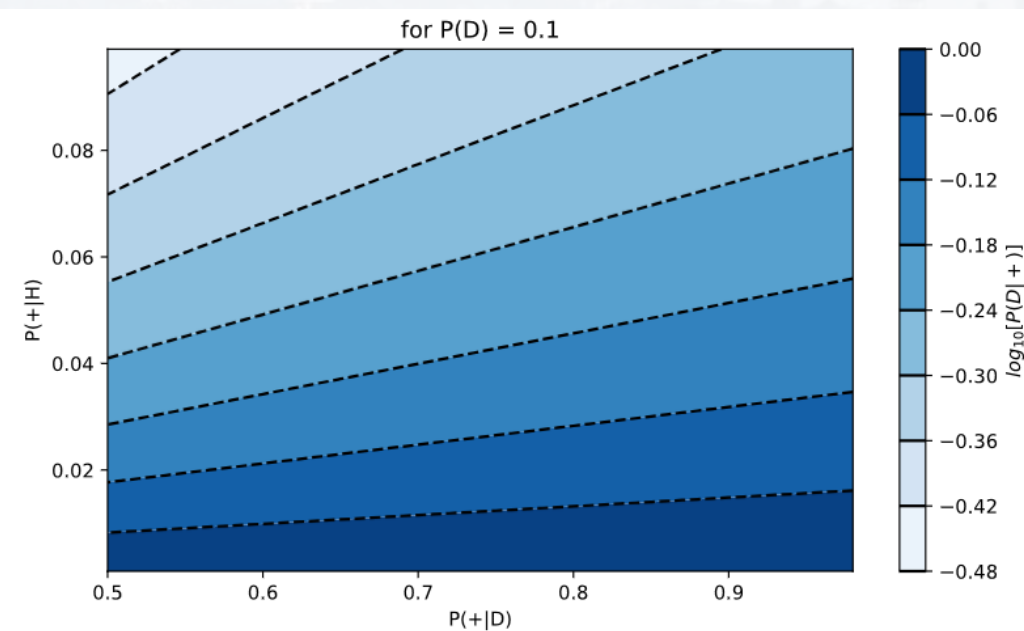
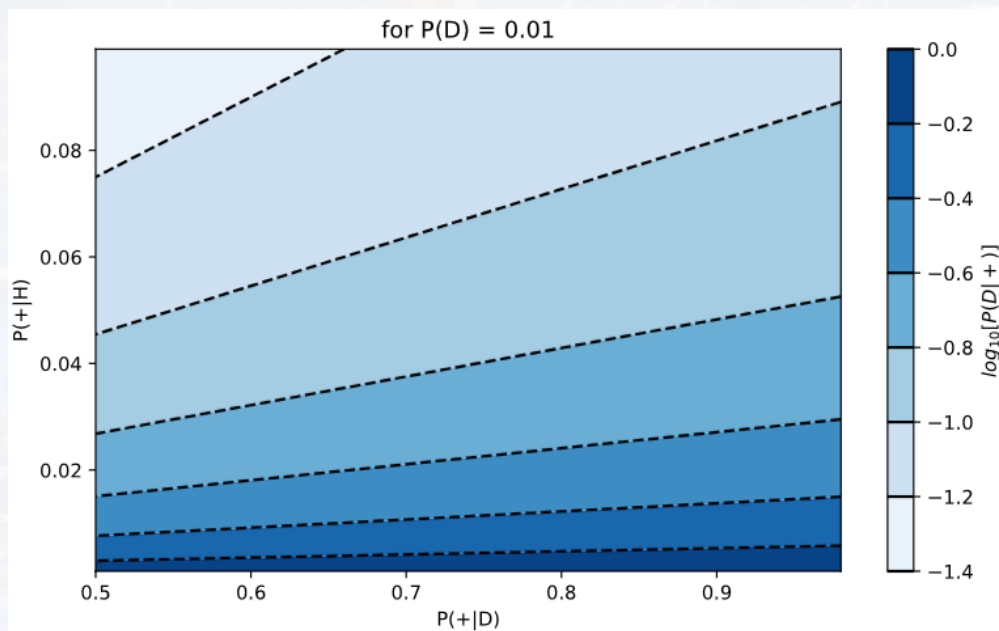
statement 3:

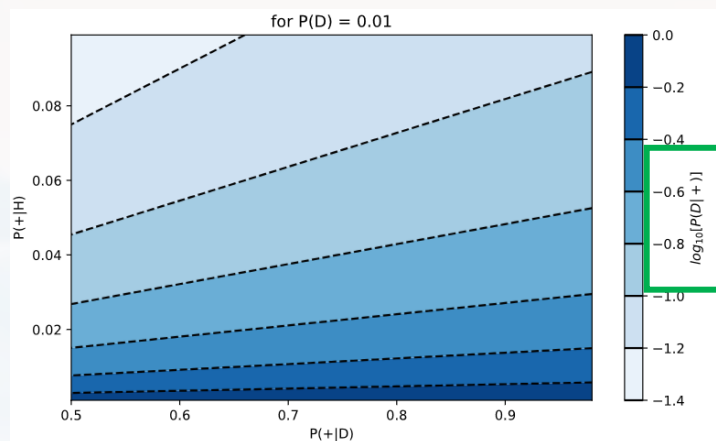
p-value or false positive rate

$P(+|H) = 5\%$

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

check: `PlotPD_Plus.py`





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

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

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



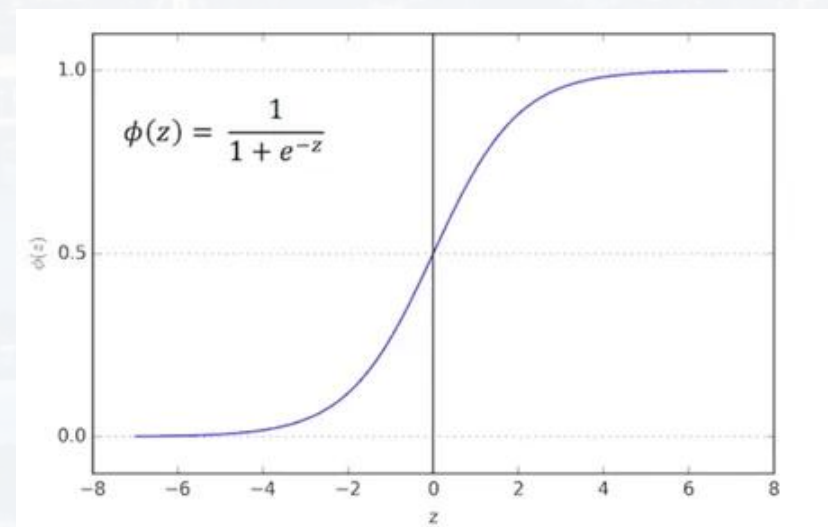
log odds ratios: $r_1 = \log \left[\frac{P(+|H)}{P(+|D)} \right]$

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

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

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

- note:**
- logistic (or logit or sigmoid) function
 - transfer function ANN (see later)
 - bound growth (Verhulst equation)





Thank you very much for your attention!

