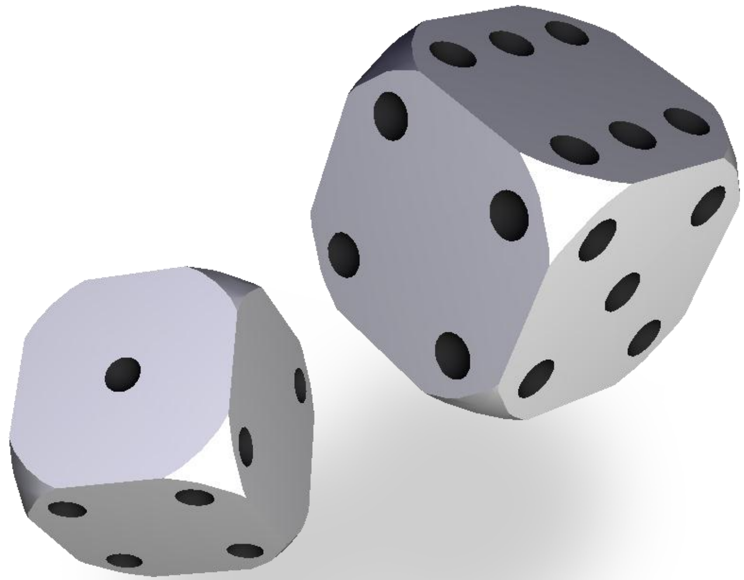
Statistical Geometry Processing

Winter Semester 2011/2012



Bayesian Statistics

UNIVERSITÄT DES SAARLANDES

M2CI CLUSTER OF EXCELLENCE

max planck institut informatik

Bayesian Statistics

Summary

- Importance
  - The only sound tool to handle uncertainty
  - Manifold applications: Web search to self-driving cars
- Structure
  - Probability: *positive*, *additive*, *normed* measure
  - Learning is density estimation
  - Large dimensions are the source of (almost) all evil
  - No free lunch: There is no universal learning strategy

Motivation

Modern AI

Classic **artificial intelligence:**

- Write a complex program with enough rules to understand the world
- This has been perceived as *not very successful*

Modern **artificial intelligence**

- Machine learning
- Learn structure from data
  - Minimal amount of “hardwired” rules
  - “Data driven approach”
- Mimics human development (training, early childhood)

Data Driven Computer Science

Statistical data analysis is everywhere:

- Cell phones (transmission, error correction)
- Structural biology
- Web search
- Credit card fraud detection
- Face recognition in point-and-shoot cameras
- ...

Probability Theory

(a very brief summary)

Probability Theory

(a very brief summary)

Part I: Philosophy

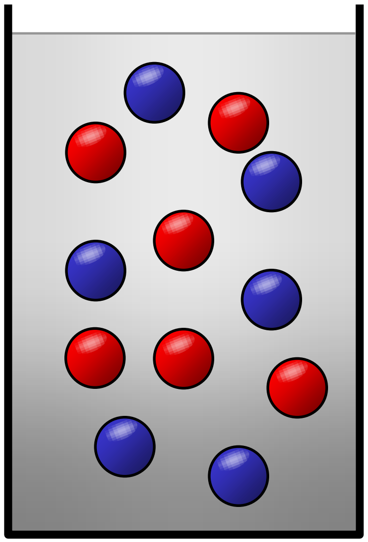
What is Probability?

Question:

- *What is probability?*

Example:

- A bin with 50 red and 50 blue balls
- Person A takes a ball
- Question to Person B:
  What is the probability for *red*?



What happened:

- Person A took a blue ball
- Not visible to person B

Philosophical Debate…

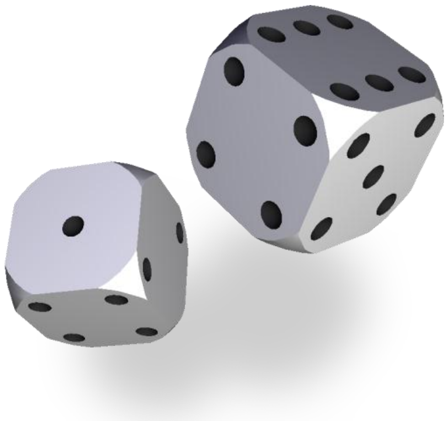
An old philosophical debate:

- What does *"probability"* actually mean?
- Can we assign probabilities to events for which the outcome is already fixed? (but we do not know it for sure)

"Fixed outcome" examples:

- Probability for life on mars
- Probability for J.F. Kennedy having been assassinated by a intra-government conspiracy
- Probability that the code you wrote is correct

Two Camps



Frequentists' (traditional) view:

- Well defined experiment
- Probability is the relative number of positive outcomes
- Only meaningful as a mean of many experiments

Bayesian view:

- Probability expresses a degree of belief
- Mathematical model of uncertainty
- Can be subjective

Mathematical Point of View

Mathematics:

- Math does not tell you what is true
- It only tells you the *consequences* if you accept other assumptions (axioms) to be true
- Mathematicians don’t do philosophy.

Mathematical definition of probability:

- Properties of probability measures
- Consistent with both views
- Defines rules for computing with probabilities
- Setting up probabilities is not a math problem

Probability Theory

(a very brief summary)

Part II: Probability Measures

Kolmogorov's Axioms

Discrete probability space:

• *Elementary events*:	$\Omega = \{\omega_1, \ldots, \omega_n\}$
• General *events*:	Subsets $A \subseteq \Omega$
• *Probability* measure:	$\Pr: \mathcal{P}(\Omega) \to \mathbb{R}$

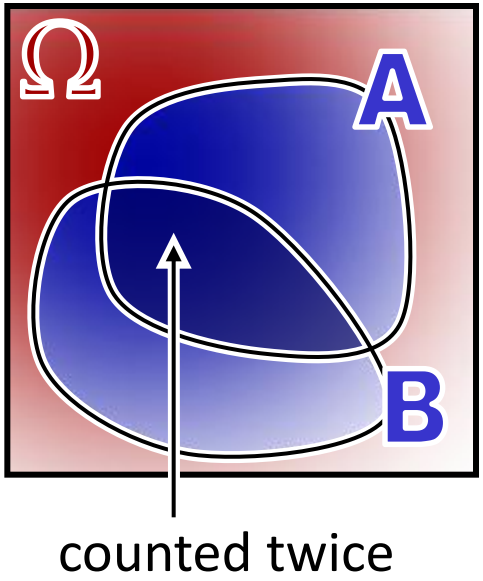
A valid probability measure must ensure:

• Positive:	$\Pr(A) \geq 0$
• Additive:	$[A \cap B = \emptyset] \Rightarrow [\Pr(A) + \Pr(B) = \Pr(A \cup B)]$
• Normed:	$\Pr(\Omega) = 1$

Other Properties Follow

Properties derived from Kolmogorov's Axioms:

- $P(A) \in [0..1]$
- $P(\bar{A}) = P(\Omega \setminus A) = 1 - P(A)$
- $P(\varnothing) = 0$
- $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$ ⇨
- …



In other words

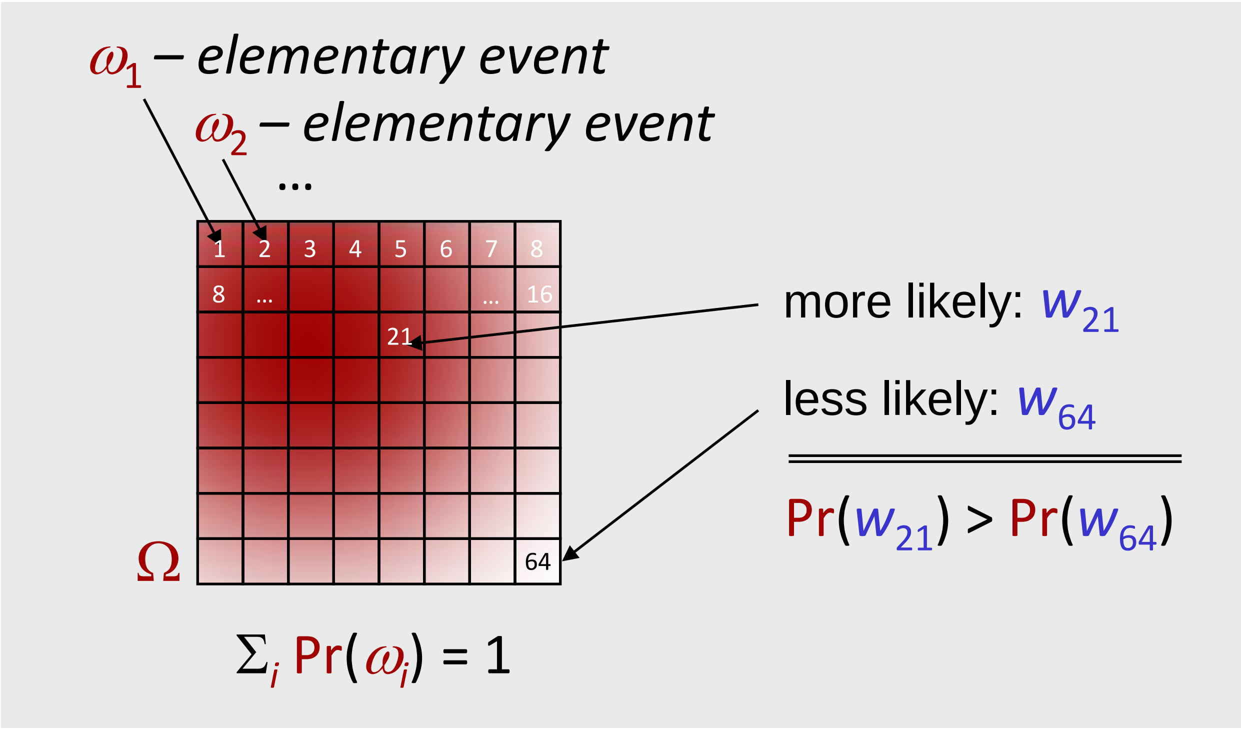
Mathematical probability is a

- *non-negative*, *normed*, *additive* measure.
  - Always ≥ 0
  - Sums to 1
  - Disjoint pieces add up

In other words

Mathematical probability is a

- *non-negative*, *normed*, *additive* measure.

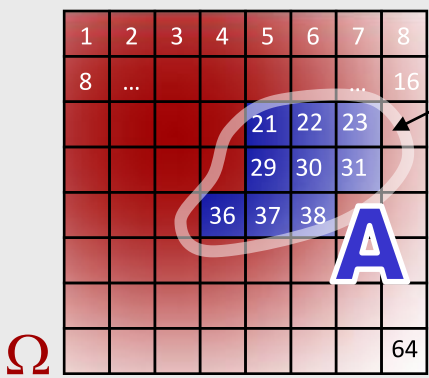


- Think of a *density* on some domain Ω

In other words

Mathematical probability is a

- *non-negative*, *normed*, *additive* measure.



A is an *event*

$$\Pr(A) = \Sigma_{i \in A} \Pr(\omega_i)$$

$$= \Pr(\omega_{21}) + \Pr(\omega_{22}) + \Pr(\omega_{23})$$
$$+ \Pr(\omega_{29}) + \Pr(\omega_{30}) + \Pr(\omega_{31})$$
$$+ \Pr(\omega_{36}) + \Pr(\omega_{37}) + \Pr(\omega_{38})$$

- Think of a *density* on some domain Ω

In other words

Mathematical probability is a

- *non-negative*, *normed*, *additive* measure.
  - Always ≥ 0
  - Sums to 1
  - Disjoint pieces add up

What does this model?

- You can always think of an area with density.
- All pieces are positive.
- Sum of densities is 1.

Discrete Models

Discrete probability space:

• *Elementary events*:	$\Omega = \{\omega_1, \ldots, \omega_n\}$
• General *events*:	Subsets $A \subseteq \Omega$
• *Probability* measure:	$\Pr: \mathcal{P}(\Omega) \to \mathbb{R}$

Probability measures:

- Sum of elementary probabilities

$$\Pr(A) = \sum_{\omega_i \in A} \Pr(\omega_i)$$

Continuous Probability Measures

Continuous probability space:

• *Elementary events*:	$\Omega \subseteq \mathbb{R}^d$
• General *events*:	"reasonable"*) subsets $A \subseteq \Omega$
• *Probability* measure:	$\Pr: \sigma(\Omega) \rightarrow \mathbb{R}$ assigns probability to subsets*) of Ω

*) not "*all*" subsets: Borel sigma algebra (details omitted)

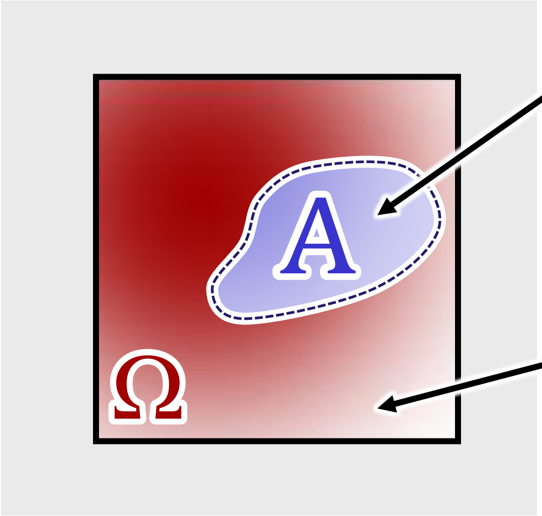
The same axioms:

• Positive:	$\Pr(A) \geq 0$
• Additive:	$[A \cap B = \varnothing] \Rightarrow [\Pr(A) + \Pr(B) = \Pr(A \cup B)]$
• Normed:	$P(\Omega) = 1$

Continuous Density

Density model

- No elementary probabilities
- Instead: density p: $\mathbb{R}^d \rightarrow \mathbb{R}^{\geq 0}$



A is an *event*

$$\Pr(A) = \int_A p(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

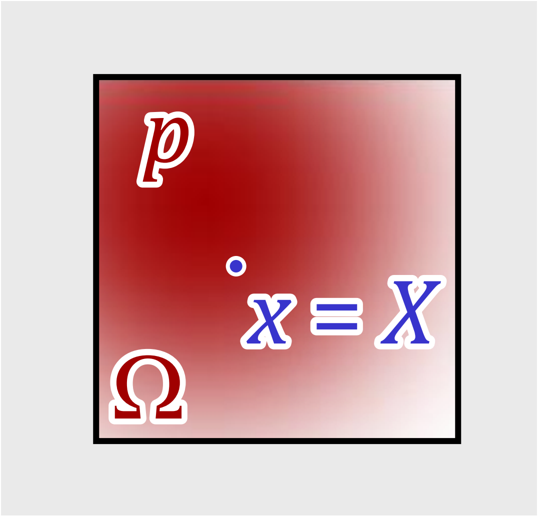
Density $p(\mathbf{x})$ with

$$p(\mathbf{x}) \geq 0 \text{ and } \int_\Omega p(\mathbf{x})\, \mathrm{d}\mathbf{x} = 1$$

Random Variables

Random Variables

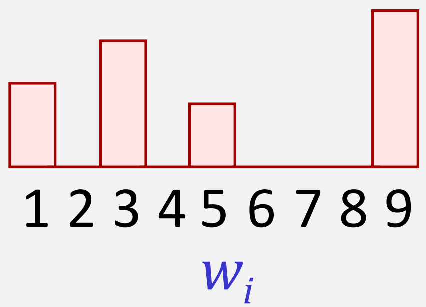
- Assign numbers or vectors from $\mathbb{R}^d$ to outcomes
- Notation:
  - random variable X
  - density $p(x) = \Pr(X = x)$
- Usually:
  Variable = domain of the density



Unified View

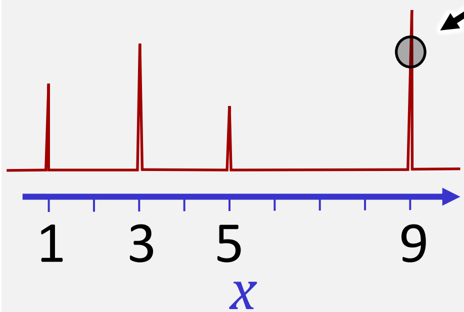
Discrete models as special case

$p(w_i)$, $w_i \in \{1,\dots,9\}$



Discrete model

$p(x)$, $x \in \mathbb{R}$



Continuous model

Dirac-Delta pulses

$p(x) = \Sigma_i\, \delta(x - x_i)\, p(w_i)$

Idealization

$\int_{\mathbb{R}^d} \delta(x)\, \mathrm{d}x = 1$

$\delta(0)$ very large

$\mathrm{d}(\mathrm{x}) = 0$ everywhere else

Probability Theory

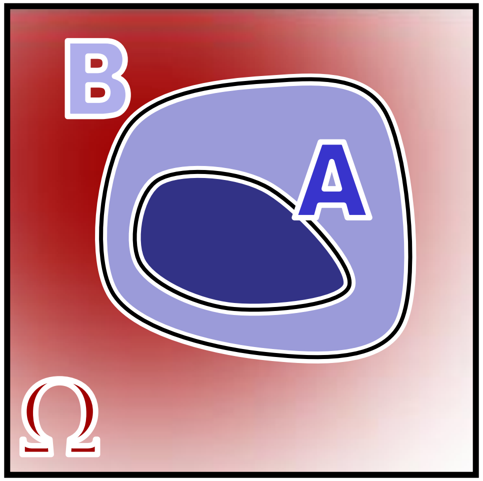
(a very brief summary)

Part III: Statistical Dependence

Conditional Probability

Conditional Probability:

- $\Pr(A \mid B)$ = Probability of A given B [is true]
- Easy to show:
  $\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B)$

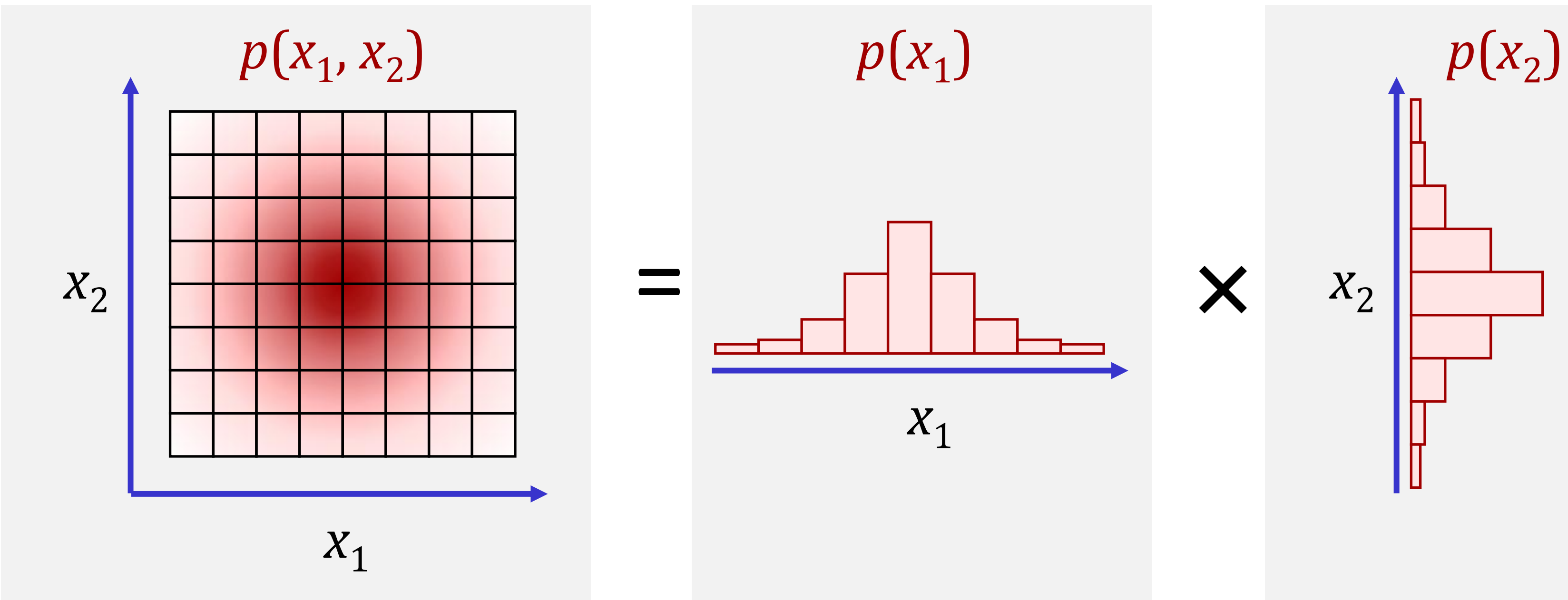


Statistical Independence

- A and B independent
  $:\Leftrightarrow \Pr(A \cap B) = \Pr(A) \cdot \Pr(B)$
- Knowing the value of A does not yield information about B (and vice versa)

Factorization

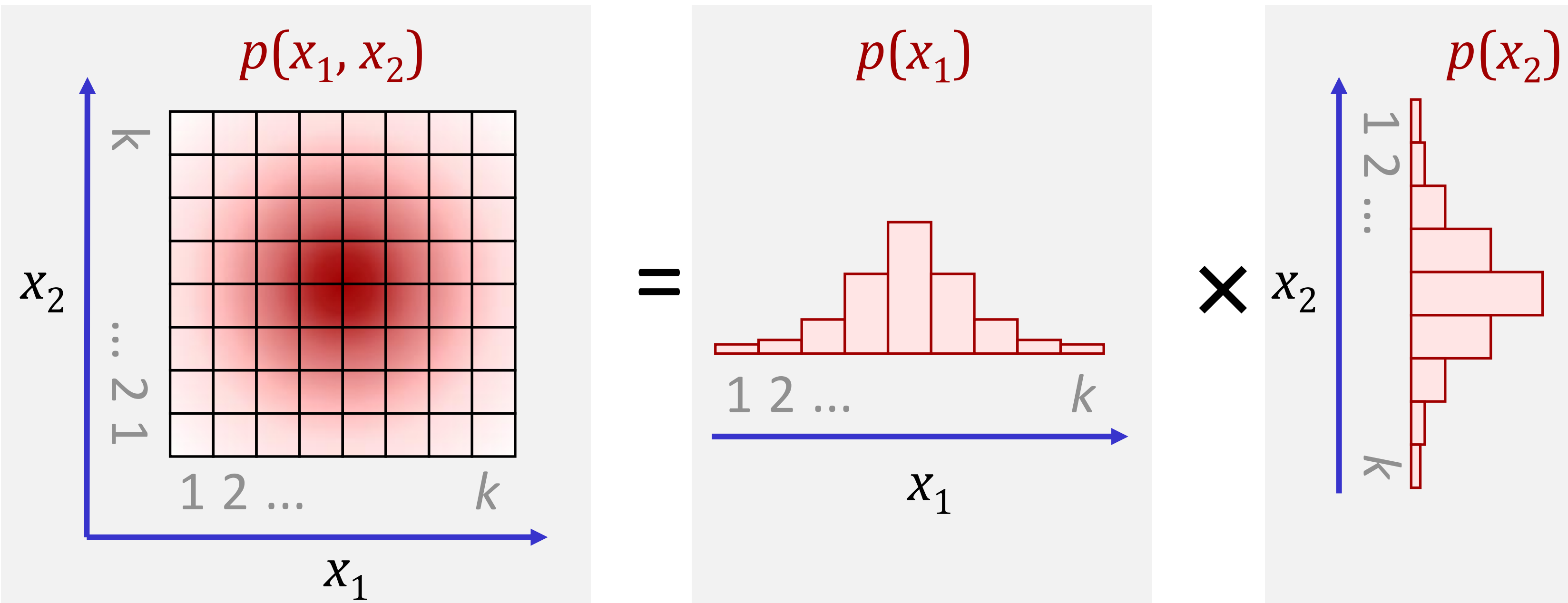
Independence = Density Factorization



$$p(x_1, x_2) = p(x_1) \times p(x_2)$$

Factorization

Independence = Density Factorization



$$p(x_1, x_2) = p(x_1) \times p(x_2)$$

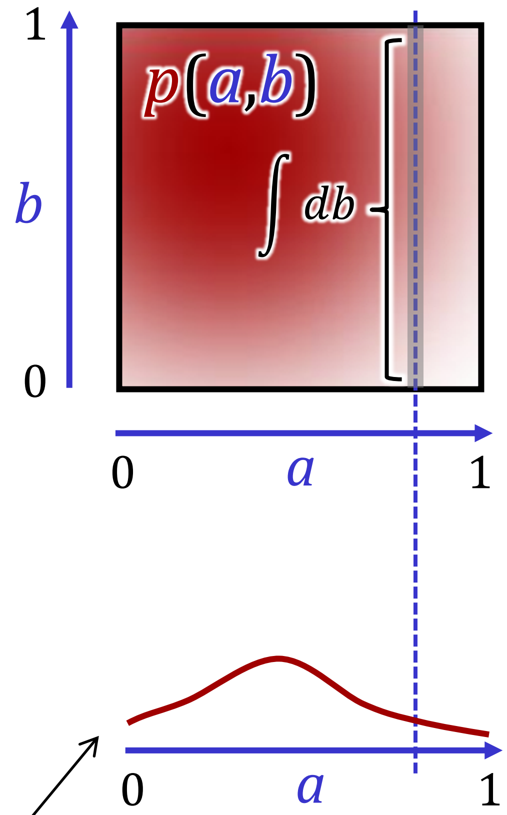
$O(k^d)$ $\qquad$ $O(d \cdot k)$

Marginals

Example

- Two random variables
  $a, b \in [0,1]$
- Joint distribution $p(a, \mathrm{b})$
- We do not know b
  (could by anything)
- What is the distribution of a?

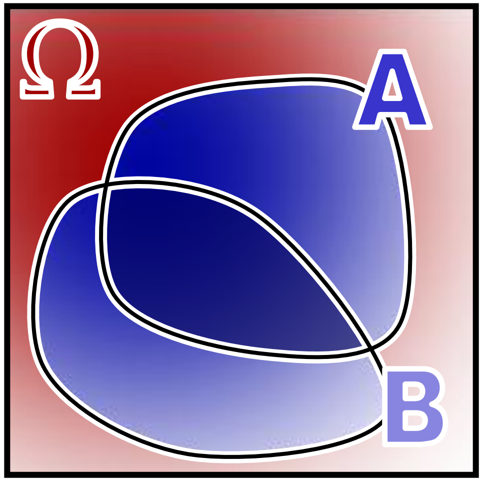
$$p(a) = \int_0^1 p(a,b)\,db$$



"Marginal Probability"

Conditional Probability

Bayes' Rule:



$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

Derivation

- $\Pr(A \cap B) = \Pr(A \mid B) \cdot \Pr(B)$
  $\Pr(A \cap B) = \Pr(B \mid A) \cdot \Pr(A)$

$$\Rightarrow \Pr(A \mid B) \cdot \Pr(B) = \Pr(B \mid A) \cdot \Pr(A)$$

Bayesian Inference

Example: Statistical Inference

- Medical test to check for a medical condition
- A: *Medical test positive?*
  - 99% correct if patient is ill
  - But in 1 of 100 cases, reports illness for healthy patients
- B: *Patient has disease?*
  - We know: One in 10 000 people have it

A patient is diagnosed with the disease:

- How likely is it for the patient to actually be sick?

Bayesian Inference

Apply Bayes' Rule:

A: *Medical test positive?*
B: *Patient has disease?*

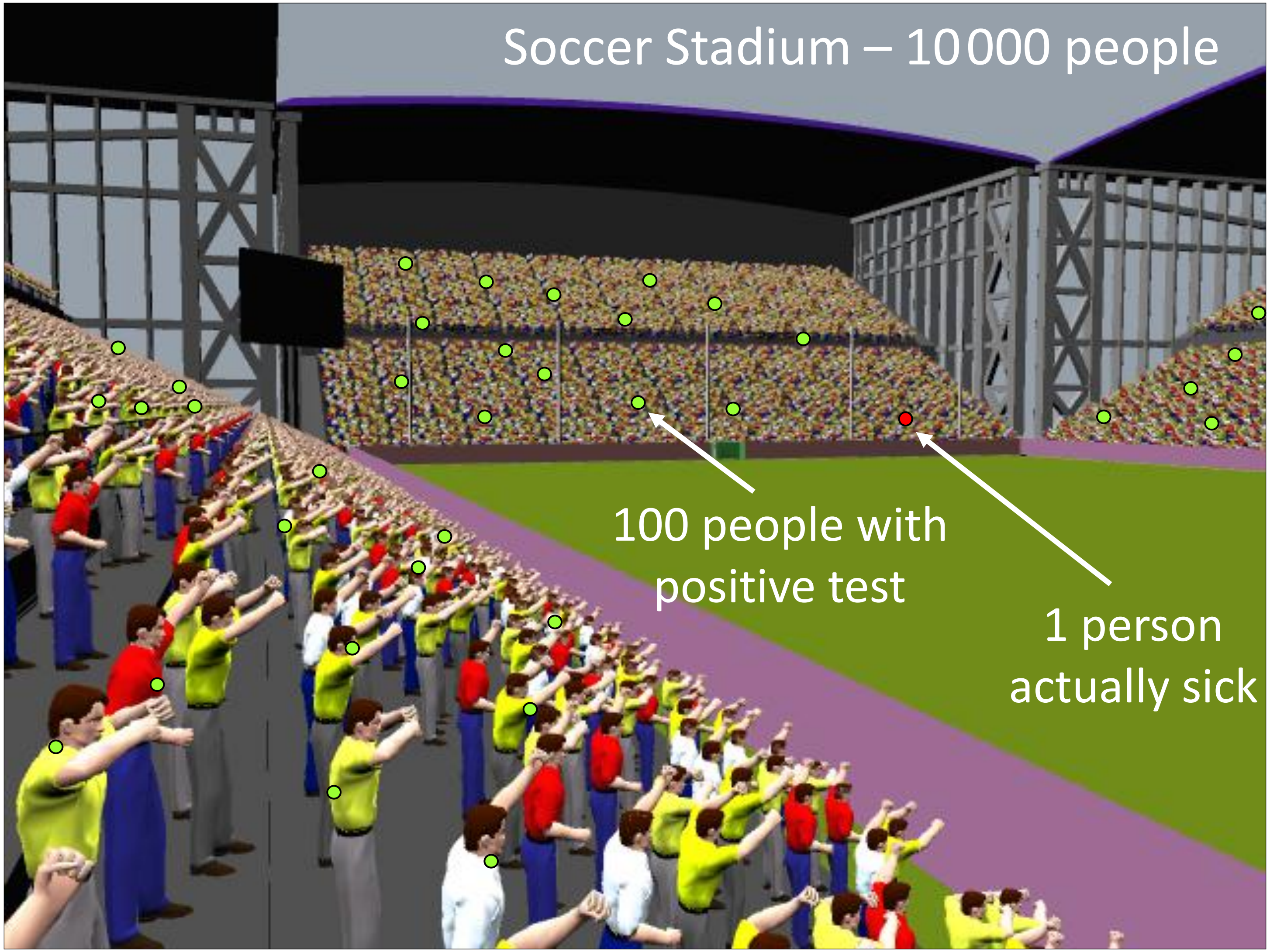
$$\Pr(B \mid A) = \frac{\Pr(A \mid B) \cdot \Pr(B)}{\Pr(A)}$$

$$\Pr(\text{disease} \mid \text{test positive}) = \frac{\Pr(\text{test pos.} \mid \text{disease}) \cdot \Pr(\text{deasease})}{\Pr(\text{test pos.} \mid \text{disease})\Pr(\text{disease}) + \Pr(\text{test pos.} \mid \overline{\text{disease}})\Pr(\overline{\text{disease}})}$$

$$= \frac{0.99 \cdot 0.0001}{0.99 \cdot 0.0001 + 0.01 \cdot 0.9999} = \frac{0.000099}{0.0100979901}$$

$$\approx 0.0098 \approx \frac{1}{100} \quad \leftarrow \text{most likely healthy}$$

Intuition


Soccer Stadium – 10 000 people
100 people with positive test
1 person actually sick

Conclusion

$$\Pr(A \mid B) = \frac{\Pr(B \mid A) \cdot \Pr(A)}{\Pr(B)}$$

Bayes' Rule:

- Used to fuse knowledge
  - "Prior" knowledge (prevalence of disease)
  - "Measurement": tests, sensor data, new information
  - Can be used repeatedly to add more information
- Standard tool for interpreting sensor measurements (Sensor fusion, reconstruction)
- Examples:
  - Image reconstruction (noisy sensors)
  - Face recognition

Chain Rule

Incremental update

- Probability can be split into chain of conditional probabilities:

$$\Pr(X_n, \dots, X_2, X_1)$$
$$= \Pr(X_n | X_{n-1}, X_{n-2}, \dots, X_1) \cdots \Pr(X_3 | X_2, X_1) \Pr(X_2 | X_1) \Pr(X_1)$$

- Example application:
  - X_i is measurement at time i
  - Update probability distribution as more data comes in
- Attention – although it might look like, this does not reduce the complexity of the joint distribution

Probability Theory

(a very brief summary)

Part IV: Uniqueness – Philosophy Again...

Cox Axioms

Are there alternatives?

- Is this the right way to define probabilities?
- Are there no other uncertainty measures?

Answer (short):

- Yes.
- Any reasonable*) probability measure has the same properties
  - Up to normalization constant; we can have $\Pr \in [0..42]$ if we like

*) reasonable – Cox axioms:
ordering $\Pr(A) > \Pr(B) > \Pr(C)$ well defined, $\Pr(\overline{A}) = f(\Pr(A))$,
$\Pr(A \cap B) = g(\Pr(A|B), \Pr(B))$ for arbitrary, fixed f, g.

What is Probability?

Principle #1: [Hertzman 2004]

"*Probability theory is nothing more than common sense reduced to calculation*"

Pierre-Simon Laplace, 1814

Principle #2,3: [Hertzman 2004]

- Given a complete model, we can compute any other probability
- Use Bayes rule to infer unknown variables from observations

Probability Theory

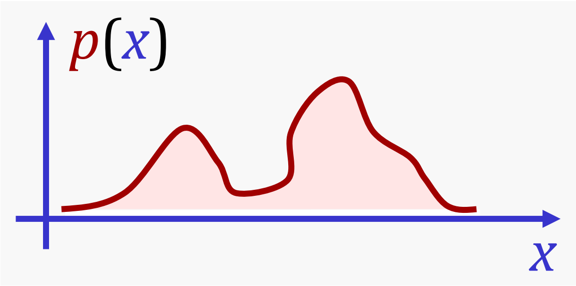
(a very brief summary)

Part IV: Characteristics of Probability Measures

Moments of Distributions

Density Function (1D)

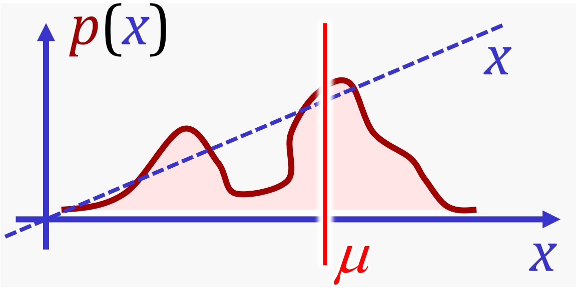
- $p: \mathbb{R} \to \mathbb{R}^{\geq 0}$



Expected Value / Mean:

- $E(p) = \mu := \langle p, x \rangle$

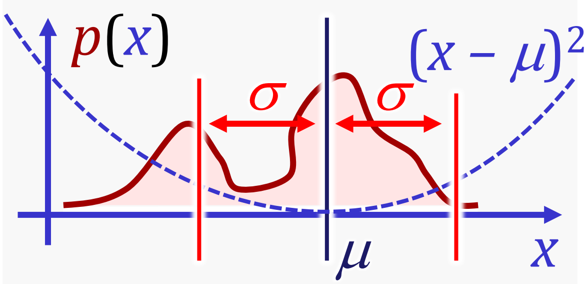
$$= \int_{\mathbb{R}} p(x) \cdot x \, dx$$



Variance:

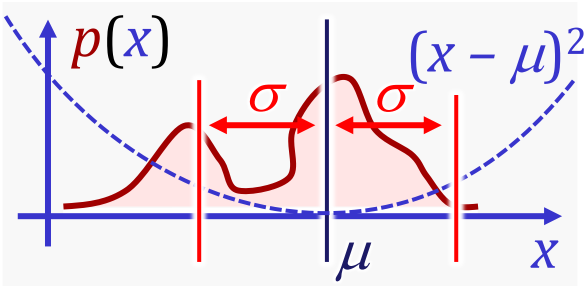
- $Var(p) = \sigma^2 := \langle p, (x - \mu)^2 \rangle$

$$= \int_{\mathbb{R}} p(x) \cdot (x - \mu)^2 \, dx$$



Standard Deviation

Bounds on spread



- Standard deviation

$$\sigma = \sqrt{Var(p)}$$

- Expected range of variations
- Bounds spread of the distribution
- Formal bound: Chebyshev's inequality

$$\Pr(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Remark: Other Moments

Higher order moments:

- $m_k(p) \coloneqq \langle p, (x-\mu)^k \rangle = \int_{\mathbb{R}} p(x) \cdot (x-\mu)^k \, dx$
- Skewness: m_3 (asymetry of the distribution)
- Kurtosis: m_4 (peakedness)

More general

- $\langle p, f_i \rangle$ with basis functions f_i, for example:
  - Fourier basis („characteristic function“)

We will not use any of this in this lecture...

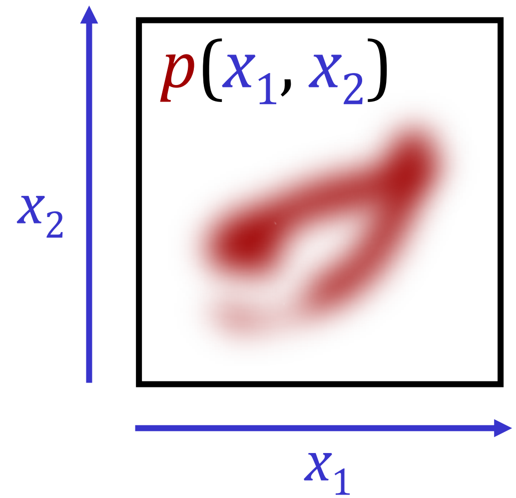
Moments of Distributions

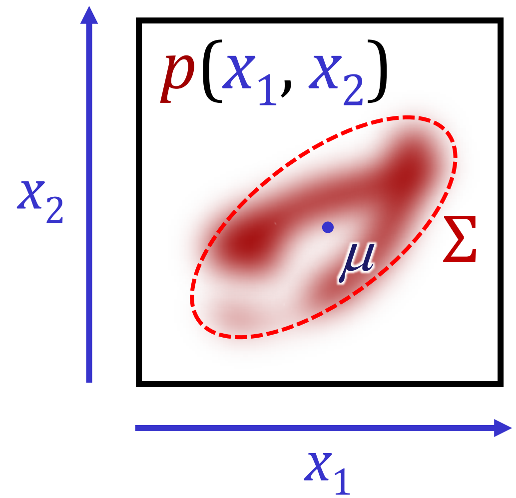
Multi-variate density function

- Density $p\colon \mathbb{R}^d \to \mathbb{R}^{\geq 0}$
- $E(p) = \mu := \langle p, x \rangle = \int_{\mathbb{R}^d} p(x) \cdot x \, dx$
- $\mathrm{Cov}(x_i, x_j) := \langle p, (x_i - \mu_i)(x_j - \mu_i) \rangle$

  $$= \int_{\mathbb{R}^d} p(x)\,(x_i - \mu_i)(x_j - \mu_i)\, dx$$

- $$\Sigma = \begin{pmatrix} \ddots & \vdots & ⋰ \\ \cdots & \mathrm{Cov}(x_i, x_j) & \cdots \\ ⋰ & \vdots & \ddots \end{pmatrix}$$





Properties

Expected Value:

- $E(X+Y) = E(X) + E(Y)$
- $E(\lambda X) = \lambda E(X)$

Variance:

- $\mathrm{Var}(\lambda X) = \lambda^2 \mathrm{Var}(X)$
- Let X, Y be *independent*, then:
  $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

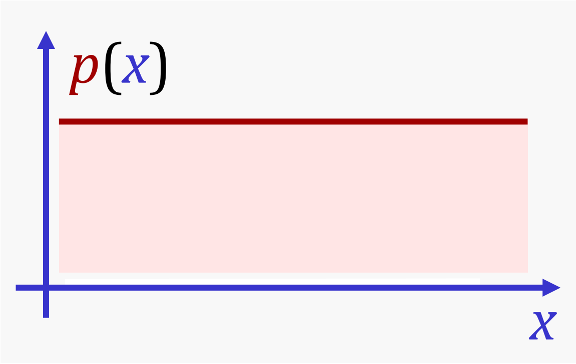
Entropy

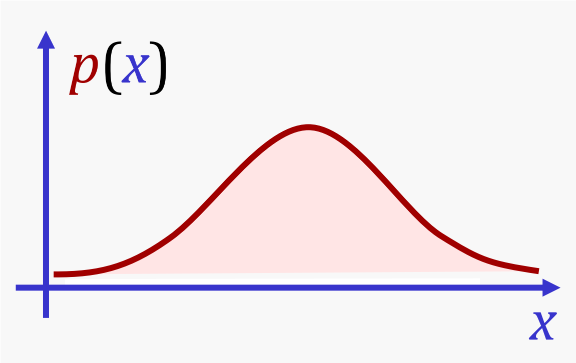
How random is the randomness?

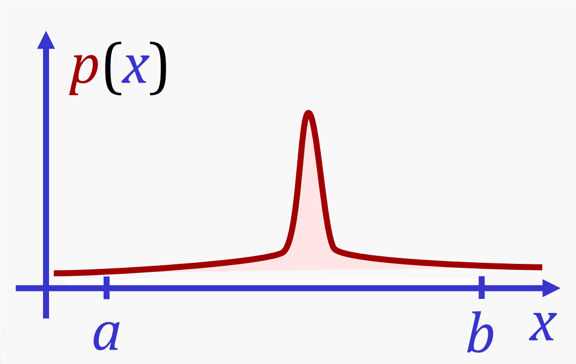
- Measure of unorderliness
- How much information remains in the events, knowing the distribution?

Idea

- Try to code the events
- Binary codes
  - short codes for frequent events
  - long codes for infrequent events







Entropy

Best solution

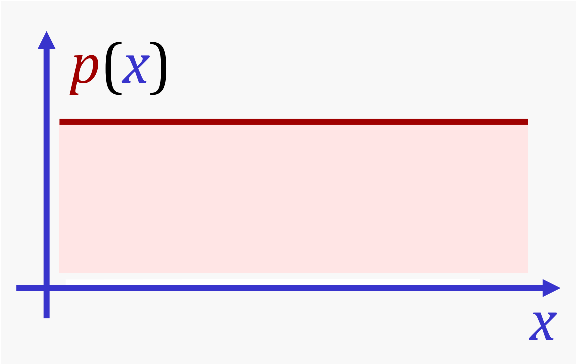
- Use codes of $\mathcal{O}(\log \frac{1}{p})$ bits for events with probability p
- Can be implemented: Huffman coding, arithmetic coding

Definition: Entropy

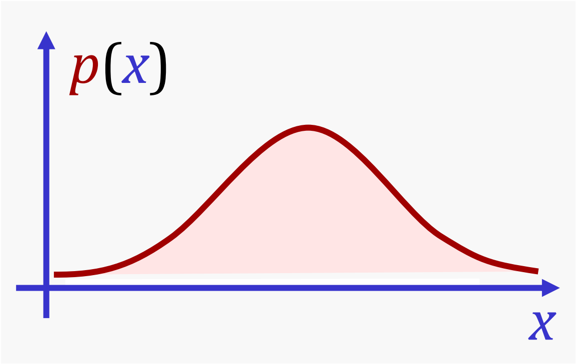
$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

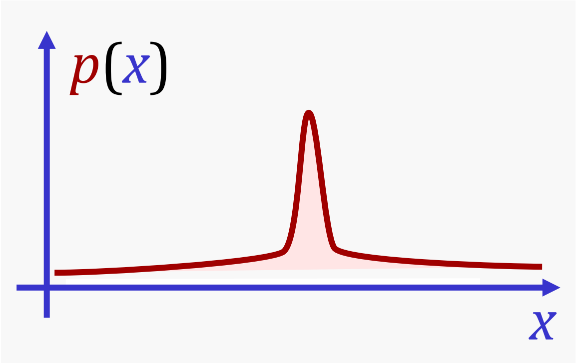
- Coding efficiency of independent events

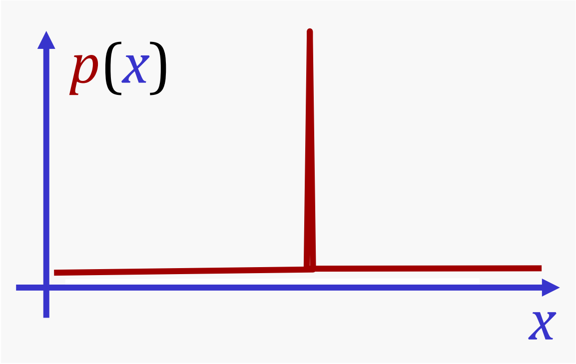
Examples



$$H = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$







$$H = 0$$

Probability Theory

(a very brief summary)

Part V: Large numbers

Law of Large Numbers

Intuition for Probabilities:

- Single outcomes are random
- But on average over a larger number of trials, the behavior is known
- It can be shown that probability measures naturally have this property

Law of Large Numbers

Let

- $X_1, X_2, \ldots, X_n$ be i.i.d. random variables (independent, identically distributed)

We look at the mean

$$\bar{X}_n = \frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)$$

(Weak) law of large numbers

$$\lim_{n \to \infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0$$

Proof

Proof:

- Additionally assumption: finite variance $\mathrm{Var}(X_i) = \sigma^2$
- The theorem then follows from
  - Additivity of variances
  - Chebyshev's bound

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i\right)\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n} \mathrm{Var}(X_i)\right) = \frac{n\sigma}{n^2} = \frac{\sigma}{n}$$

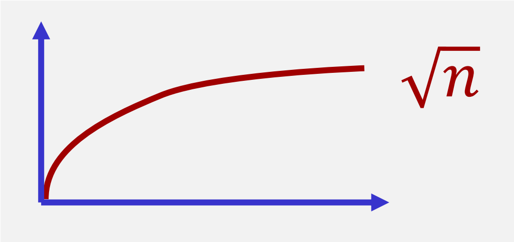
$$\Rightarrow \sigma(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}$$

- Chebyshev: $\Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Additional Insight

Averaging of independent trials

- Reduces the variance
- For independent sampling, convergence rate is $\frac{1}{\sqrt{n}}$
- This is usually lousy...
  - Rapid progress first
  - Then takes forever to converge



Central Limit Theorem

Why are so many phenomena normal-distributed?

- Let $X_1, \ldots, X_n$ be real (1D) random variables with means μ_i and *finite* variances σ_i^2.
- Then the distribution of the mean

$$\frac{\sum_{i=1}^{n} X_i \; - \; \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \quad \rightarrow \quad \mathcal{N}(0,1)$$

  converges to a normal distribution.

Multi-dimensional variant

- Similar result for multi-dimensional case

Probability Theory

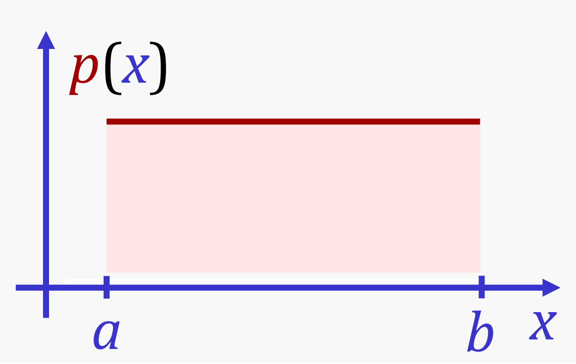
(a very brief summary)

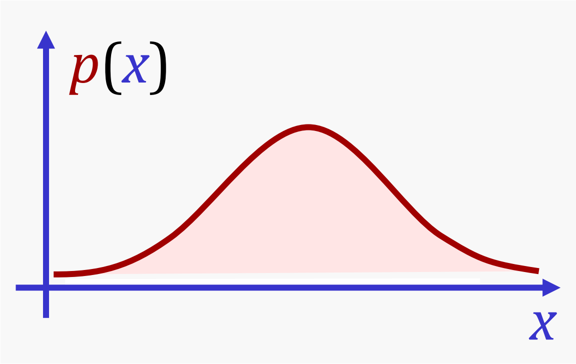
Part VI: Gaussian Distributions

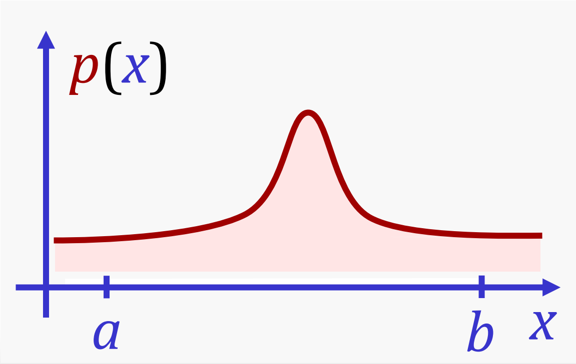
Well-known probability distributions

Important distributions

- Uniform distribution
  - Only defined for finite domains
  - Maximum entropy among all distributions
- Gaussian / normal distribution
  - Infinite domains
  - Maximizes entropy for fixed variance
- Heavy tail distributions
  - “Outlier robust”







Gaussians

Gaussian Normal Distribution

- Two parameters: μ, σ
- Density:

$$\mathcal{N}_{\mu,\sigma}(x) \coloneqq \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Mean: μ
- Variance: σ^2

Gaussian normal distribution

Log Space

Neg-log-density:

$$\log \mathcal{N}_{\mu,\sigma}(x) := \frac{(x-\mu)^2}{2\sigma^2} + \frac{1}{2}\ln(2\pi\sigma^2)$$
$$\sim \frac{1}{2\sigma^2}(x-\mu)^2$$

Calculations in log-space:

- Densities of products of Gaussians are Sums of quadratic polynomials
- Calculations simplified in log-space
  - Exception: Sum of Gaussians do not work

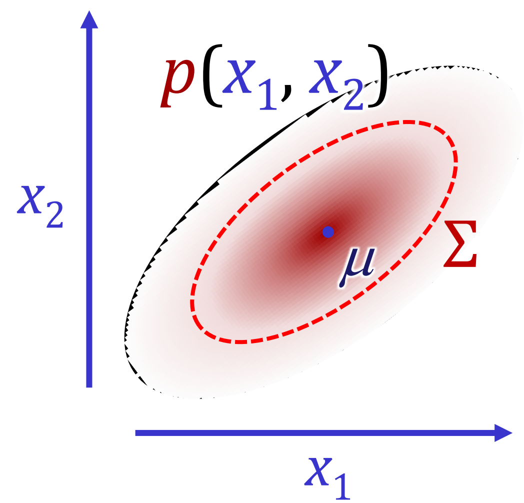
Multi-Variate Gaussians

Gaussian Normal Distribution in *d* Dimensions

- Two parameters: $\boldsymbol{\mu}$ (d-dim-vector), Σ ($d \times d$ matrix)
- Density:

$$\mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{x}) := \left(\frac{1}{(2\pi)^{-\frac{d}{2}}\det(\boldsymbol{\Sigma})^{-\frac{1}{2}}}\right) e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

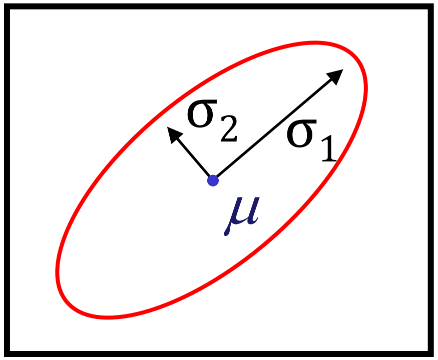
- Mean: $\boldsymbol{\mu}$
- Covariance Matrix: Σ



Log Space

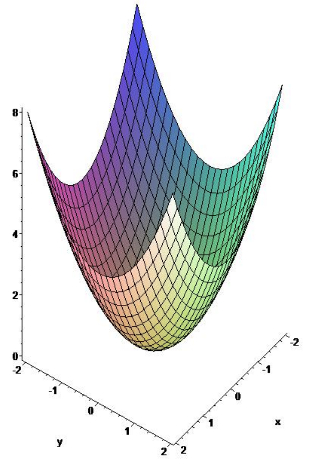
Neg-Log Density:

- $\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})+const$
- Quadratic multivariate polynomial



Consequences:

- Optimization (maximum probability density) by solving a linear system
- Gaussians are ellipsoids
  - Eigenvectors of Σ are main axes (*principal component analysis*, PCA)
  - Eigenvalues are extremal variances



More Rules for Gaussians

More Rules for Computations with Gaussians

- Products of Gaussians are Gaussians
  - Algorithm: Add quadratic polynomials
  - Variance can only decrease
- Marginals ("projections") of Gaussians are Gaussians
  - Unknown values: Leave out dimensions in $\boldsymbol{\mu}$, Σ
  - Known values: Schur complement
- Affine mappings of Gaussians are Gaussians
  - Algorithm: apply map to argument $\mathbf{x}$, yields different quadric
- General sums of Gaussians do not have closed-form log-densities

More Rules for Gaussians

Coordinate Transforms

- General Gaussians as affine transforms of unit Gaussians
  - Quadric $\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})+c$
  - Main axis transform:

$$\boldsymbol{\Sigma}^{-1}=\mathbf{U}\mathbf{D}\mathbf{U}^{\mathrm{T}}=\mathbf{U}\begin{pmatrix}\sigma_1^{-2} & & \\ & \sigma_2^{-2} & \\ & & \ddots\end{pmatrix}\mathbf{U}^{\mathrm{T}}$$

$$\boldsymbol{\Sigma}^{-\frac{1}{2}}=\mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^{\mathrm{T}}=\mathbf{U}\begin{pmatrix}\sigma_1^{-1} & & \\ & \sigma_1^{-1} & \\ & & \ddots\end{pmatrix}\mathbf{U}^{\mathrm{T}}$$

More Rules for Gaussians

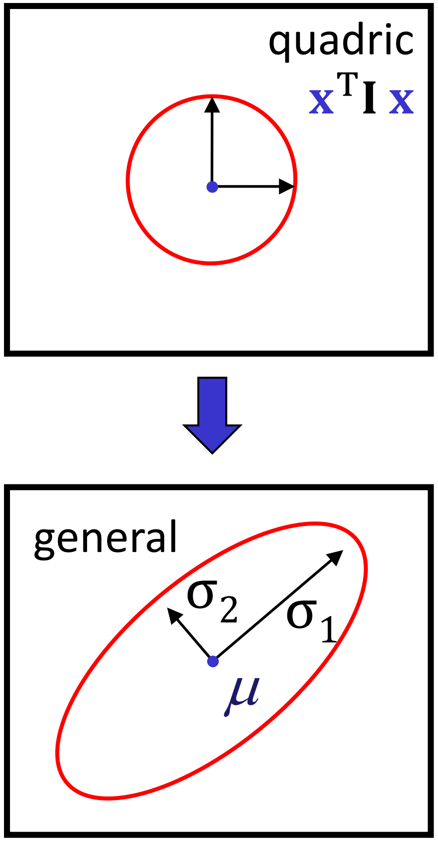
Unit Gaussian:

- We get:

$$\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\left(\Sigma^{-\frac{1}{2}}\right)^{\mathrm{T}}\left(\Sigma^{-\frac{1}{2}}\right)(\mathbf{x}-\boldsymbol{\mu})+c$$

$$=\frac{1}{2}\left(\left(\Sigma^{-\frac{1}{2}}\right)\mathbf{x}-\left(\Sigma^{-\frac{1}{2}}\right)\boldsymbol{\mu}\right)^{\mathrm{T}}\left(\left(\Sigma^{-\frac{1}{2}}\right)\mathbf{x}-\left(\Sigma^{-\frac{1}{2}}\right)\boldsymbol{\mu}\right)+c$$

- This is a unit Quadric / Gaussian $\mathbf{x}^{\mathrm{T}}\mathbf{I}\,\mathbf{x}$
  - rotated to Coordinate frame $\Sigma^{-\frac{1}{2}}$
  - and translated accordingly by $\left(\Sigma^{-\frac{1}{2}}\right)\boldsymbol{\mu}$



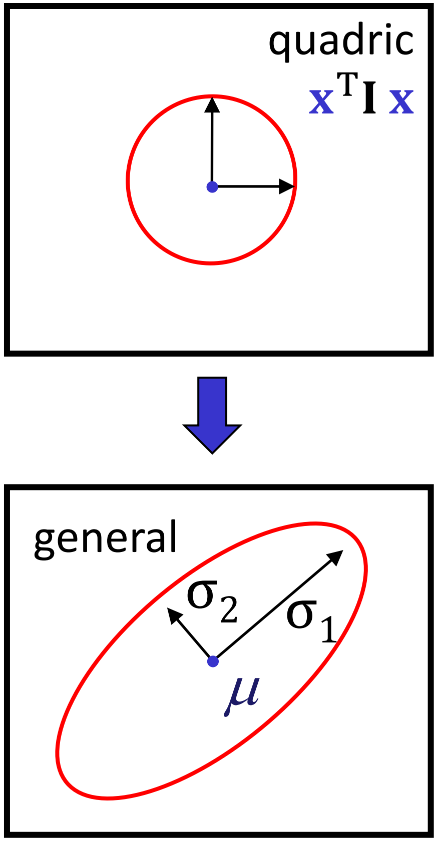
More Rules for Gaussians

Unit Gaussian:

- In addition, we have to recompute the (log) normalization factor

$$c = \ln\left(\frac{1}{(2\pi)^{-\frac{d}{2}}\det(\boldsymbol{\Sigma})^{-\frac{1}{2}}}\right)$$

to ensure a unit integral



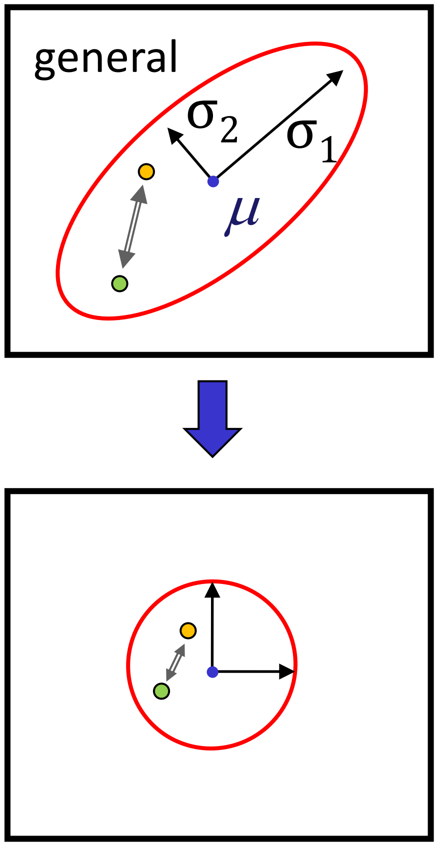
Rule of thumb:

- All Gaussians are related by
  - Translation
  - Rotation & non-uniform scaling
  - Adapting the density to integrate to 1

Mahalanobis Distance

Given:

- A Gaussian distribution with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
- Sample point $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$



Mahalanobis distance of $\mathbf{x}$:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

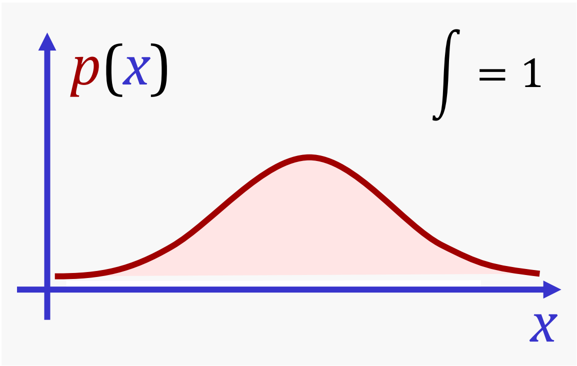
Interpretation:

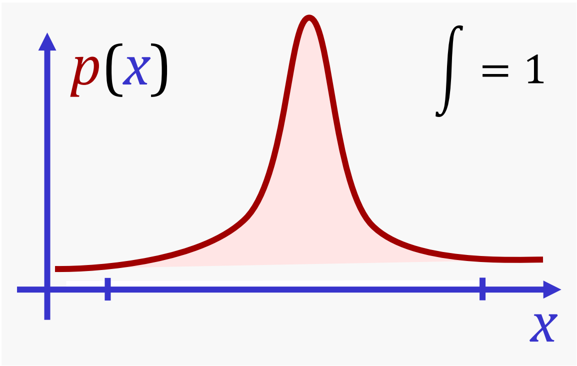
- Measures distances in "unit Gaussian space"
- One unit = one standard deviation

Applications

Example

- Given a sample from and a Gaussian distribution
- How likely is this sample from that distribution?
- Density value not a good measure
  - Absolute density depends on breadth





Estimation from Data

Task

- Data $\mathbf{d}_1, \ldots, \mathbf{d}_n$ generated w/Gaussian distribution (i.i.d.)
- Estimate parameters

Maximum Likelihood Estimation

- Most likely parameters: $\text{argmax}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} P(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{d}_1, \ldots, \mathbf{d}_n)$

$$\boldsymbol{\mu}_{ml} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{d}_i$$

mean

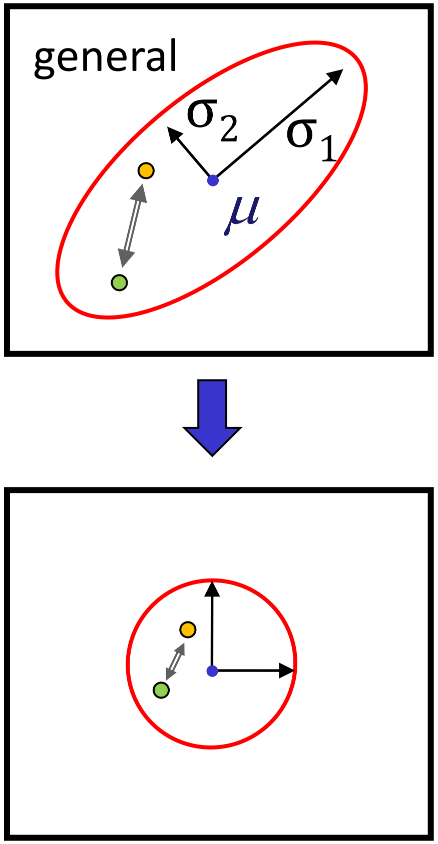
$$\boldsymbol{\Sigma}_{ml} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{d}_i - \boldsymbol{\mu})(\mathbf{d}_i - \boldsymbol{\mu})^{\mathrm{T}}$$

covariance

Mahalanobis Distance

Given:

- A Gaussian distribution with parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$
- Sample point $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$



Mahalanobis distance of $\mathbf{x}$:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

$$D_M(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})}$$

Interpretation:

- Measures distances in “unit Gaussian space”
- One unit = one standard deviation

Conclusions

Bayesian Statistics

- Uncertain captured in numbers
- Mathematics gives us the rules to derive consequences of our assumptions

The rest of the theory

- Formal tools to work with uncertainty