ETH zürich

D INFK

Probabilistic Artificial Intelligence

Introduction

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)
Institute for Machine Learning

What is AI?

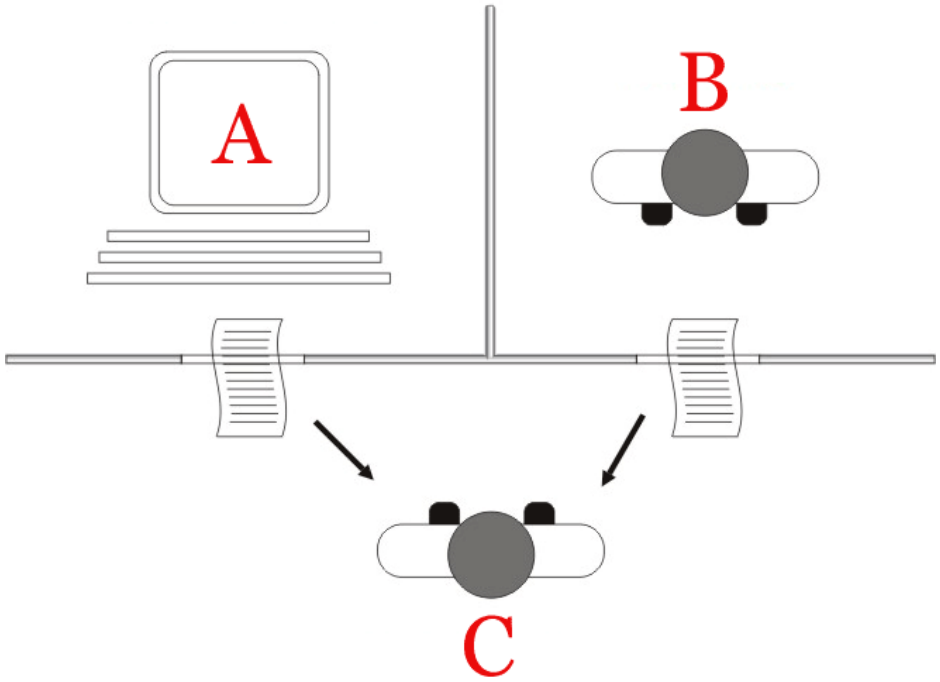
"The science and engineering of making intelligent machines"

(John McCarthy, '56)

What does intelligence mean?

https://en.wikipedia.org/wiki/John_McCarthy_(computer_scientist)#/media/File:John_McCarthy_Stanford.jpg

Acting Intelligently: The Turing Test



- Turing (‘50): Computing Machinery and Intelligence
- Predicted that by 2000, machine has 30% of fooling a lay person for 5 minutes

https://commons.wikimedia.org/wiki/File:Turing_test_diagram.png

AI Today

- ~~Build systems that act intelligently ("Strong AI")~~
- Build systems that solve tasks commonly associated with requiring human-level intelligence
- Amenable to engineering principles, empirical evaluation
- Involves / builds on / integrates
  - optimization, algorithms, control theory, logic, probability & statistics, game theory, machine learning …

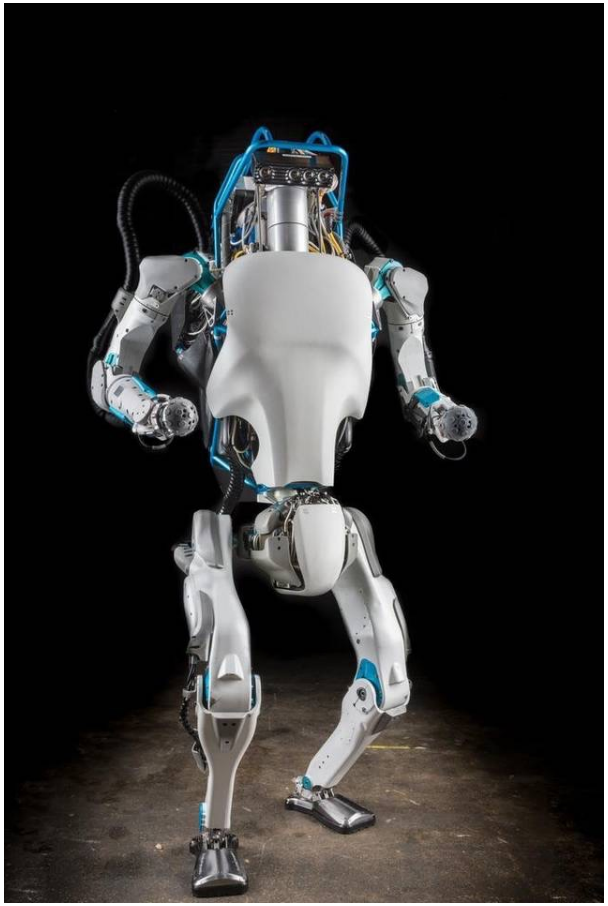
- "Strong / General AI" still inspiration for the field!

What if We Had Intelligent Machines?

- What will happen to our jobs?
- What about misuse of AI?
- How to ensure AI systems act ethically?
- Will machines surpass human intelligence?
- What will we do with superintelligent machines?
- What will they do with us?
- …

Humanoid Robotics

Honda ASIMO



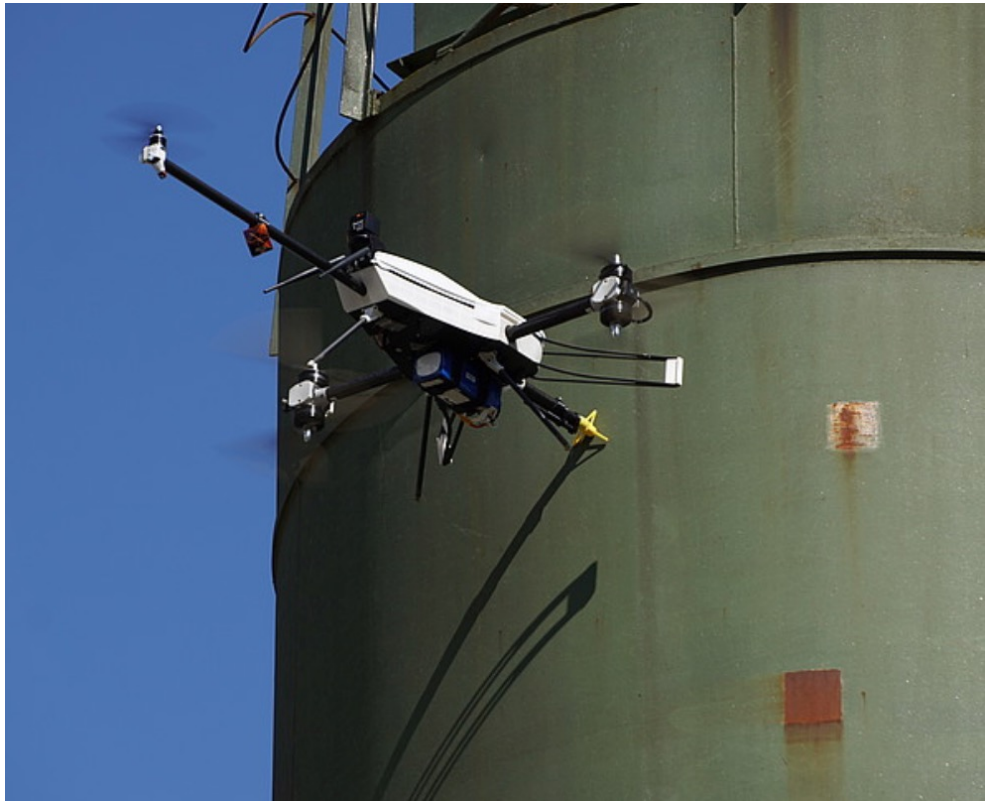
Boston Dynamics Atlas

http://www.kansascity.com/news/business/technology/917xpi/picture62197987/ALTERNATES/FREE_640/atlas%20from%20boston%20dynamics

Robotics

ANYmal

[rsl.ethz.ch]

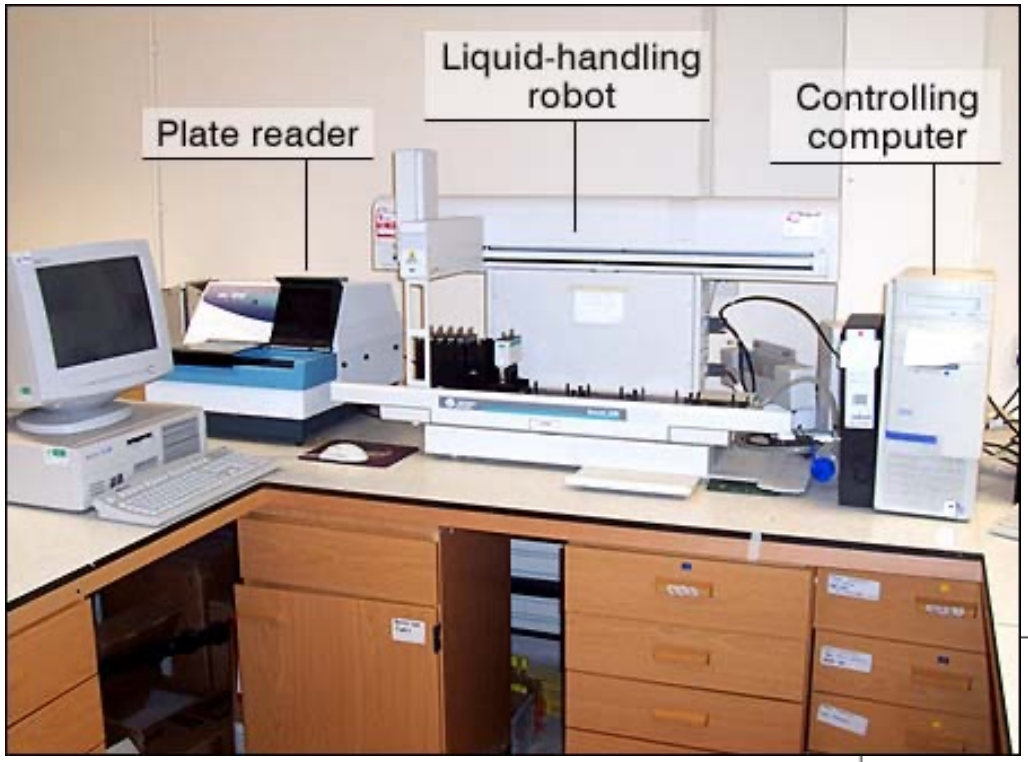


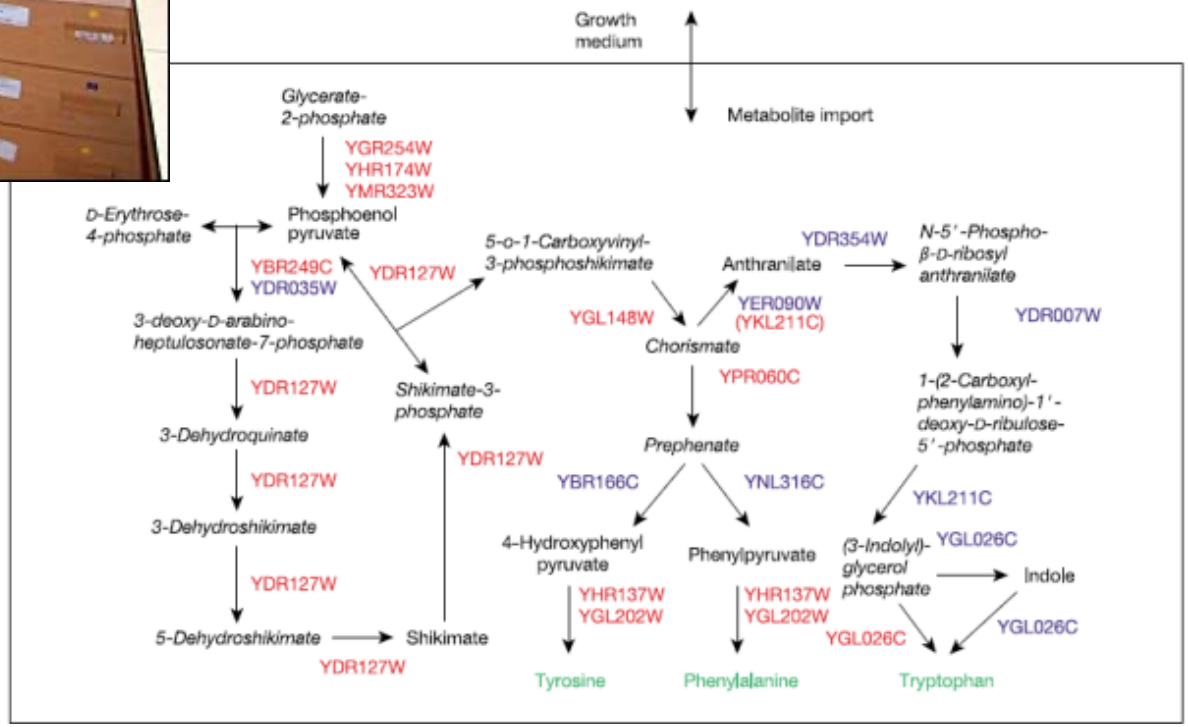
Voliro

[https://www.wysszurich.uzh.ch/projects/wyss-zurich-projects/voliro]

Robot Scientists

[King et al, Nature ’04, Science ‘09]





Autonomous Driving

CMU's Boss

Stanford's Stanley

- 1994, >1000km 3-lane highway w heavy traffic
- DARPA Grand Challenges:
  - 2005: drive 150 mile in the Mojave desert
  - 2007: drive 60 mile in traffic in urban environment
- Tesla / Waymo / Uber / … self-driving car projects

https://www.cmu.edu/news/image-archive/Boss.jpg

Games

IBM’s Deep Blue wins 6 game match against Garry Kasparov (’97)

https://media.wired.com/photos/5932a09d2a990b06268abee5/master/pass/deep-blue-kasparov.jpg

Games

- Go:
  - 2008: MoGo beats Pro (8P) in 9x9 game
  - 2016: AlphaGo wins 5th final game against Lee Sedol
  - 2017: Alpha Zero learns to play go (chess, shogi) from scratch
- Poker:
  - 2017: DeepStack and Libratus outperform pro Texas No limit Hold'em
  - 2019: Pluribus outperform pros at 6 player no-limed Hold'em

Computer games / eSports

[TheVerge 6.8.'18]

- 2018: DeepMind's AlphaStar defeats top pro player
- 2019: OpenAI Five Bot defeats pro human experts in Dota 2

NLP: “Answer Questioning”

February ’11 IBM Watson beat Brad Rutter, Ken Jennings in a $1M competition.

https://upload.wikimedia.org/wikipedia/en/9/9b/Watson_Jeopardy.jpg

Large language models & Generative AI

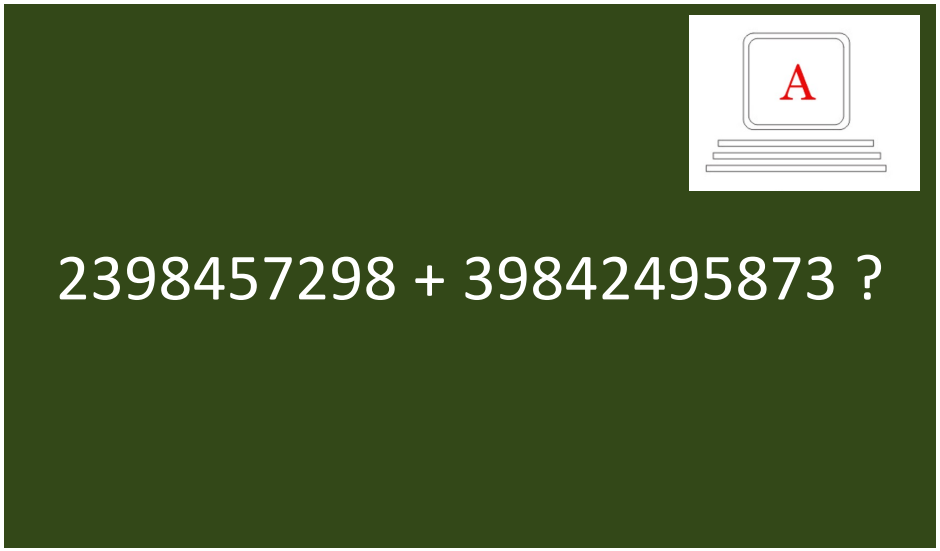
KR

Why should I take the course "Probabilistic Artificial Intelligence"?

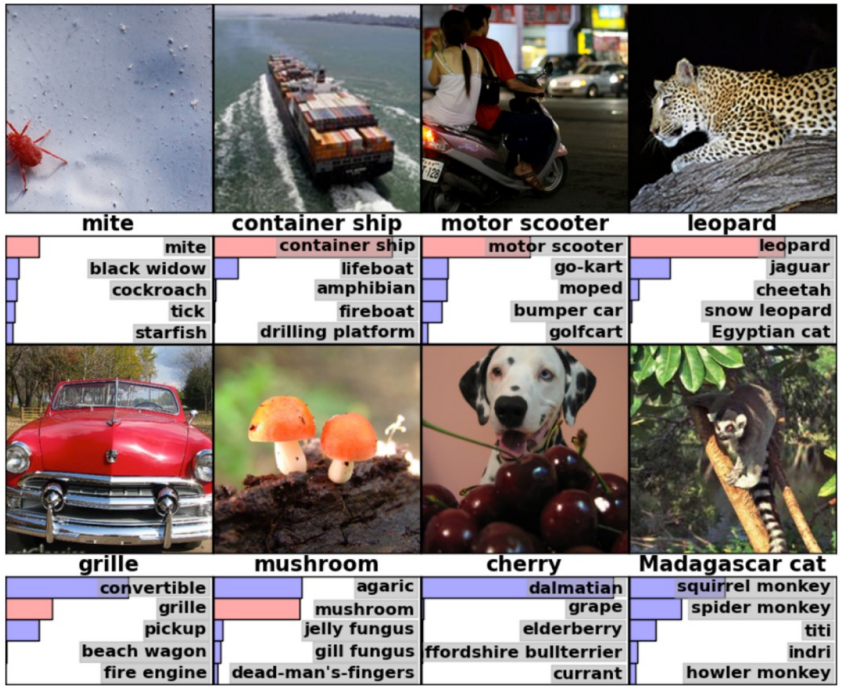
Taking a course in "Probabilistic Artificial Intelligence" can offer you several benefits, depending on your goals and interests. Here are some reasons why you might consider enrolling in such a course:

1. **Fundamental Understanding**: Probabilistic AI is a subfield of artificial intelligence that focuses on using probabilistic methods and models to solve complex problems. It provides a fundamental understanding of how uncertainty and randomness can be incorporated into AI systems. This knowledge is essential for building robust and reliable AI applications.
2. **Versatility**: Probabilistic AI techniques are applicable across various domains and industries. Whether you are interested in healthcare, finance, robotics, natural language processing, or any other field, probabilistic AI methods can be adapted to address a wide range of real-world problems.
3. **Decision-Making**: Probabilistic AI equips you with tools to make informed decisions in uncertain environments. You'll learn how to model and reason about uncertainty, which is crucial in situations where outcomes are not deterministic. This is particularly relevant in fields like autonomous vehicles, medical diagnosis, and risk assessment.
4. **Machine Learning**: Many probabilistic AI techniques are closely related to machine learning, especially in areas like Bayesian machine learning

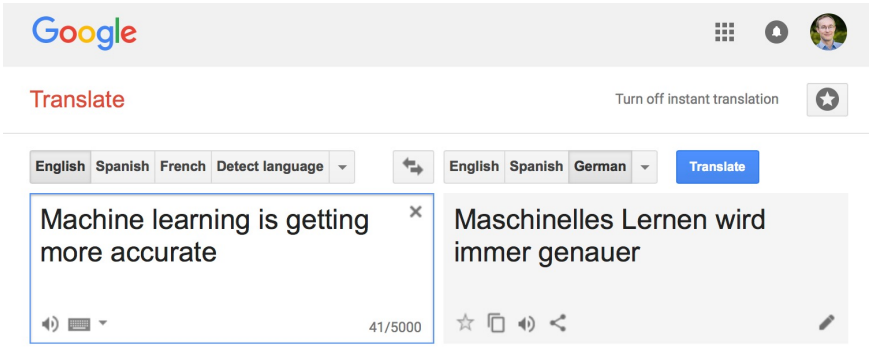
What is easy? What is hard?

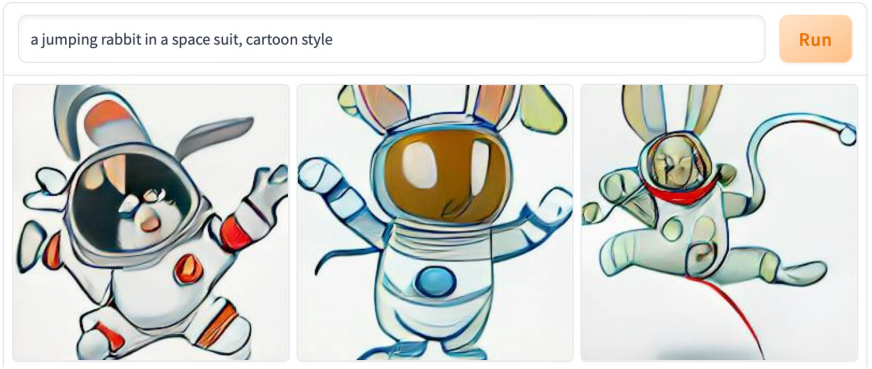


Breakthroughs in ML



Krizhevsky et al. ImageNet Classification with Deep Convolutional Neural Networks '12





https://huggingface.co/spaces/dalle-mini/dalle-mini

KR Why should I take the course "Probabilistic Artificial Intelligence"?

Taking a course in "Probabilistic Artificial Intelligence" can offer you several benefits, depending on your goals and interests. Here are some reasons why you might consider enrolling in such a course:

1. **Fundamental Understanding**: Probabilistic AI is a subfield of artificial intelligence that focuses on using probabilistic methods and models to solve complex problems. It provides a fundamental understanding of how uncertainty and randomness can be incorporated into AI systems. This knowledge is essential for building robust and reliable AI applications.
2. **Versatility**: Probabilistic AI techniques are applicable across various domains and industries. Whether you are interested in healthcare, finance, robotics, natural language processing, or any other field, probabilistic AI methods can be adapted to address a wide range of real-world problems.
3. **Decision-Making**: Probabilistic AI equips you with tools to make informed decisions in uncertain environments. You'll learn how to model and reason about uncertainty, which is crucial in situations where outcomes are not deterministic. This is particularly relevant in fields like autonomous vehicles, medical diagnosis, and risk assessment.
4. **Machine Learning**: Many probabilistic AI techniques are closely related to machine learning, especially in areas like Bayesian machine learning

Topics covered

- **Probabilistic foundations of AI**

  ➔ data-driven reasoning and
  decision making under uncertainty

- Bayesian learning (GPs, Bayesian deep learning, variational inference, MCMC)
- Bandits & Bayesian optimization
- Planning under uncertainty (MDPs, POMDPs)
- (Deep) Reinforcement learning
- Applications (in class and in project)

Other ML Courses @ ETHZ

- Advanced Machine Learning (Fall)
- Deep Learning (Fall)
- Reliable and Interpretable AI (Fall)
- Natural Language Processing (Fall)

- Foundations of Reinforcement Learning (Spring)
- Computational Intelligence Lab (Spring)
- Statistical Learning Theory (Spring)
- Machine Perception (Spring)
- Large Language Models (Spring)
- …

Overview

- *Instructor*:
  Andreas Krause (krausea@ethz.ch)

- *Teaching assistants*: (pai24-info@inf.ethz.ch)
  Scott Sussex – *Head TA*

  Arad Mohammadi, Arghavan Kassraie, Armin Lederer, Bhavya Sukhija, Cynthia Chen, Dennis Jueni, Erya Guo, Fan Sun, Frederike Luebeck, Jin Cheng, Lars Lorch, Leander Diaz-Bone, Liyuan Li, Manish Prajapat, Marco Bagatella, Mohammadreza Karimi, Mojmir Mutni, Morteza Sadat, Nicolas Dutly, Parnian Kassraie, Patrik Okanovic, Paul Streli, Riccardo De Santi, Tanyu Jiang, Viacheslav Borovitskiy, Ya-ping Hsieh, Yarden As, Yilmazcan Özyurt

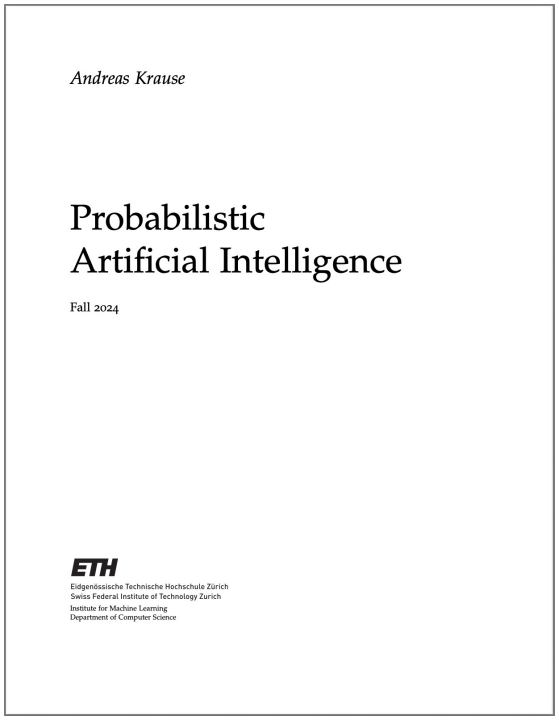
- *Administrative assistant*:
  Rita Klute (rita.klute@inf.ethz.ch)

Course material

- Slides & information available on course webpage: https://las.inf.ethz.ch/teaching/pai-f24

- Lecture notes https://las.inf.ethz.ch/courses/pai-f24/script/main.pdf

  - Feedback appreciated via issue/merge requests: https://gitlab.inf.ethz.ch/OU-KRAUSE/pai-script



Further Reading

- Relevant books:
  - S. Russell, P. Norvig: Artificial Intelligence, A Modern Approach (4th edition)
  - C. Rasmussen, C.K.I. Williams: Gaussian Processes in Machine Learning http://www.gaussianprocess.org/gpml/
  - K. Murphy: Machine Learning: A Probabilistic Perspective
  - R. Sutton, A. Barto: Reinforcement Learning http://incompleteideas.net/book/RLbook2020.pdf
  - C. Szepesvári: Algorithms for Reinforcement Learning https://sites.ualberta.ca/~szepesva/papers/RLAlgsInMDPs.pdf
  - T. Lattimore, C. Szepesvári: Bandit Algorithms https://tor-lattimore.com/downloads/book/book.pdf
- Articles referenced on webpage

Background & Prequisites

- **Required**:
  - Solid basic knowledge in probability, linear algebra, calculus, algorithms and programming.
  - Introduction to Machine Learning (or similar)
    Material online: https://las.inf.ethz.ch/teaching/introml-s24

- We review necessary background, but will move quickly…

Coursework

- Grade based on written **session exam**
  (need to pass course project to be admitted to exam)

- **~ Six homeworks** (not graded)
  - Primarily using the Moodle environment
- **Exercise sessions:** Thursdays 16-18 (HG F7 + recorded)
  - Reviewing relevant material
  - Discussion of homework solutions
  - **Will start next week**

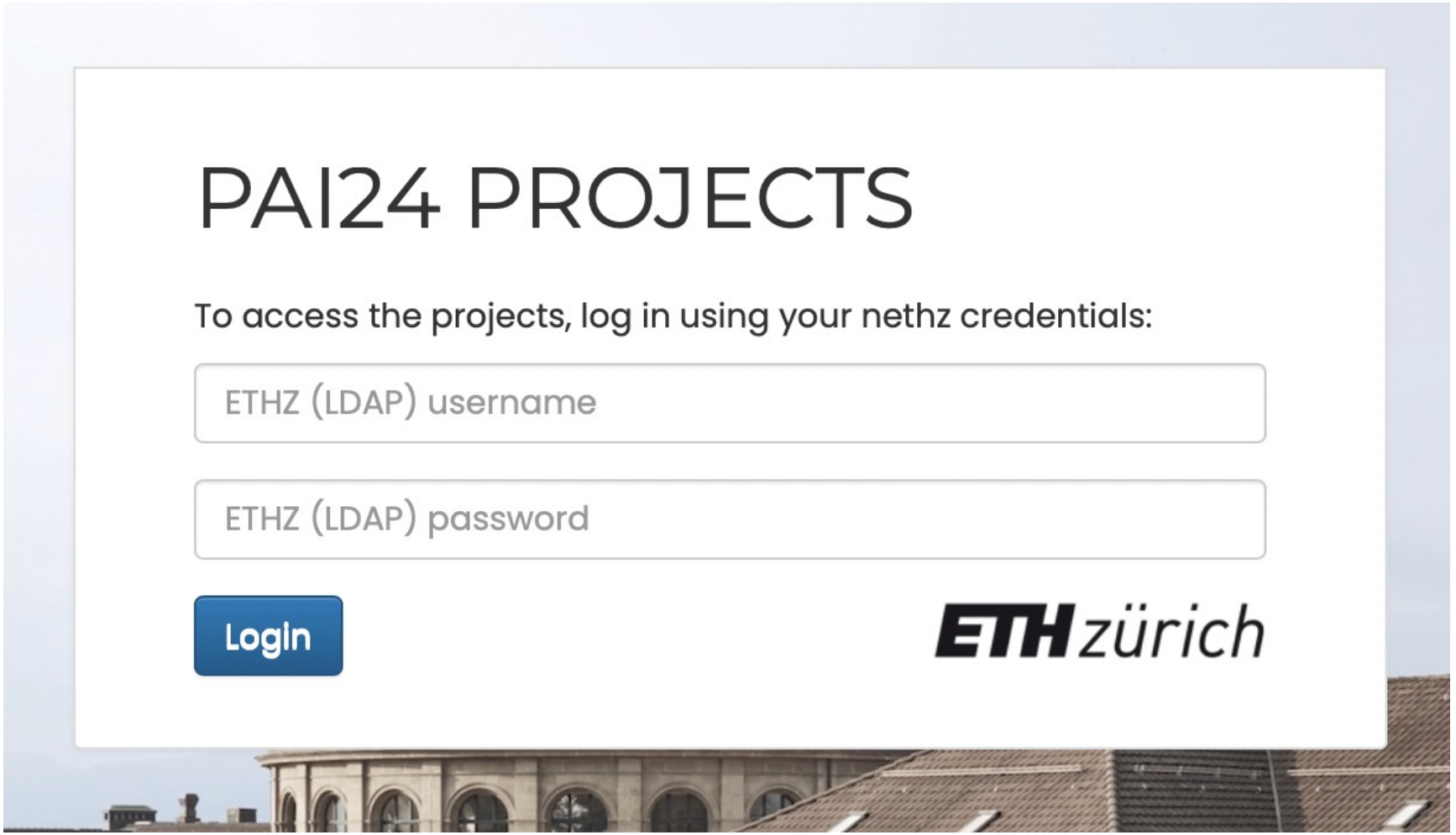
Lecture Format

- Focus on physical participation
- In principle possible to attend the class (lectures, exercises, projects) entirely virtually (except exam)
- Opportunities for asking questions:
  - During class: in person, via EduApp chat (https://eduapp-app1.ethz.ch)
  - During virtual Q&A session on Mondays 17:15-18:00 (Zoom link on course webpage)
  - Asynchronously: Moodle forum

Course Project

- Hands on experience on Probabilistic machine learning, Bayesian optimization & Reinforcement learning
- Teams of (up to) 3 students
- Four smaller projects (+1 ungraded warmup)
- Details and grading explained here:
- https://las.inf.ethz.ch/courses/pai-f24/project/pai24projectinfo.pdf
- 3 out of 4 projects must be passed for exam admission
- Best project award for each task
- Possible solutions discussed during Q&A session

https://project.las.ethz.ch



ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

inf | Informatik
Computer Science

Probabilistic Artificial Intelligence

Brief Recap of Probability

Prof. Andreas Krause
Learning and Adaptive Systems (las.ethz.ch)
Institute for Machine Learning

Review: Probability

- Formally: Probability Space $(\Omega, \mathcal{F}, P)$
  - Set of atomic events: Ω
  - Set of all non-atomic events: $\mathcal{F} \subseteq 2^{\Omega}$

    $\mathcal{F}$ is a σ-Algebra (closed under complements and countable unions)

$$\Omega \in \mathcal{F}$$

$$A \in \mathcal{F} \Rightarrow \Omega \setminus A \in \mathcal{F}$$

$$A_1, \ldots, A_n, \cdots \in \mathcal{F} \Rightarrow \bigcup_i A_i \in \mathcal{F}$$

  - Probability measure $P\colon \mathcal{F} \to [0, 1]$

    For $A \in \mathcal{F}$, $P(A)$ is the probability that event A happens

Probability Axioms

Normalization: $P(\Omega) = 1$

Non-negativity: $P(A) \geq 0 \text{ for all } A \in \mathcal{F}$

σ-additivity:

$$\forall A_1, \ldots A_n, \ldots \in \mathcal{F} \text{ disjoint}: P\left(\bigcup_{I=1}^{\infty} A_i\right) = \sum_{I=1}^{\infty} P(A_i)$$

Interpretation of probabilities

- Philosophical debate…
- Frequentist interpretation
  - $P(A)$ is relative frequency of A in repeated experiments
  - Can be difficult to assess with limited data
- Bayesian interpretation
  - $P(A)$ is "degree of belief" that A will occur
  - Where does this belief come from?
  - Many different flavors (subjective, objective, pragmatic, …)
- For now, assume probabilities are known

Random Variables

- Events are cumbersome to work with
- Let D be some set (e.g., the integers)
- A random variable X is a mapping $X : \Omega \rightarrow D$
- For some $x \in D$, we say

$$P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

"probability that variable X assumes state x"

Specifying Probability Distributions through RVs

- Bernoulli distribution: "(biased) coin flips"

  $$D = \{H, T\}$$

  Specify $P(X = H) = p$. Then $P(X = T) = 1 - p$.

  Note: can identify atomic ev. ω with $\{X = H\}, \{X = T\}$

- Binomial distribution counts no. heads S in n flips

- Categorical distribution: "(biased) m-sided dice"

  $$D = \{1, \ldots, m\}$$

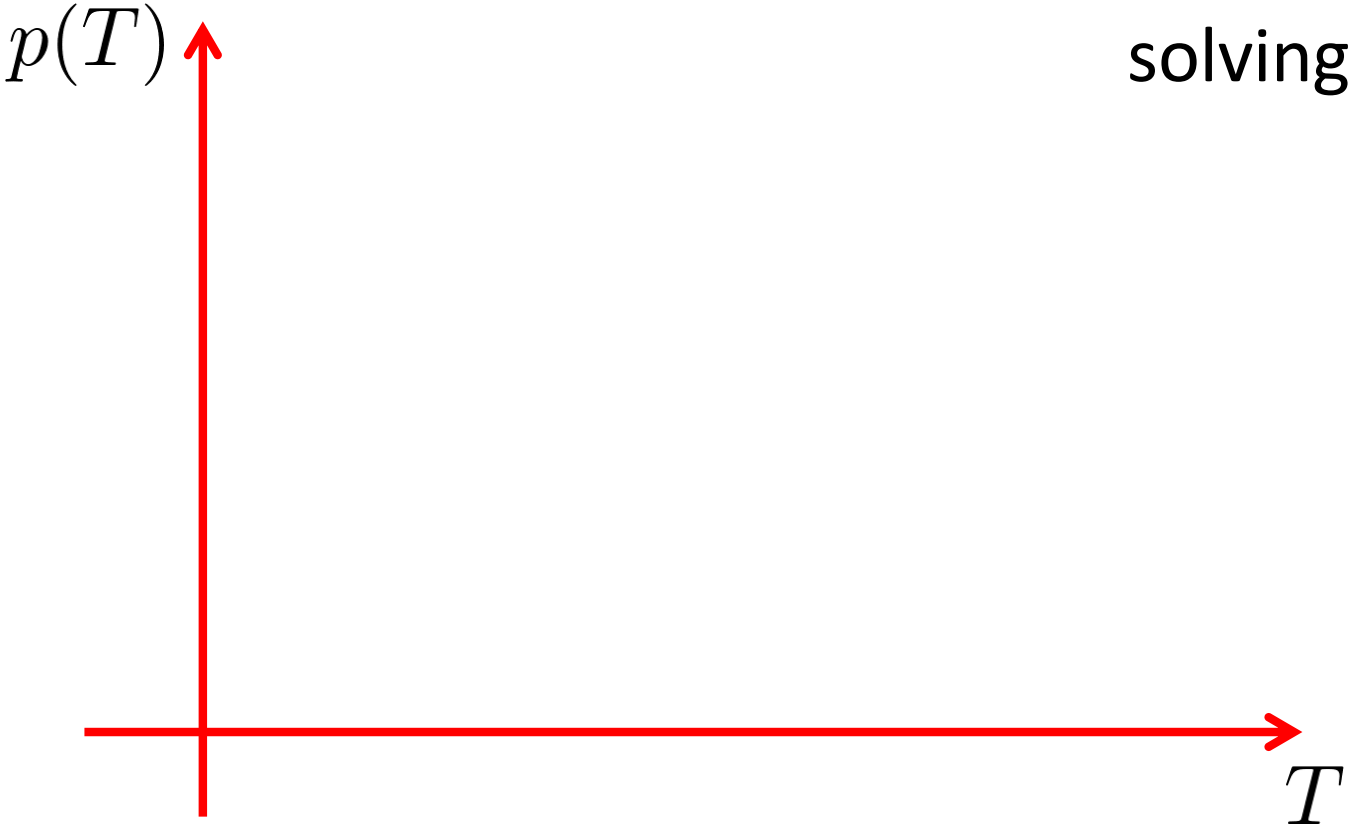
  Specify $P(X = i) = p_i$, s.t. $p_i \geq 0, \sum_i p_i = 1$

- Multinomial distribution counts the number of outcomes for each side for n throws

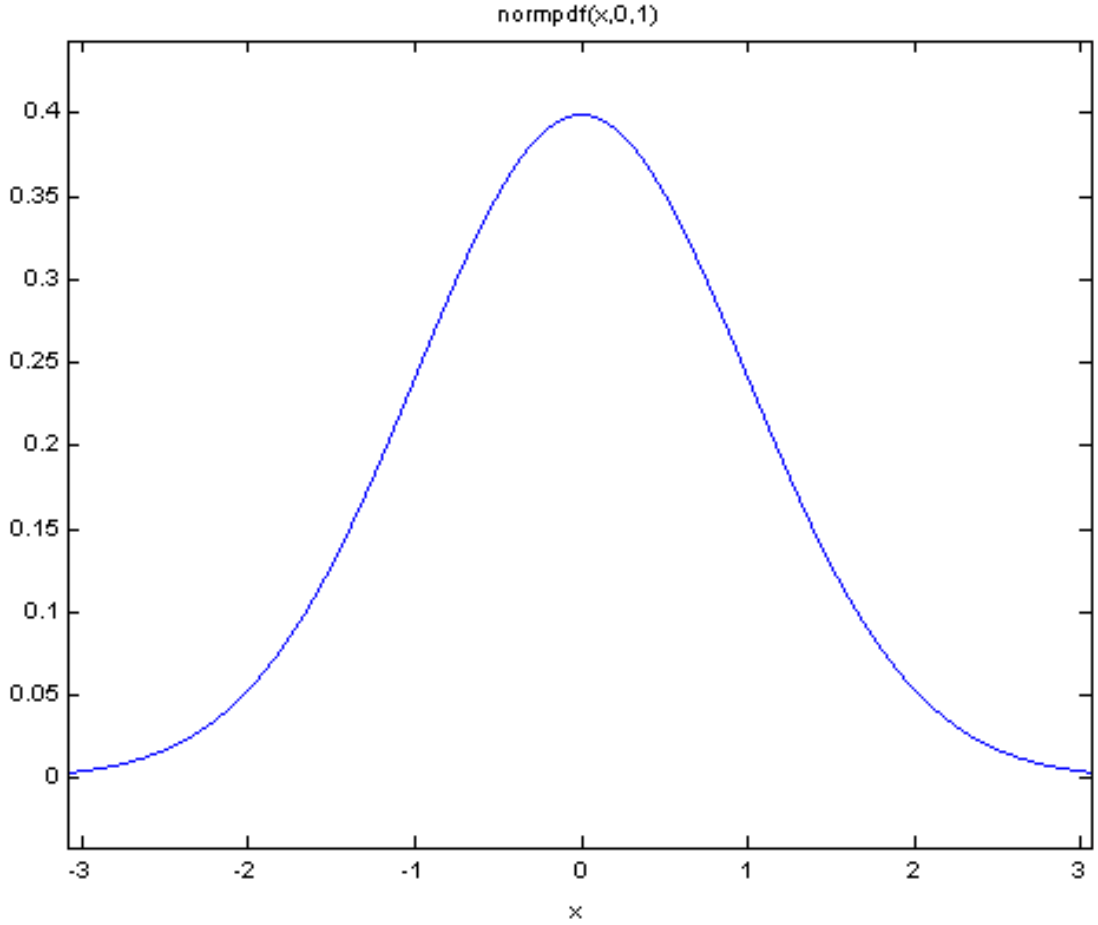
Continuous Distributions

- Probability *density*

T = time taken for solving the homework



Example: Gaussian Distribution



- σ = Std. dev.
- μ = mean

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

Joint Distributions

- Instead of random variable, have random vector

$$\mathbf{X} = [X_1(\omega), \ldots, X_n(\omega)]$$

- Can specify $P(X_1 = x_1, \ldots, X_n = x_n)$ directly (atomic events are assignments $x_1, \ldots, x_n$)
- Joint distribution describes relationship among all variables
- Example:

	toothache		no toothache	
	catch	no catch	catch	no catch
cavity	.108	.012	.072	.008
no cavity	.016	.064	.144	.576

Conditional Probability

- Formal definition:

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \text{ if } P(b) \neq 0$$

- Product rule $P(a \wedge b) = P(a \mid b)P(b)$
- For random vars: $P(A, B) = P(A \mid B)P(B)$
  (set of equations, one for each realization of A, B)

- Product rule for multiple RVs:

The Two Rules for Joint Distributions

Sum rule (Marginalization)

$$P(X_{1:i-1}, X_{i+1:n}) = \sum_{x_i} P(X_{1:i-1}, x_i, X_{i+1:n})$$

Product rule (chain rule)

$$P(X_1, \ldots, X_n) = P(X_1)P(X_2 \mid X_1) \ldots P(X_n \mid X_1, \ldots, X_{n-1})$$

Posterior Inference

- Suppose we know:
  - Prior probability $P(C)$

cavity	no cavity
.1	.9

  - Likelihood $P(T \mid C)$

	toothache	no toothache
cavity	.9	.1
no cavity	.01	.99

- How do we get $P(\text{cavity} \mid \text{toothache})$?

Bayes' Rule

Given:

- Prior P(X)
- Likelihood P(Y | X)

Bayes' rule computes posterior

$$P(X \mid Y) = \frac{P(X)P(Y \mid X)}{\sum_{X=x} P(X = x)P(Y \mid X = x)}$$

Independent RVs

- Random variables $X_1 \ldots X_n$ are called independent if

$$P(X_1 = x_1, \ldots, X_n = x_n) = P(x_1)P(x_2)\cdots P(x_n)$$

- Independence is a very strong requirement.
  Is there something weaker?

Conditional Independence (example)

- If I know there's a *cavity*, knowing *toothache* won't help predict whether the probe *catches*

$$P(\text{catch} \mid \text{cavity}, \text{toothache}) = P(\text{catch} \mid \text{cavity})$$

- for all values of *catch, cavity* and *toothache*

Key Concept: Conditional Independence

- Rand. vars. X and Y conditionally independent given Z iff for all x, y, z :

$$P(X = x, Y = y \mid Z = z) = P(X = x \mid Z = z)P(Y = y \mid Z = z)$$

- If $P(Y = y \mid Z = z) > 0$, that is equivalent to

$$P(X = x \mid Z = z, Y = y) = P(X = x \mid Z = z)$$

Similarly for sets of random variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$

We write:

$$\mathbf{X} \perp \mathbf{Y} \mid \mathbf{Z}$$

Problems with High-dim. Distributions

- Suppose we have n binary variables
- How many parameters do we need to specify $P(X_1 = x_1, \ldots, X_n = x_n)$?

X_1	X_2	…	X_{n-1}	X_n	P(X)
0	0	…	0	0	.01
0	0	…	1	0	.001
0	0	…	1	1	.213
…	…	…	…	…	
1	1	…	1	1	.0003

2^n-1 parameters! ☹

More Problems: Computing Marginals

- Suppose we have joint distribution $P(X_1, \ldots, X_n)$
- Then (acc. to sum rule)

$$P(X_i = x_i) = \sum_{x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n} P(x_1, \ldots, x_n)$$

- If all X_i are binary: How many terms does this sum have?

More Problems: Conditional Queries

- Suppose we have joint distribution $P(X_1, \ldots, X_n)$
- Compute distribution of some variables given values for others:

$$P(X_1 = \cdot \mid X_7 = x_7) = \frac{P(X_1 = \cdot, X_7 = x_7)}{P(X_7 = x_7)} = \frac{1}{Z} P(X_1 = \cdot, X_7 = x_7)$$

Challenges with high-dimensional distributions

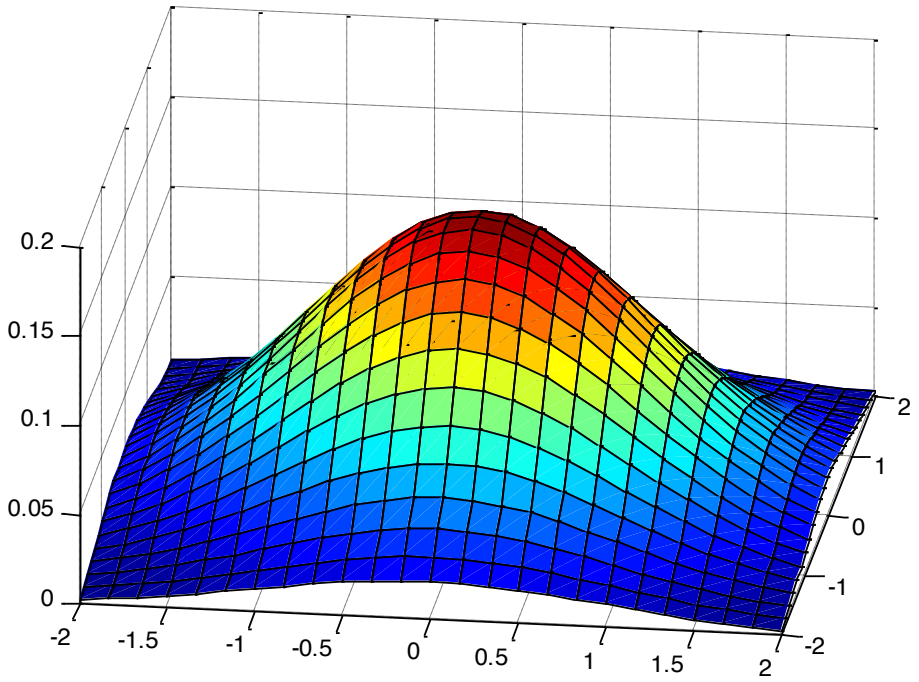
- Representation (parametrization)
- Learning (estimation)
- Inference (prediction)

- In the following, we will start with a particular family of distributions where these problems are tractable
  - Multivariate Gaussians

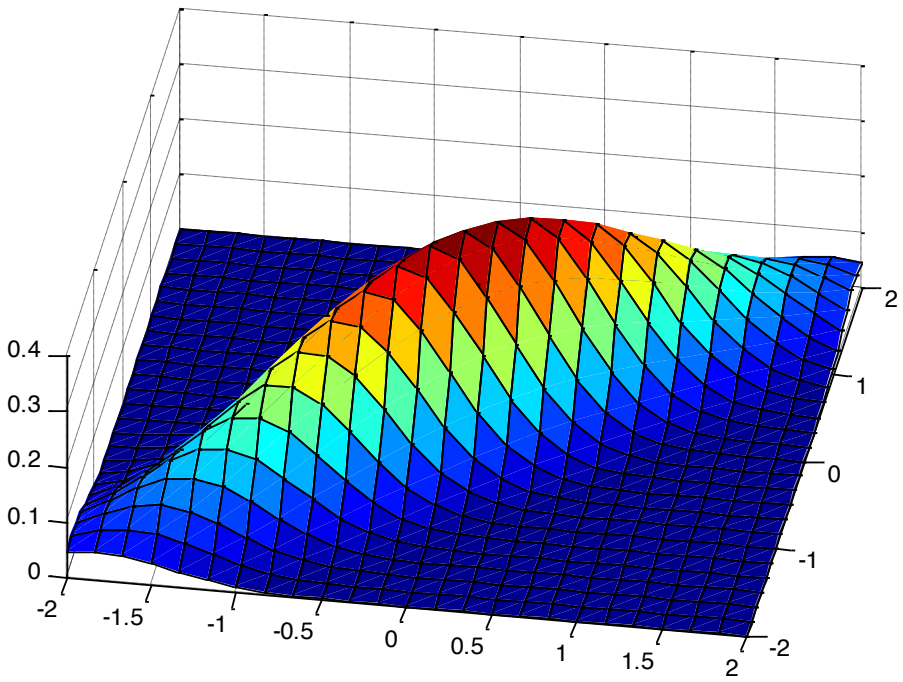
- We will discuss how to do Bayesian learning and inference with Gaussians (and Gaussian processes)
- Then generalize to more complex models, approximate inference etc. (Bayesian neural nets etc.)

Example: Multivariate Gaussian

$$p(x) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) \qquad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

Multivariate Gaussian distribution

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \vdots & & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix}$$

Thus joint distribution over n variables requires only $O(n^2)$ parameters!

Fact: Gaussians are independent iff they are uncorrelated

Bayesian inference in Gaussian Distributions

- Suppose we have a Gaussian random vector

$$\mathbf{X} = \mathbf{X}_V = [X_1, \ldots, X_d] \sim \mathcal{N}(\mu_V, \Sigma_{VV})$$

- Hereby $V = \{1,\ldots,d\}$ is an index set.
- Suppose we consider a subset of the variables

$$A = \{i_1, \ldots, i_k\}$$

- The marginal distribution of variables indexed by *A* is:

$$\mathbf{X}_A = [X_{i_1}, \ldots, X_{i_k}] \sim \mathcal{N}(\mu_A, \Sigma_{AA})$$

Conditional distributions

- Suppose we have a Gaussian random vector

$$\mathbf{X} = \mathbf{X}_V = [X_1, \ldots, X_d] \sim \mathcal{N}(\mu_V, \Sigma_{VV})$$

- Further, suppose we take two disjoint subsets of V

$$A = \{i_1, \ldots, i_k\} \qquad B = \{j_1, \ldots, j_m\}$$

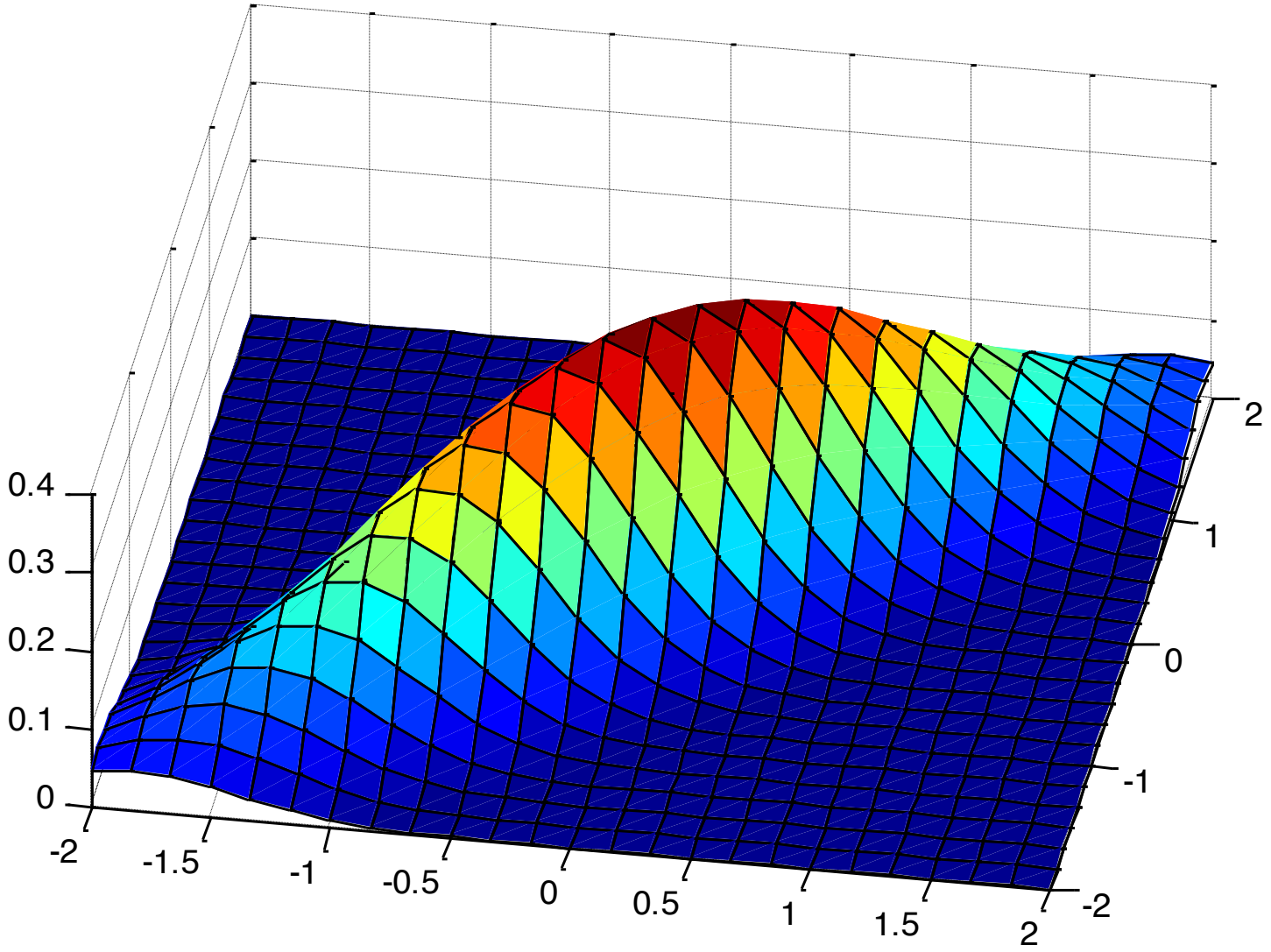
- The conditional distribution

$$p(\mathbf{X}_A \mid \mathbf{X}_B = \mathbf{x}_B) = \mathcal{N}(\mu_{A|B}, \Sigma_{A|B})$$

is Gaussian, where

$$\mu_{A|B} = \mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(\mathbf{x}_B - \mu_B)$$
$$\Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}$$

Illustration



Multiples of Gaussians are Gaussian

- Suppose we have a Gaussian random vector

$$\mathbf{X} = \mathbf{X}_V = [X_1, \ldots, X_d] \sim \mathcal{N}(\mu_V, \Sigma_{VV})$$

- Take a matrix $M \in \mathbb{R}^{m \times d}$
- Then the random vector $\mathbf{Y} = \mathbf{M}\mathbf{X}$ is Gaussian:

$$\mathbf{Y} \sim \mathcal{N}(\mathbf{M}\mu_V, \mathbf{M}\Sigma_{VV}\mathbf{M}^T)$$

Sums of Gaussians are Gaussian

- Suppose we have two independent Gaussian random vectors

$$\mathbf{X} = \mathbf{X}_V = [X_1, \ldots, X_d] \sim \mathcal{N}(\mu_V, \Sigma_{VV})$$
$$\mathbf{X}' = \mathbf{X}'_V = [X'_1, \ldots, X'_d] \sim \mathcal{N}(\mu'_V, \Sigma'_{VV})$$

- Then the random vector $\mathbf{Y} = \mathbf{X} + \mathbf{X}'$ is Gaussian:

$$\mathbf{Y} \sim \mathcal{N}(\mu_V + \mu'_V, \Sigma_{VV} + \Sigma'_{VV})$$

Conditional Linear Gaussians

- If X, Y are jointly Gaussian, then $p(X = x \mid Y = y)$ is Gaussian, with mean linearly dependent on y

- Thus random variable X can be viewed as a linear function of Y with independent Gaussian noise added

- The converse also holds.

Conditional Linear Gaussians

- If X, Y are jointly Gaussian, then $p(X = x \mid Y = y)$ is Gaussian, with mean linearly dependent on *y*

$$p(X = x \mid Y = y) = N\left(x; \mu_{X|Y=y}, \sigma^2_{X|Y}\right)$$
$$\mu_{X|Y=y} = \mu_X + \sigma_{XY}\sigma_Y^{-2}(y - \mu_Y)$$
$$\sigma_{X|Y} = \sigma_X^2 - \sigma_{XY}^2\sigma_Y^{-2}$$

- Thus random variable *X* can be viewed as a linear function of *Y* with independent Gaussian noise added

$$X = aY + b + \epsilon, \text{ where } \epsilon \sim N(0, \sigma^2_{X|Y})$$
$$\text{and } a = \sigma_{XY}\sigma_Y^{-2}, b = \mu_X - \sigma_{XY}^2\sigma_Y^{-2}\mu_y$$

- The converse also holds.

Outlook

- Multivariate Gaussians have important properties:
  - Compact representation of high-dimensional joint distributions
  - Closed form inference

- In the following, will discuss how we can make use of these properties for Bayesian learning