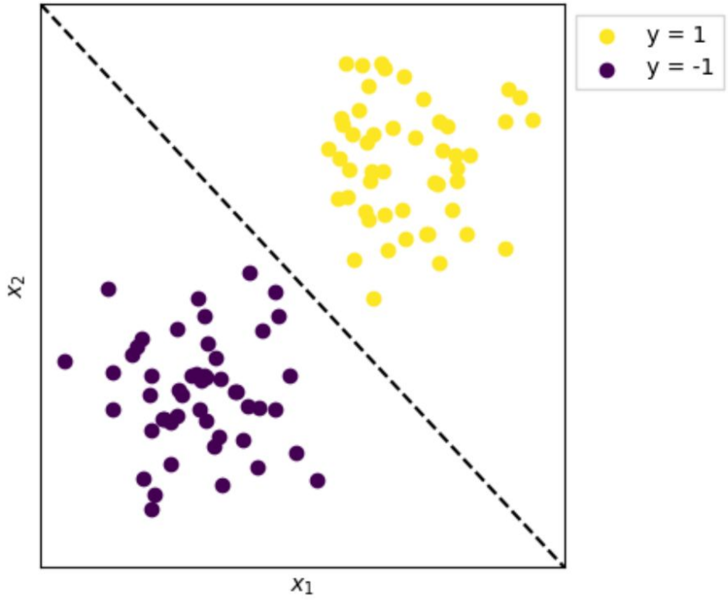
# CS 480/680
# Introduction to Machine Learning

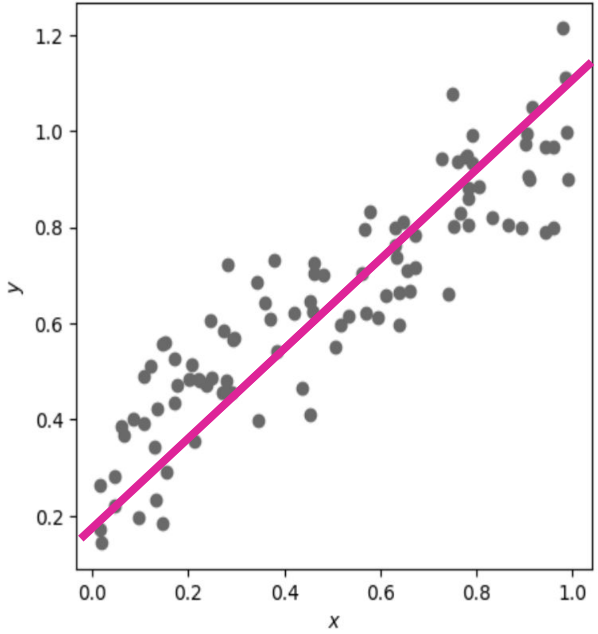
## Lecture 3
## Maximum Likelihood Estimation and Entropy

Kathryn Simone
17 September 2024

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

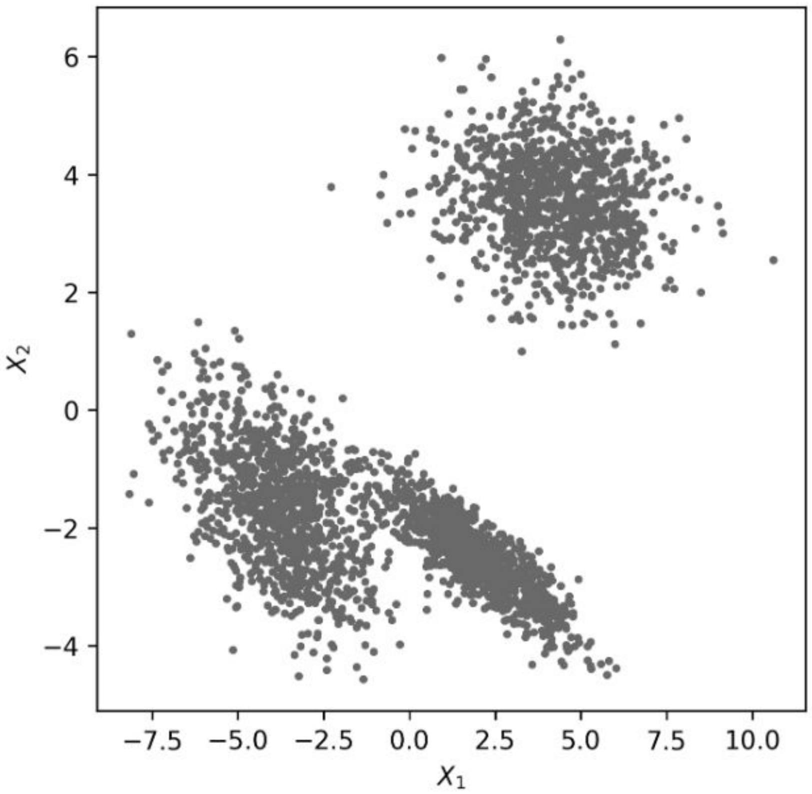
# Classification and regression are supervised learning tasks




UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

# Unsupervised learning concerns pattern identification



| Lecture | Date | Topics |
|---|---|---|
| 0 | 05/09/2024 | Introduction + Administrative Remarks |
| 1 | 10/09/2024 | Halfspaces the Perceptron Algorithm |
| 2 | 12/09/2024 | Linear Regression and Convexity |
| 3 | 17/09/2024 | Maximum Likelihood Estimation |
| 4 | 19/09/2024 | k-means Clustering |
| 5 | 24/09/2024 | k-NN Classification and Logistic Regression |
| 6 | 26/09/2024 | Hard-margin SVM |
| 7 | 01/10/2024 | Soft-margin SVM |
| 8 | 03/10/2024 | Kernel methods |
| 9 | 08/10/2024 | Decision Trees |
| 10 | 10/10/2024 | Bagging and Boosting |
| | 15/10/2024 | NO LECTURE - MIDTERM BREAK |
| | 17/10/2024 | NO LECTURE- MIDTERM BREAK |
| 11 | 22/10/2024 | Expectation Maximization Algorithm |
| 12 | 24/10/2024 | MLPs and Fully-Connected NNs |
| | 29/10/2024 | NO LECTURE - MIDTERM EXAM |
| 13 | 31/10/2024 | Convolutional Neural Networks |
| 14 | 05/11/2024 | Recurrent Neural Networks |
| 15 | 07/11/2024 | Attention and Transformers |
| 16 | 12/11/2024 | Graph Neural Networks (Time permitting) |
| 17 | 14/11/2024 | VAEs and GANs |
| 18 | 19/11/2024 | Flows |
| 19 | 21/11/2024 | Contrastive Learning (Time permitting) |
| 20 | 26/11/2024 | Robustness |
| 21 | 28/11/2024 | Privacy (Saber Malekmohammadi) |
| 22 | 03/12/2024 | Fairness |

UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

# Key Questions

I. How can we represent and sample from a distribution?

II. How do you estimate the parameters and evaluate the model?

III. Could this also apply to supervised learning?

IV. Summary + Housekeeping

# Lecture Objectives

At the end of the lecture, we should be able to:

★ Identify the probability density function and parameterization of widely used distributions and relate it to their use in constructing likelihood functions.
★ Construct the likelihood function for a dataset and maximize it to find the maximum likelihood estimates (MLE) of the parameters.
★ Define and apply information theoretic measures such as entropy and KL divergence to characterize and compare distributions.
★ Reformulate the linear regression objective using likelihood principles, and demonstrate that it can be viewed as a special case of MLE.

# Key Questions

I. **How can we represent and sample from a distribution?**

II. How do you estimate the parameters and evaluate the model?

III. Could this also apply to supervised learning?

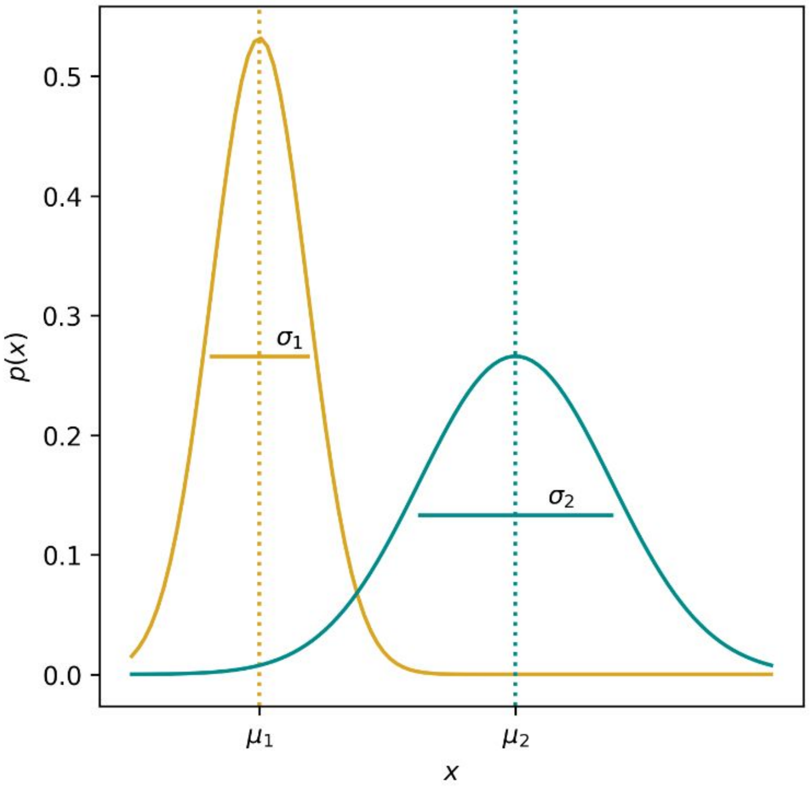
IV. Summary + Housekeeping

# The PDF of a univariate Gaussian (normal) distribution

The probability density at a point $x$ under a Gaussian distribution is given by:

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

Where $\mu$ and $\sigma^2$ parameters referring to the mean (or center) and variance, respectively.



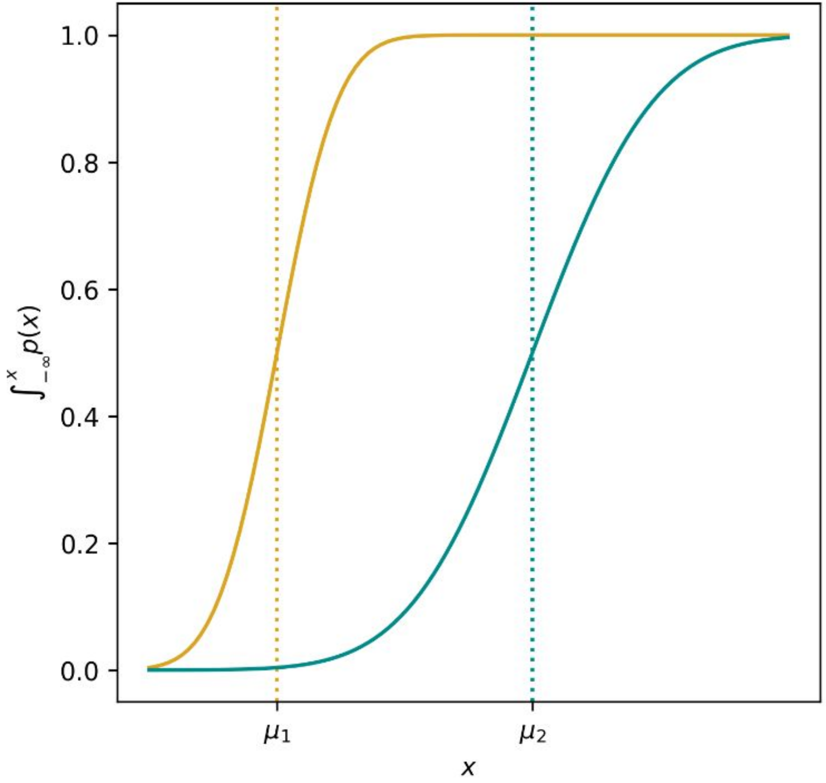
# The CDF of a univariate Gaussian (normal) distribution

Probability density functions must satisfy

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

Cumulative distribution function (CDF):

$$\Pr[X \leq x] = \int_{-\infty}^{x} p(x)$$



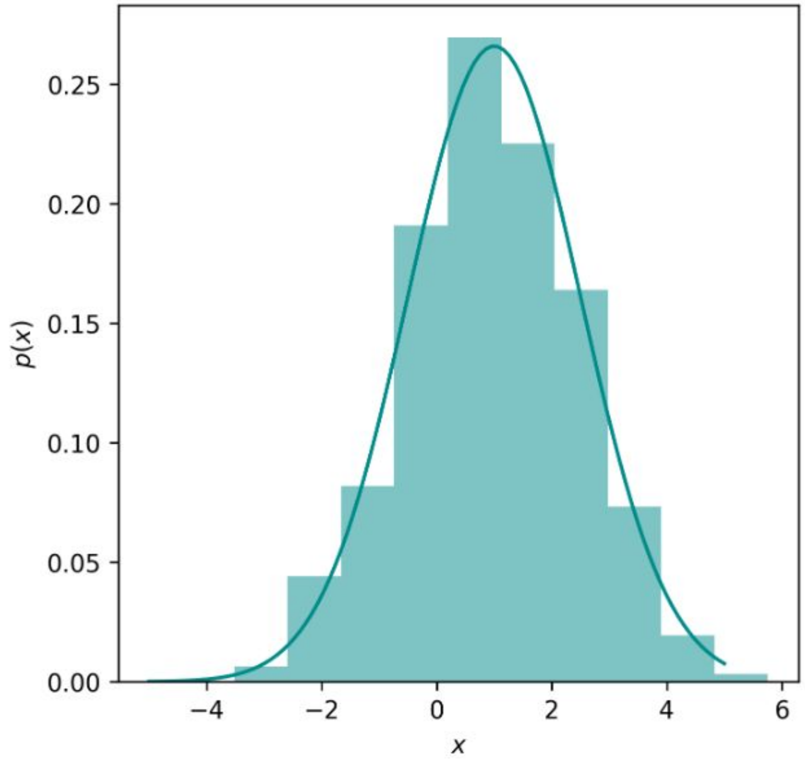
# Expectation and the first moment

We denote a continuous random variable $X$ that follows a normal distribution with mean $\mu$ and variance $\sigma^2$ as

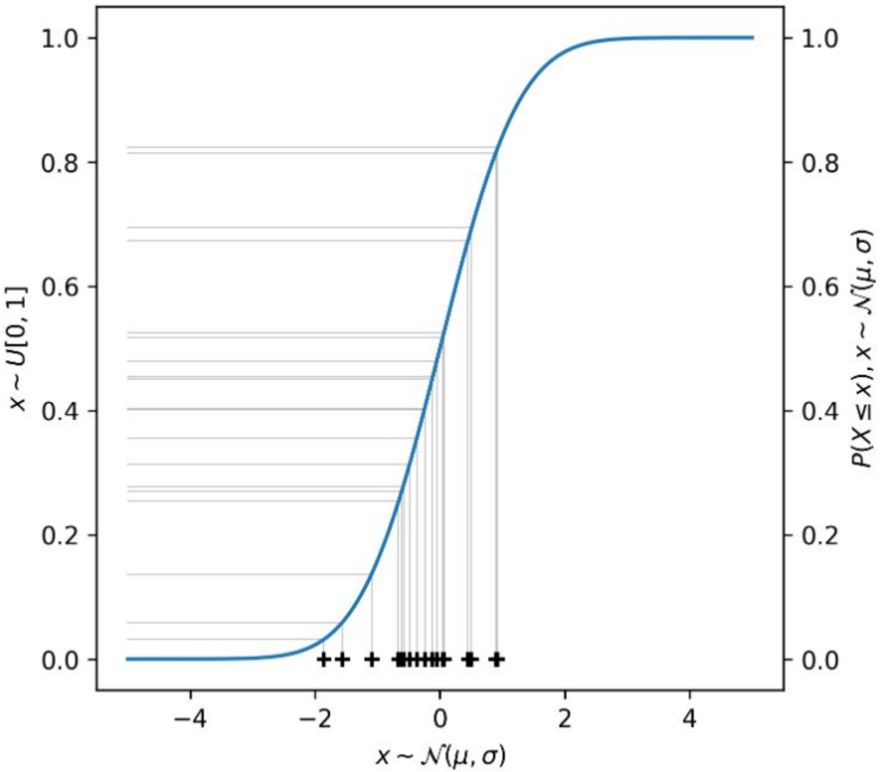
$$X \sim \mathcal{N}(\mu, \sigma^2).$$

The expectation of $X$ is given by:

$$E[X] = \int xp(x)dx = \mu$$
$$\approx \frac{1}{n}\sum_{i=1}^{n} X_i$$



# Inverse transform sampling from a parameterized distribution




UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

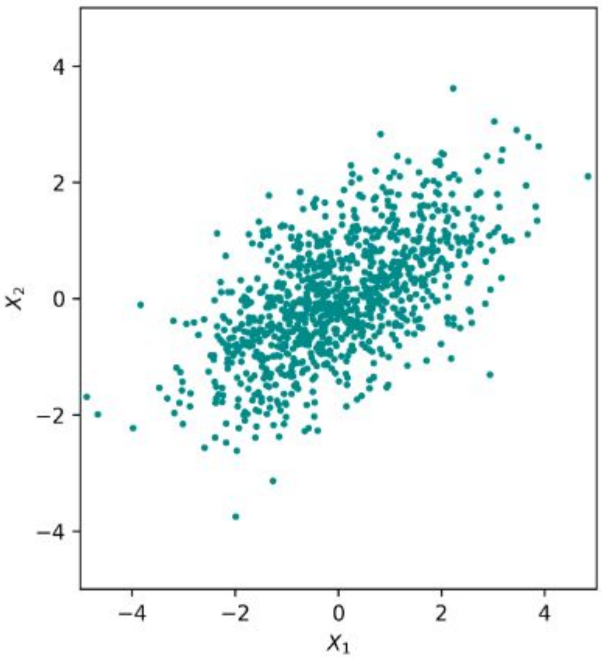
# Covariance: generalization for multidimensional data

A random vector $\boldsymbol{X}, \boldsymbol{X} \in \mathbb{R}^d$ has covariance matrix

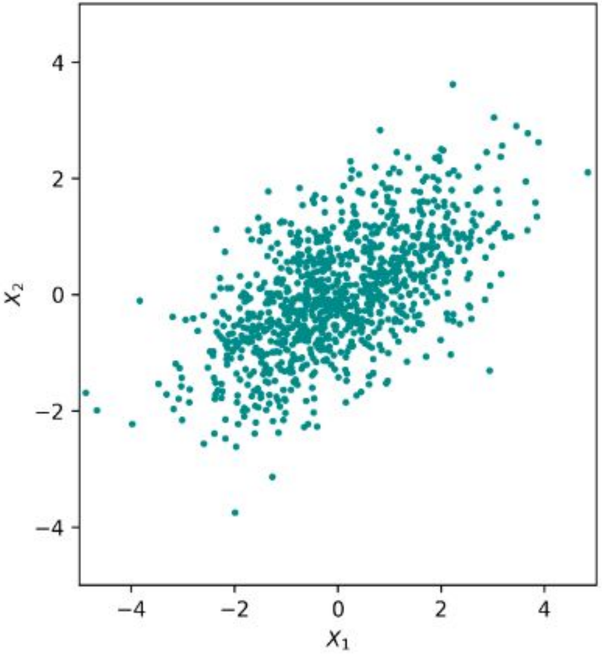
$$
\begin{aligned}
\boldsymbol{\Sigma} &= \operatorname{Cov}(\boldsymbol{X}) \\
&= E[(\boldsymbol{X} - E[\boldsymbol{X}])^T(\boldsymbol{X} - E[\boldsymbol{X}])] \\
&= \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \dots & \operatorname{Cov}(X_1, X_d) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \dots & \operatorname{Cov}(X_2, X_d) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_d, X_1) & \operatorname{Cov}(X_d, X_2) & \dots & \operatorname{Var}(X_d) \end{bmatrix},
\end{aligned}
$$

which is symmetric and positive semidefinite.

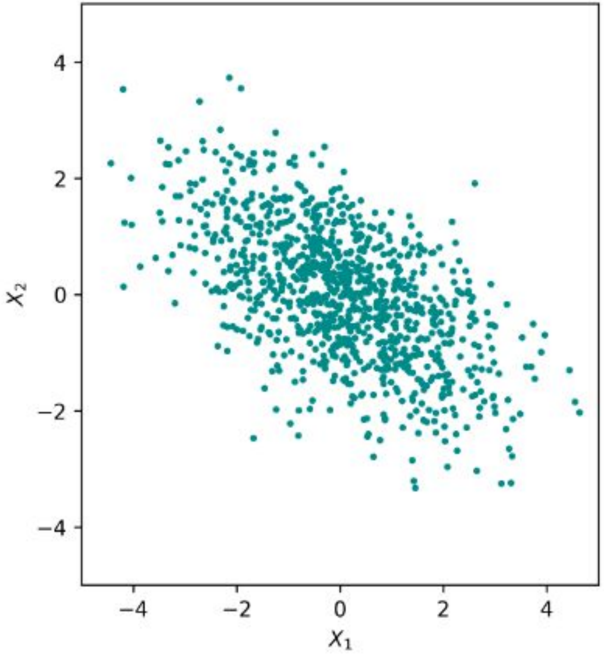
$$
\begin{aligned}
\boldsymbol{X} &\sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&\sim \frac{1}{\sqrt{2\pi^d \det(\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}
\end{aligned}
$$



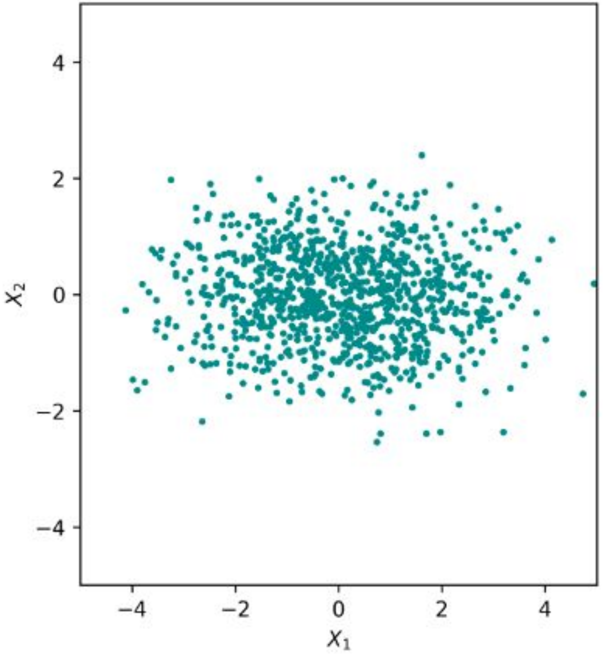
# Covariance matrix examples




```
[[2.05764803 0.86381417]
 [0.86381417 1.04250702]]
```




```
[[ 2.23043927 -0.85443664]
 [-0.85443664  1.15488608]]
```




```
[[ 2.66695315 -0.05916988]
 [-0.05916988  0.65315518]]
```

# Key Questions

I. How can we represent and sample from a distribution?

II. **How do you estimate the parameters and evaluate the model?**

III. Could this also apply to supervised learning?

IV. Summary + Housekeeping

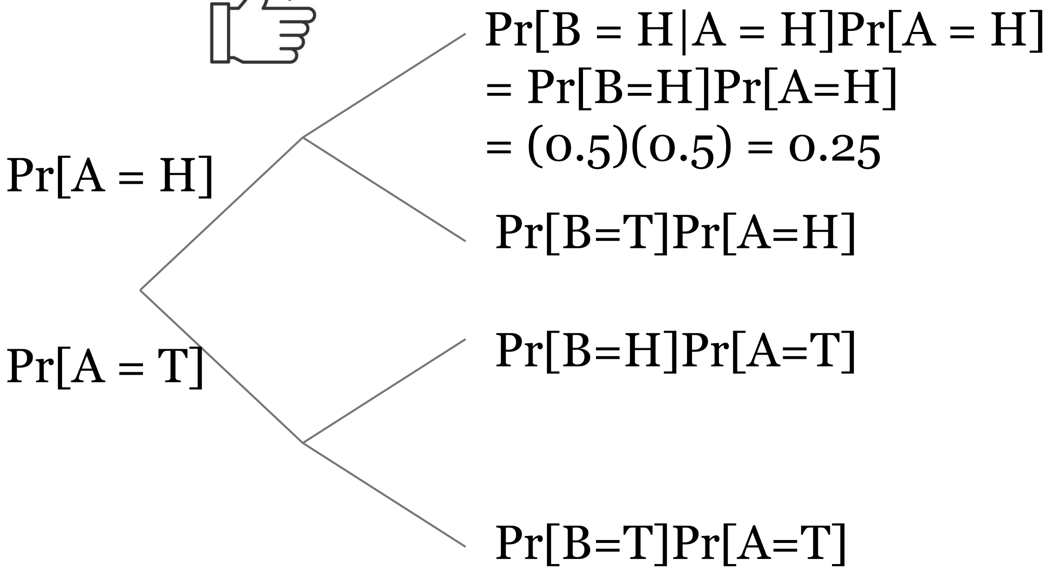
# The joint distribution describes behavior of combined densities

Suppose we have two independent univariate random variables, $X_1$ and $X_2$. Their joint probability distribution,

$$p(X_1, X_2) = p(X_2 \mid X_1)p(X_1),$$

describes the probability of the variables occurring together. If variables $X_1$ and $X_2$ are *independent*, this simplifies to:

$$p(X_1, X_2) = p(X_2)p(X_1).$$



# Example: estimating the parameters of a distribution

Suppose we have a set of n observations $\{x_1, x_2, \ldots, x_n\}$, $x_i \in \mathbb{R}$, and we assume that they are realizations of a univariate Gaussian (normal) distribution with some mean $\mu$ and variance $\sigma^2$:

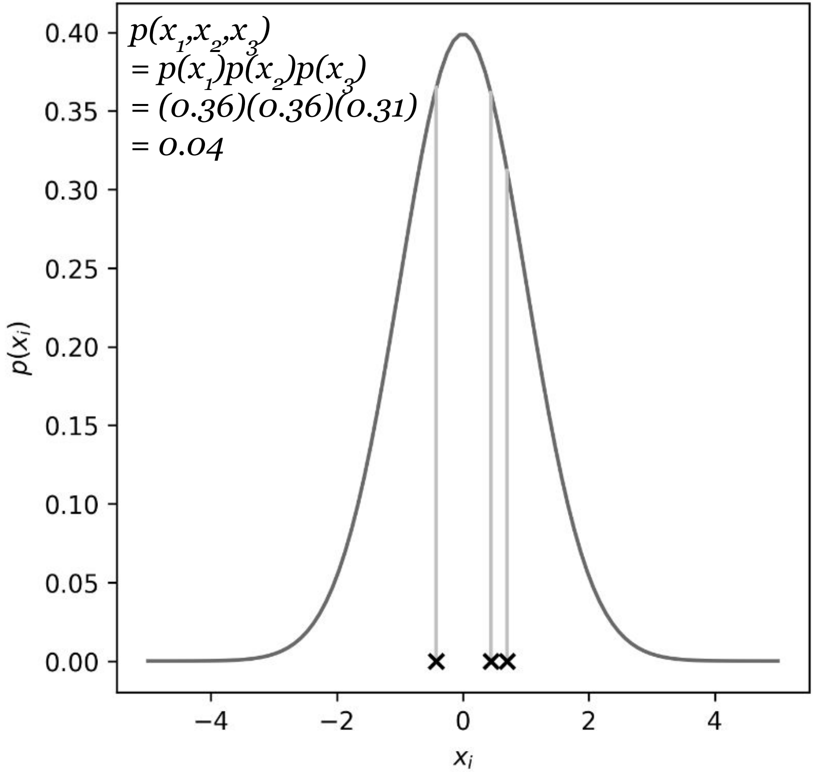
$$x_i \sim \mathcal{N}(\mu, \sigma^2)$$

The probability density for each observation $x_i$ is

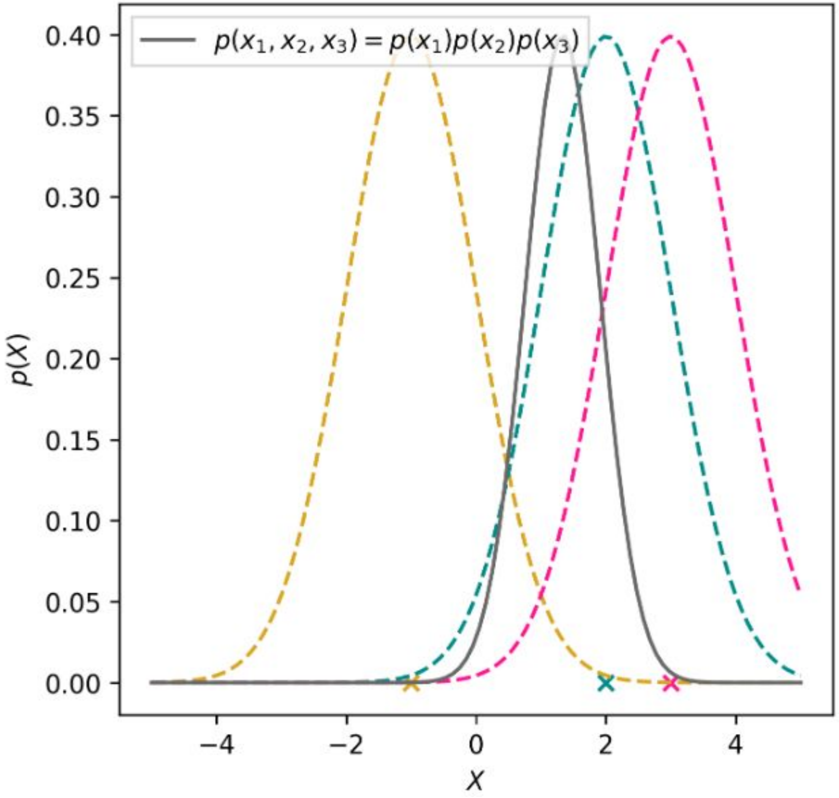
$$p(x_i \mid \mu, \sigma^2) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2},$$

with joint density

$$p(\boldsymbol{x} \mid \mu, \sigma^2) = \prod_{i=1}^{n} p(x_i \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}.$$



# Visual interpretation of joint probability distribution



# Likelihood considers the probability of parameters, given data

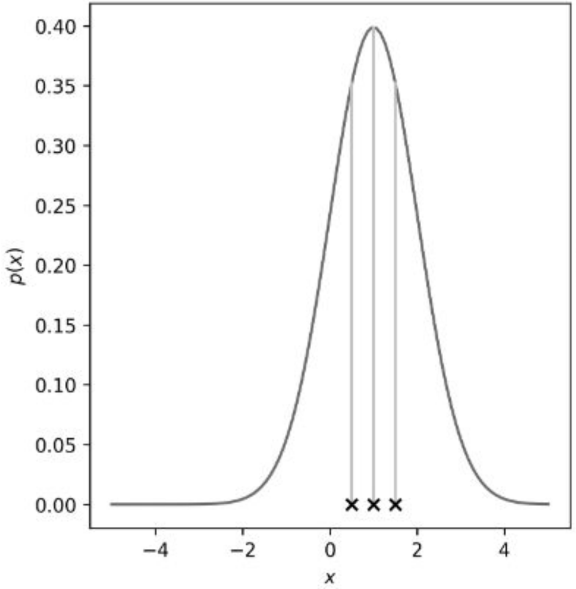
Joint density:

$$p(\boldsymbol{x} \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

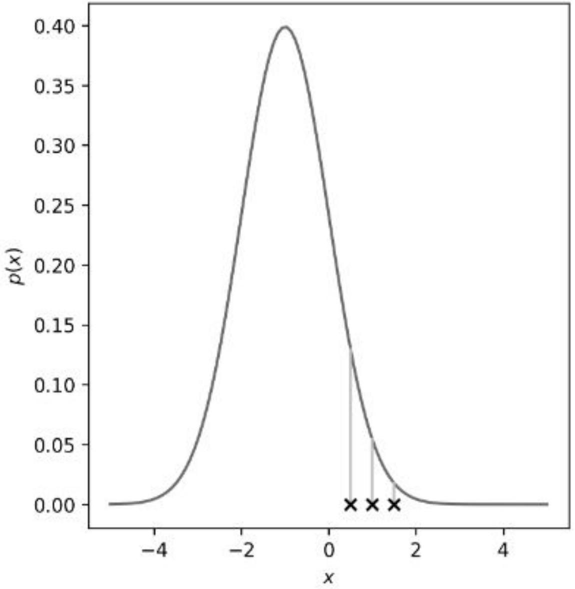
Likelihood:

$$\mathcal{L}(\mu, \sigma^2 \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

$\mu = 1, \sigma = 1$

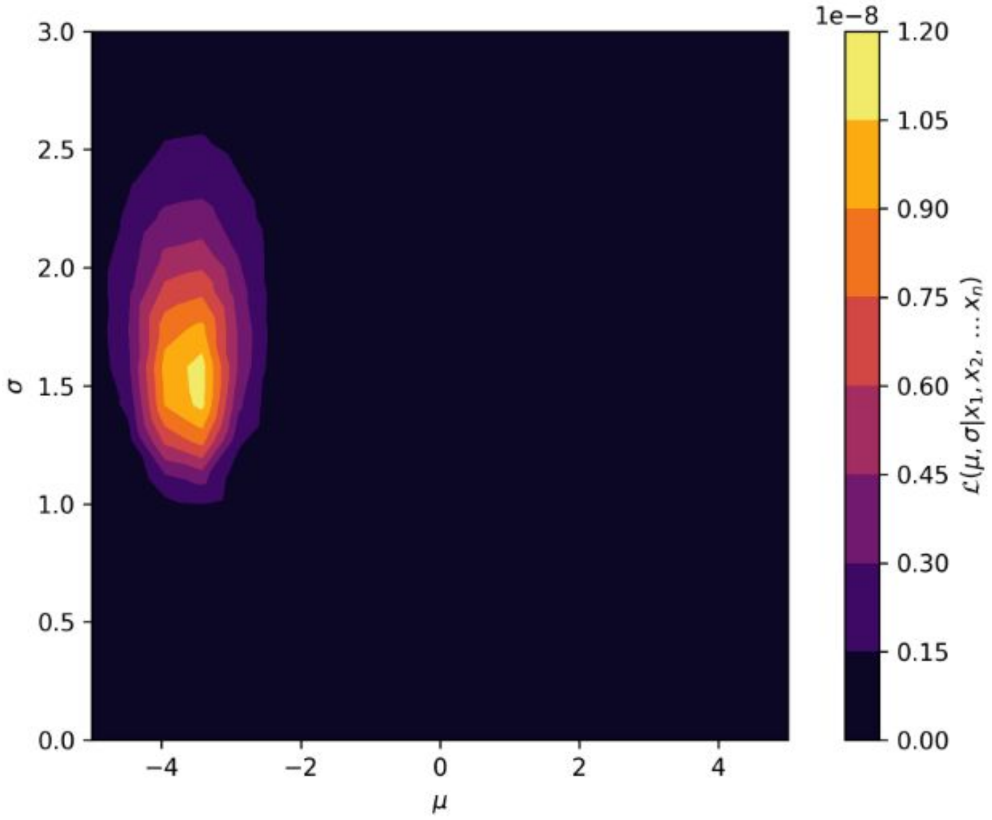



$\mu = -1, \sigma = 1$



# Finding the Maximum Likelihood Estimate (MLE)

$$\mathcal{L}(\mu, \sigma^2 \mid \boldsymbol{x}) = \prod_{i=1}^{n} \frac{1}{\sigma^2} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

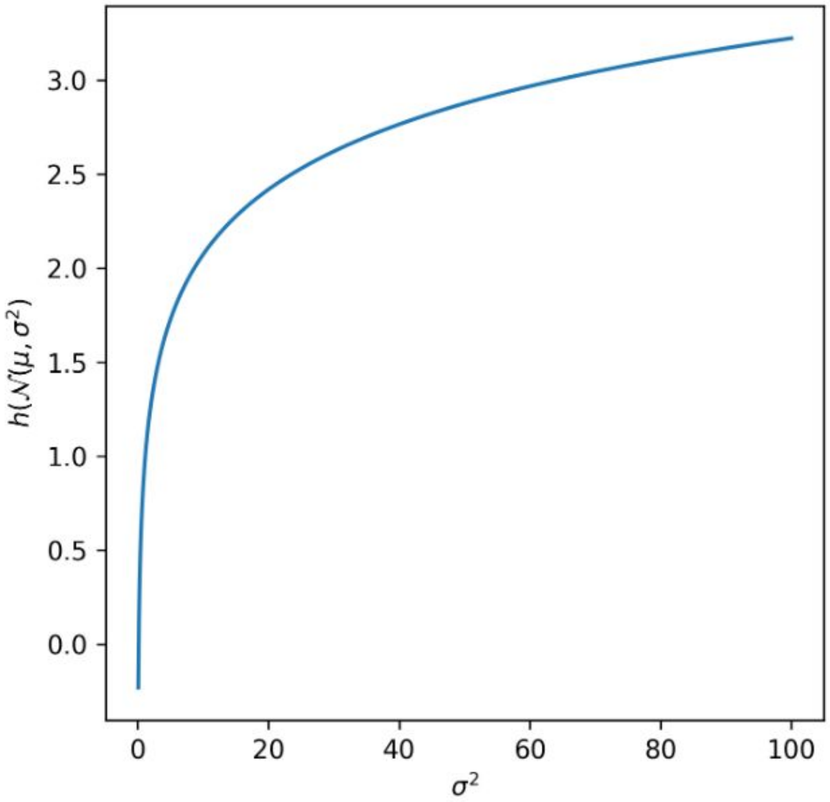


# Entropy characterizes the uncertainty in a distribution

# Example: Entropy of a Gaussian (normal) distribution

$$
\begin{aligned}
h(X) &= E[-\log p(X)] \\
&= -\int_{\mathcal{X}} p(x)\log p(x)dx
\end{aligned}
$$

$$h(X \sim \mathcal{N}(\mu, \sigma^2)) = \frac{1}{2}\log[2\pi e\sigma^2]$$

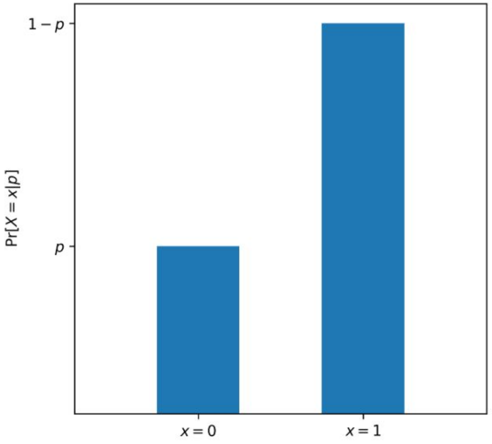


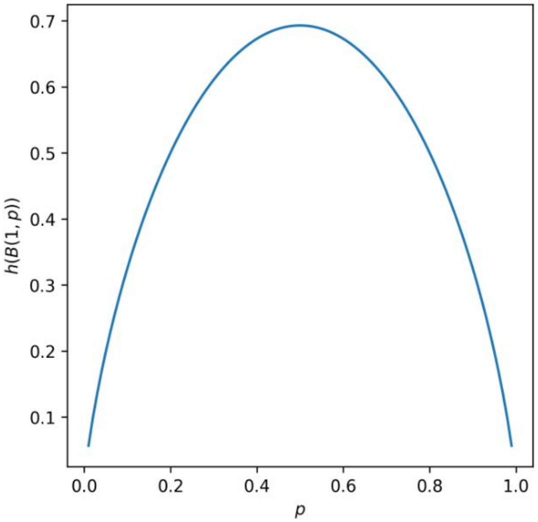
# Example: Entropy of a Bernoulli random variable

Bernoulli random variable:

$$\Pr[X = x] = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \end{cases}$$

where $0 \leq p \leq 1$.

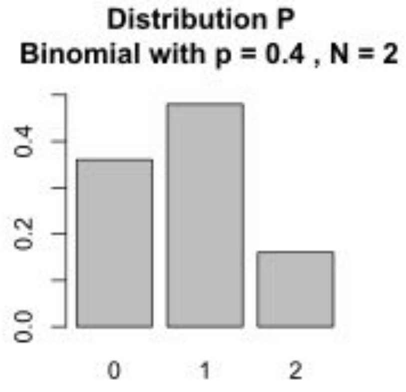


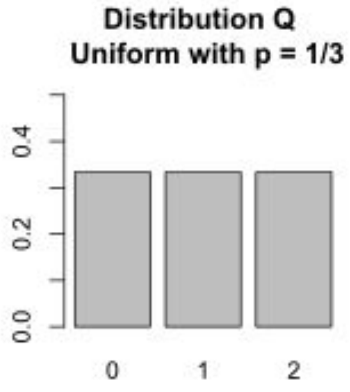



# Kullback-Leibler divergence measures dissimilarity between a reference and model distribution (aka relative entropy)

$$D_{\mathrm{KL}}(P \,||\, Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \text{ (Discrete Random Variable)}$$

$$D_{\mathrm{KL}}(P \,||\, Q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx \text{ (Continuous Random Variable)}$$




| | P | Q | Plog(P/Q) |
|---|---|---|---|
| X = 0 | 0.36 | ≅ 0.33 | ≅ 1.08 |
| X = 1 | 0.48 | ≅ 0.33 | ≅ 1.44 |
| X = 2 | 0.16 | ≅ 0.33 | ≅ -0.11 |
| | | | $D_{KL}$ ≅ 0.085 |

Source: Wikipedia

# Key Questions

I. How can we represent and sample from a distribution?

II. How do you estimate the parameters and evaluate the model?

**III. Could this also apply to supervised learning?**

IV. Summary + Housekeeping

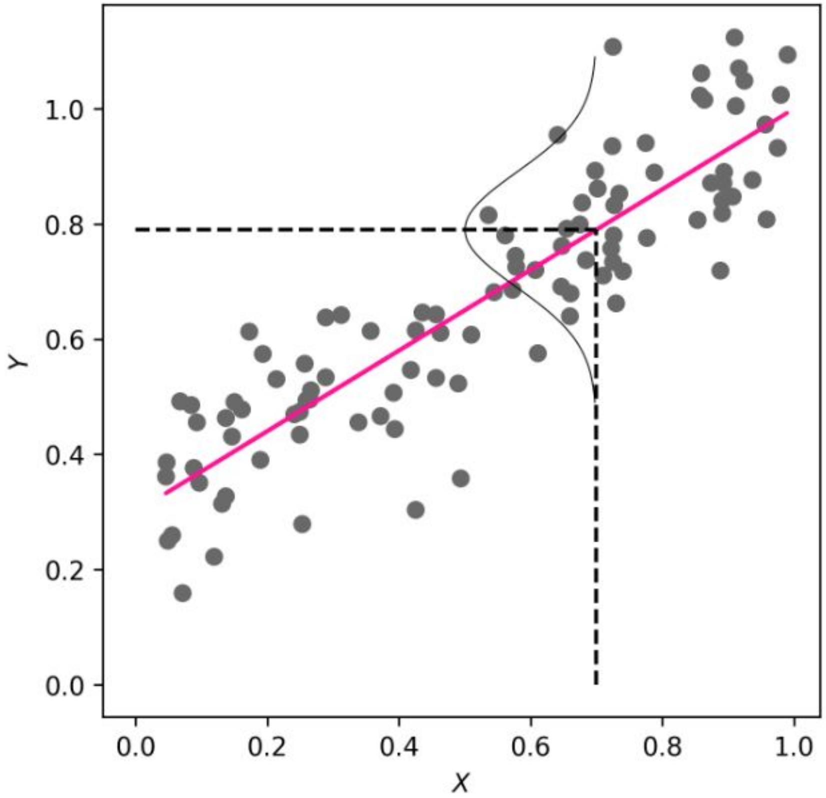
# Interpreting the linear regression problem with MLE

Consider a random variable $Y$ that follows a normal distribution with mean $\mu = w^T X$, where $X$ is another random variable, and variance $\sigma^2$:

$$Y \sim \mathcal{N}(w^T X, \sigma^2)$$
$$y_i = w^T x_i + \mathcal{N}(0, \sigma^2)$$
$$\text{c.f. } X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\mathcal{L}(\boldsymbol{y} \mid \boldsymbol{x}, w, \sigma^2) = \prod_{i=1}^{n} p(y_i \mid x_i, w, \sigma^2)$$

$$\mathcal{L}(z \mid A, w, \sigma^2) = \prod_{i=1}^{n} p(z_i \mid a_i, w, \sigma^2)$$



# Maximizing the likelihood with respect to the parameters $\boldsymbol{w}$

expected / mean

expected / mean

$$
\begin{aligned}
\mathcal{L}(z \mid \; A, w, \sigma^2) &= \prod_{i=1}^{n} p(z_i \mid a_i, w, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}(z_i - w^T a_i)^2} \\
\log\mathcal{L}(z \mid \; A, w, \sigma^2) &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw) \\
\text{argmax}_w \log\, \mathcal{L} &= \text{argmin}_w - \log\, \mathcal{L} \\
-\log\, \mathcal{L} &= \frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw) \\
\frac{\partial - \log\, \mathcal{L}}{\partial w} &= \frac{1}{2\sigma^2}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw) \\
&= \frac{1}{2\sigma^2}\left(w^T A^T A w - 2A^T w^T \boldsymbol{z} + \boldsymbol{z}^T \boldsymbol{z}\right) \\
\frac{\partial - \log\, \mathcal{L}}{\partial w} &= 0 \\
\implies \; w &= (A^T A)^{-1} A^T z
\end{aligned}
$$

# What about the variance?

$$\frac{\partial - \log \mathcal{L}}{\partial \sigma} = \frac{\partial}{\partial \sigma}\left[\frac{n}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw)\right]$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw)$$

$$\frac{\partial - \log \mathcal{L}}{\partial \sigma} = 0$$

$$\implies \frac{n}{\sigma} = \frac{1}{\sigma^3}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw)$$

$$\implies \sigma^2 = \frac{1}{n}(\boldsymbol{z} - Aw)^T(\boldsymbol{z} - Aw)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(a_i w - z_i)^2$$

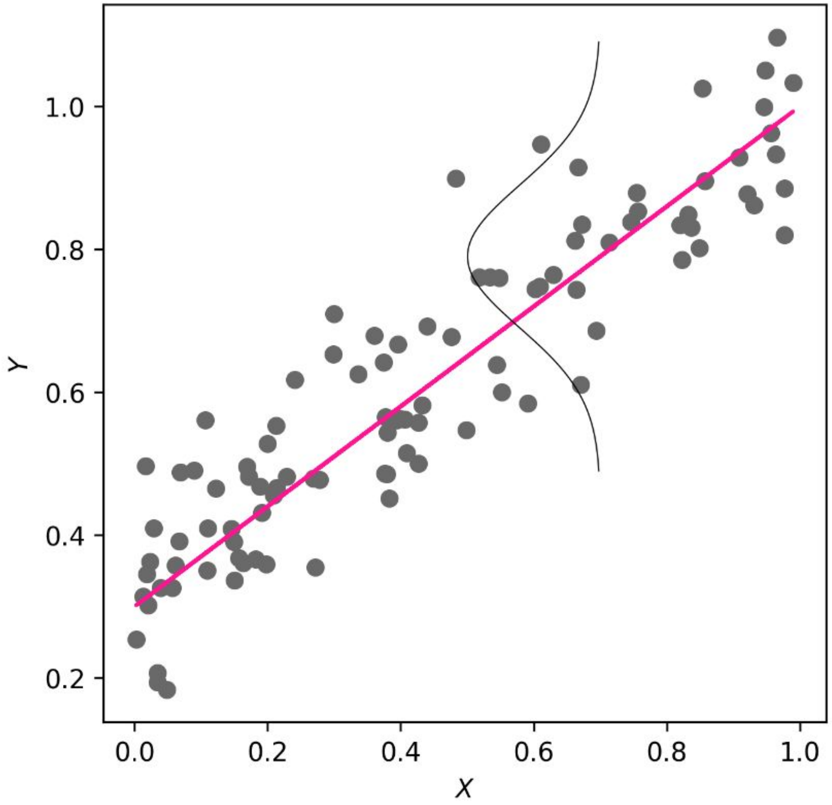
If $y = \log[f(x)]$, then

$$\frac{dy}{dx} = \frac{1}{f(x)}f'(x)$$

$$\implies \frac{d}{d\sigma}\left(-\frac{n}{2}\log(2\pi\sigma^2)\right) = -\frac{n}{2}\frac{1}{2\pi\sigma^2}4\pi\sigma$$

$$= -\frac{n}{\sigma}$$

# Under assumption of normally distributed errors, least-squares regression can be viewed as maximizing likelihood




UNIVERSITY OF WATERLOO | FACULTY OF MATHEMATICS

# Key Questions

I. How can we represent and sample from a distribution?

II. How do you estimate the parameters and evaluate the model?

III. Could this also apply to supervised learning?

IV. **Summary + Housekeeping**

# Lecture Objectives

At the end of the lecture, we should be able to:

- ★ Identify the probability density function and parameterization of widely used distributions and relate it to their use in constructing likelihood functions.
- ★ Construct the likelihood function for a dataset and maximize it to find the maximum likelihood estimates (MLE) of the parameters.
- ★ Define and apply information theoretic measures such as entropy and KL divergence to characterize and compare distributions.
- ★ Reformulate the linear regression objective using likelihood principles, and demonstrate that it can be viewed as a special case of MLE.

# Last lecture's slides have been updated

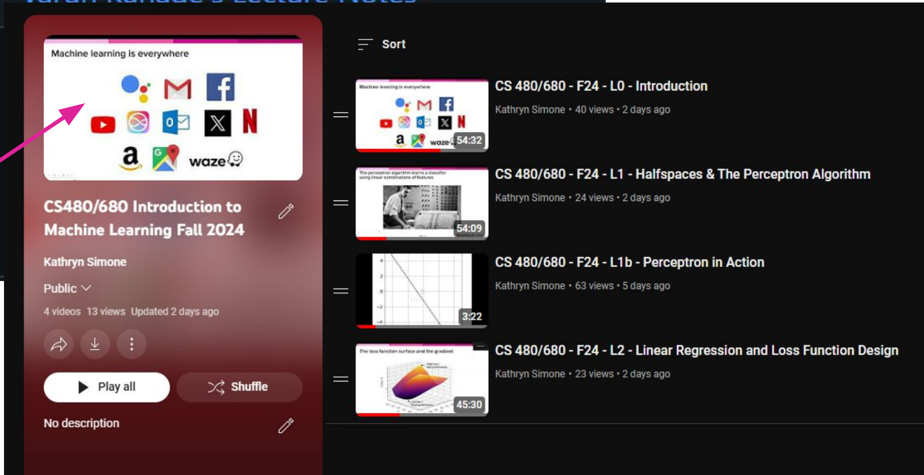
Errata

- On the slide titled, "Equivalent notation of loss to leverage the gradient" a reference was made to constructing a loss *matrix*. This has been corrected to "We can write the total loss, $L$ as ...".

- A previous version of the slide deck had a slide titled "If a function is convex, its second derivative is positive." This statement was incorrect because a convex function requires the second derivative to be non-negative ($\geq 0$, not strictly positive, $> 0$). Additionally, a function may be convex but not necessarily everywhere twice differentiable. The corrected statement reads: "A twice-differentiable function of more than one variable is convex *if and only if* its Hessian is everywhere positive semidefinite," emphasizing that this condition must hold for all points in the function's domain. This clarification highlights that having a positive semidefinite Hessian matrix everywhere in the domain is a sufficient and necessary condition for convexity in the context of twice-differentiable functions. However, convexity as a broader property does not inherently require the function to be twice differentiable or the Hessian to be defined everywhere.

# Lecture videos linked from course homepage, playlist also on YouTube

| LECTURE | TITLE | MATERIALS | SUPPLEMENTARY READINGS |
|---|---|---|---|
| 0 | Logistics & Introduction | Slides<br>Video Lecture | N/A |
| 1 | Halfspaces & The Perceptron Algorithm | Slides<br>Video Lecture<br>Perceptron Video | UML Section 9.1<br>ESL Section 4.5<br>Yaoliang Yu's Lecture Notes<br>Varun Kanade's Lecture Notes |
| 2 | Linear Regression & Loss Function Design | Slides<br>Video Lecture | |



| Lecture | Date | Topics |
|---|---|---|
| ~~0~~ | ~~05/09/2024~~ | ~~Introduction + Administrative Remarks~~ |
| ~~1~~ | ~~10/09/2024~~ | ~~Halfspaces the Perceptron Algorithm~~ |
| ~~2~~ | ~~12/09/2024~~ | ~~Linear Regression and Convexity~~ |
| ~~3~~ | ~~17/09/2024~~ | ~~Maximum Likelihood Estimation~~ |
| 4 | 19/09/2024 | k-means Clustering |
| 5 | 24/09/2024 | k-NN Classification and Logistic Regression |
| 6 | 26/09/2024 | Hard-margin SVM |
| 7 | 01/10/2024 | Soft-margin SVM |
| 8 | 03/10/2024 | Kernel methods |
| 9 | 08/10/2024 | Decision Trees |
| 10 | 10/10/2024 | Bagging and Boosting |
| | 15/10/2024 | NO LECTURE - MIDTERM BREAK |
| | 17/10/2024 | NO LECTURE- MIDTERM BREAK |
| 11 | 22/10/2024 | Expectation Maximization Algorithm |
| 12 | 24/10/2024 | MLPs and Fully-Connected NNs |
| | 29/10/2024 | NO LECTURE - MIDTERM EXAM |
| 13 | 31/10/2024 | Convolutional Neural Networks |
| 14 | 05/11/2024 | Recurrent Neural Networks |
| 15 | 07/11/2024 | Attention and Transformers |
| 16 | 12/11/2024 | Graph Neural Networks (Time permitting) |
| 17 | 14/11/2024 | VAEs and GANs |
| 18 | 19/11/2024 | Flows |
| 19 | 21/11/2024 | Contrastive Learning (Time permitting) |
| 20 | 26/11/2024 | Robustness |
| 21 | 28/11/2024 | Privacy (Saber Malekmohammadi) |
| 22 | 03/12/2024 | Fairness |

Non-parametric methods:

- Kernel Density Est.
- K-means Clustering
- K-NN Classification

| Lecture | Date | Topics |
|---|---|---|
| 0 | 05/09/2024 | Introduction + Administrative Remarks |
| 1 | 10/09/2024 | Halfspaces the Perceptron Algorithm |
| 2 | 12/09/2024 | Linear Regression and Convexity |
| 3 | 17/09/2024 | Maximum Likelihood Estimation |
| 4 | 19/09/2024 | Non-parametric Methods |
| 5 | 24/09/2024 | Logistic Regression |
| 6 | 26/09/2024 | Hard-margin SVM |
| 7 | 01/10/2024 | Soft-margin SVM |
| 8 | 03/10/2024 | Kernel methods |
| 9 | 08/10/2024 | Decision Trees |
| 10 | 10/10/2024 | Bagging and Boosting |
| | 15/10/2024 | NO LECTURE - MIDTERM BREAK |
| | 17/10/2024 | NO LECTURE- MIDTERM BREAK |
| 11 | 22/10/2024 | Expectation Maximization Algorithm |
| 12 | 24/10/2024 | MLPs and Fully-Connected NNs |
| | 29/10/2024 | NO LECTURE - MIDTERM EXAM |
| 13 | 31/10/2024 | Convolutional Neural Networks |
| 14 | 05/11/2024 | Recurrent Neural Networks |
| 15 | 07/11/2024 | Attention and Transformers |
| 16 | 12/11/2024 | Graph Neural Networks (Time permitting) |
| 17 | 14/11/2024 | VAEs and GANs |
| 18 | 19/11/2024 | Flows |
| 19 | 21/11/2024 | Contrastive Learning (Time permitting) |
| 20 | 26/11/2024 | Robustness |
| 21 | 28/11/2024 | Privacy (Saber Malekmohammadi) |
| 22 | 03/12/2024 | Fairness |

# On the horizon

Questions? Ask Saber! :)

**Table 2: Grading Scheme**

| Assessment | Assessment Date | Weighting (CS480) | Weighting (CS680) |
|---|---|---|---|
| Assignment 1 | September 27 | 7.5% | 7.5% |
| Assignment 2 | October 14 | 7.5% | 7.5% |
| Assignment 3 | November 8 | 7.5% | 7.5% |
| Assignment 4 | November 22 | 7.5% | 7.5% |
| Exams | | | |
| Midterm | October 29 | 30% | 15% |
| Final | TBD | 40% | 30% |
| Project (CS 680 only) | | | |
| Pitch | September 19 | N/A | 2% |
| Proposal | October 8 | N/A | 8% |
| Report | December 3 | N/A | 15% |
| Total | | 100% | 100% |

Thursday! →