

# Math Review

Professor Ameet Talwalkar

Slide Credit: Professor Fei Sha

# Outline

- 1 Overview
- 2 Review on Probability
- 3 Review on Statistics
- 4 An integrative example

# How to grasp machine learning well

## Three pillars to machine learning

- Statistics
- Linear Algebra
- Optimization

## Resources to study them

- Suggested Reading:
  - ▶ Chapter 2 of MLAPA book
  - ▶ Linear Algebra Review and Reference by Zico Kolter and Chuong Do (<http://www.cs.cmu.edu/~zkolter/course/15-884/linalg-review.pdf>)
  - ▶ Convex Optimization Review by Zico Kolter and Honglak Lee (<http://www.cs.cmu.edu/~./15381/slides/cvxopt.pdf>)
- Wikipedia (some information might not be 100% accurate, though)

# Outline

- 1 Overview
- 2 Review on Probability
- 3 Review on Statistics
- 4 An integrative example

# Probability: basic definitions

**Sample Space:** a set of all possible outcomes or realizations of some random trial.

*Example:* Toss a coin twice; the sample space is  $\Omega = \{HH, HT, TH, TT\}$ .

**Event:** A subset of sample space

*Example:* the event that at least one toss is a head is  $A = \{HH, HT, TH\}$ .

**Probability:** We assign a real number  $P(A)$  to each event  $A$ , called the probability of  $A$ .

**Probability Axioms:** The probability  $P$  must satisfy three axioms:

- 1  $P(A) \geq 0$  for every  $A$ ;
- 2  $P(\Omega) = 1$ ;
- 3 If  $A_1, A_2, \dots$  are disjoint, then  $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

# Random Variables

**Definition:** A random variable is a function that maps from the sample space to the reals ( $X : \Omega \rightarrow R$ ), i.e., it assigns a real number  $X(\omega)$  to each outcome  $\omega$ .

*Example:*  $X$  returns 1 if a coin is heads and 0 if a coin is tails.  $Y$  returns the number of heads after 3 flips of a fair coin.

Random variables can take on many values, and we are often interested in the distribution over the values of a random variable, e.g.,  $P(Y = 0)$

# Common Distributions

Discrete variable	Probability function	Mean	Variance
<b>Uniform</b> $X \sim U[1, \dots, N]$	$1/N$	$\frac{N+1}{2}$	
<b>Binomial</b> $X \sim Bin(n, p)$	$\binom{n}{x} p^x (1-p)^{(n-x)}$	$np$	
<b>Geometric</b> $X \sim Geom(p)$	$(1-p)^{x-1} p$	$1/p$	
<b>Poisson</b> $X \sim Poisson(\lambda)$	$\frac{e^{-\lambda} \lambda^x}{x!}$	$\lambda$	
Continuous variable	Probability density function	Mean	Variance
<b>Uniform</b> $X \sim U(a, b)$	$1/(b-a)$	$(a+b)/2$	
<b>Gaussian</b> $X \sim N(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$	$\mu$	
<b>Gamma</b> $X \sim \Gamma(\alpha, \beta) (x \geq 0)$	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$	$\alpha\beta$	
<b>Exponential</b> $X \sim exponen(\beta)$	$\frac{1}{\beta} e^{-\frac{x}{\beta}}$	$\beta$	

# Distribution Function

**Definition:** Suppose  $X$  is a random variable,  $x$  is a specific value that it can take,

*Cumulative distribution function (CDF)* is the function  $F : \mathcal{R} \rightarrow [0, 1]$ , where  $F(x) = P(X \leq x)$ .

If  $X$  is discrete  $\Rightarrow$  *probability mass function*:  $f(x) = P(X = x)$ .

If  $X$  is continuous  $\Rightarrow$  *probability density function* for  $X$  if there exists a function  $f$  such that  $f(x) \geq 0$  for all  $x$ ,  $\int_{-\infty}^{\infty} f(x)dx = 1$  and for every  $a \leq b$ ,

$$P(a \leq X \leq b) = \int_a^b f(x)dx.$$

If  $F(x)$  is differentiable everywhere,  $f(x) = F'(x)$ .



# Expectation

## Expected Values

- Discrete random variable  $X$ ,  $E[g(X)] = \sum_{x \in \mathcal{X}} g(x) f(x)$ ;
- Continuous random variable  $X$ ,  $E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)$

**Mean and Variance**  $\mu = E[X]$  is the mean;  $var[X] = E[(X - \mu)^2]$  is the variance.

We also have  $var[X] = E[X^2] - \mu^2$ .

# Multivariate Distributions

## Definition:

$$F_{X,Y}(x, y) := P(X \leq x, Y \leq y),$$

and

$$f_{X,Y}(x, y) := \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y},$$

**Marginal Distribution of  $X$**  (Discrete case):

$$f_X(x) = P(X = x) = \sum_y P(X = x, Y = y) = \sum_y f_{X,Y}(x, y)$$

or  $f_X(x) = \int_y f_{X,Y}(x, y) dy$  for continuous variable.

# Conditional Probability and Bayes Rule

**Conditional probability** of  $X$  given  $Y = y$  is

$$f_{X|Y}(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

**Bayes Rule:**

$$P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

# Independence

**Independent Variables**  $X$  and  $Y$  are *independent* if and only if:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

or  $f_{X,Y}(x, y) = f_X(x)f_Y(y)$  for all values  $x$  and  $y$ .

**IID variables:** *Independent and identically distributed* (IID) random variables are drawn from the same distribution and are all mutually independent.

**Linearity of Expectation:** Even if  $X_1, \dots, X_n$  are not independent,

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i].$$

# Outline

- 1 Overview
- 2 Review on Probability
- 3 Review on Statistics**
- 4 An integrative example

# Statistics

Suppose  $X_1, \dots, X_n$  are random variables:

**Sample Mean:**

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i$$

**Sample Variance:**

$$S_{N-1}^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2.$$

If  $X_i$  are iid:

$$E[\bar{X}] = E[X_i] = \mu,$$

$$\text{Var}(\bar{X}) = \sigma^2 / N,$$

$$E[S_{N-1}^2] = \sigma^2$$

# Point Estimation

**Definition** The *point estimator*  $\hat{\theta}_N$  is a function of samples  $X_1, \dots, X_N$  that approximates a parameter  $\theta$  of the distribution of  $X_i$ .

**Sample Bias:** The bias of an estimator is

$$\text{bias}(\hat{\theta}_N) = E_{\theta}[\hat{\theta}_N] - \theta$$

An estimator is *unbiased estimator* if  $E_{\theta}[\hat{\theta}_N] = \theta$

# Example

Suppose we have observed  $N$  realizations of the random variable  $X$ :

$$x_1, x_2, \dots, x_N$$

Then,

- Sample mean  $\bar{X} = \frac{1}{N} \sum_n x_n$  is an unbiased estimator of  $X$ 's mean.
- Sample variance  $S_{N-1}^2 = \frac{1}{N-1} \sum_n (x_n - \bar{X})^2$  is an unbiased estimator of  $X$ 's variance
- Sample variance  $S_N^2 = \frac{1}{N} \sum_n (x_n - \bar{X})^2$  is *not* an unbiased estimator of  $X$ 's variance



# Outline

- 1 Overview
- 2 Review on Probability
- 3 Review on Statistics
- 4 An integrative example**

# ***Outline***

**Maximum likelihood estimation**

**Optimization**

**Convexity**

# Maximum likelihood estimation (MLE)



## Intuitive example

Estimate a coin toss

I have seen 3 flips of heads, 2 flips of tails, what is the chance of heads (or tails) of my next flip?

Model

Each flip is a Bernoulli random variable  $X$

$X$  can take only two values: 1 (heads), 0 (tails)



$$p(X = 1) = \theta$$



$$p(X = 0) = 1 - \theta$$

**Parameter to be identified from data**

# Principles of MLE

## 5 (independent) trials

### Observations



$$X_1 = 1$$



$$X_2 = 0$$



$$X_3 = 1$$



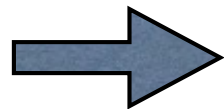
$$X_4 = 1$$



$$X_5 = 0$$

### Likelihood of all the 5 observations

$$\theta \times (1 - \theta) \times \theta \times \theta \times (1 - \theta)$$



$$\mathcal{L} = \theta^3 (1 - \theta)^2$$

### Intuition

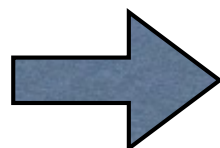
choose  $\theta$  such that  $\mathcal{L}$  is maximized

# Maximizing the likelihood

## Solution



$$\mathcal{L} = \theta^3 (1 - \theta)^2$$



$$\theta^{MLE} = \frac{3}{3 + 2}$$

**(Detailed derivation later)**

## Intuition

**Probability of head is the percentage of heads in the total flips.**

# More generally,

**Model (ie, assuming how data is distributed)**

$$X \sim P(X; \theta)$$


**Training data (observations)**

$$\mathcal{D} = \{x_1, x_2, \dots, x_N\}$$

**Maximum likelihood estimate**

$$\mathcal{L}(\mathcal{D}) = \prod_{i=1}^N P(x_i; \theta) \quad \theta^{MLE} = \arg \max_{\theta} \mathcal{L}(\mathcal{D})$$

**log-likelihood**


$$= \arg \max_{\theta} \sum_{i=1}^N \log P(x_i; \theta)$$

# Ex: estimate parameters of Gaussian distribution

## Model with unknown parameters

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

## Observations

$$\mathcal{D} = \{x_1, x_2, \dots, x_N\}$$

## Log-likelihood

$$\ell(\mu, \sigma) = \sum_{n=1}^N \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi}\sigma \right\}$$

# Solution

We will solve the following later

$$\arg \max_{\mu, \sigma} \ell(\mu, \sigma) = \sum_{n=1}^N \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi\sigma} \right\}$$

But the solution is given in the below

$$\mu = \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n \qquad \sigma^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})^2$$



# *Caveats for complicated models*

## **No closed-form solution**

**Use numerical optimization**

**many easy-to-use, robust packages are available**

**Stuck in local optimum (more on this later)**

**Restart optimization with random initialization**

## **Computational tractability**

**Can be difficult to compute likelihood  $\mathcal{L}(\mathcal{D})$  exactly**

**Need to approximate**

# Optimization

Given an objective function

$$f(x)$$

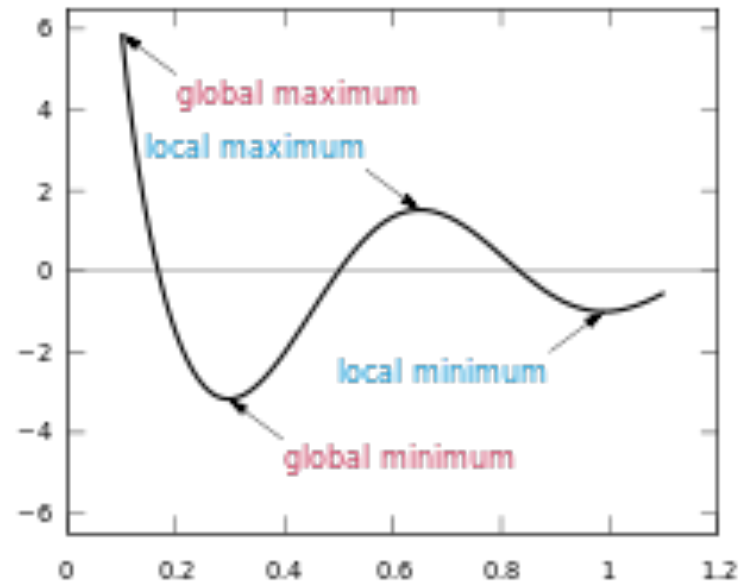
how do we find its minimum

$$\min f(x)$$

optionally, under constraints

$$\text{such that } g(x) = 0$$

difference between  
global and local optimal

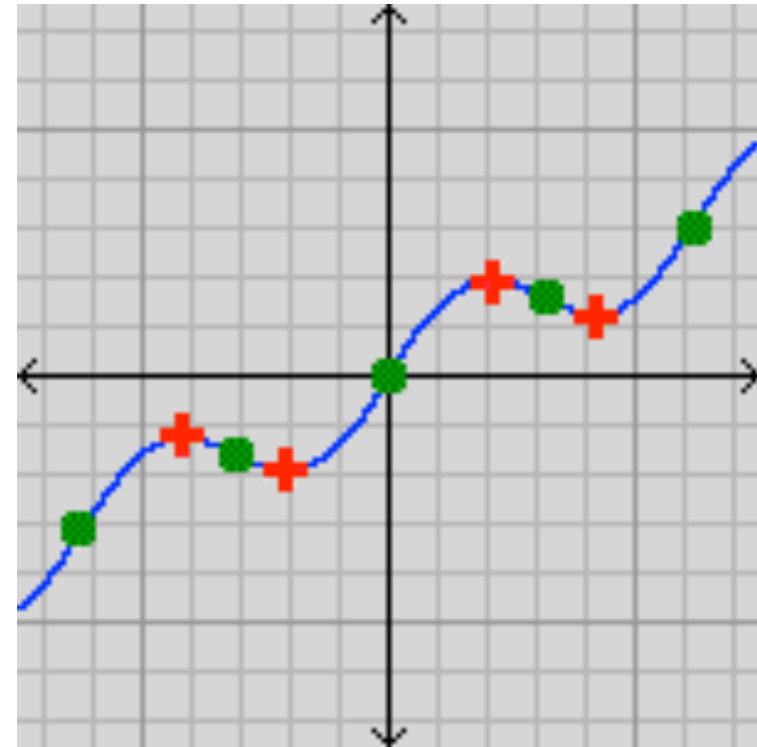


# Unconstrained optimization

## Fermat's Theorem

Local optima occurs at stationary points, namely, where gradients vanish

$$f'(x) = 0$$



# Simple example

What is the minimum of

$$f(x) = x^2$$

Gradient is

$$f'(x) = 2x$$

Set the gradient to zero

$$f'(x) = 0 \rightarrow x = 0$$

**Namely,  $x = 0$  is locally optimum (minimum and global, actually)**

# Remember the MLE of tossing coin?

## 5 (independent) trials

### Observation



$$X_1 = 1$$



$$X_2 = 0$$



$$X_3 = 1$$



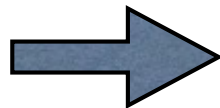
$$X_4 = 1$$



$$X_5 = 0$$

### Likelihood of all the 5 observations

$$\theta \times (1 - \theta) \times \theta \times \theta \times (1 - \theta)$$



$$\mathcal{L} = \theta^3 (1 - \theta)^2$$

# Maximizing the likelihood

the objective function is

$$L(\theta) = \theta^3(1 - \theta)^2$$

The gradient is

$$L'(\theta) = 3\theta^2(1 - \theta)^2 - 2\theta^3(1 - \theta)$$

Set gradient to zero

$$L'(\theta) = 0 \rightarrow \theta = \frac{3}{3 + 2}$$

# ***Wait a second***

**The gradient also vanishes if  $\theta = 0$**

$$L'(\theta) = 3\theta^2(1 - \theta)^2 - 2\theta^3(1 - \theta)$$

**Obviously,  $\theta = 0$  does not maximize  $L(\theta)$**

**Stationary points are only **necessary** for (local) optimum**

# Multivariate optimization

## Log-likelihood for Gaussian distribution

$$\arg \max_{\mu, \sigma} \ell(\mu, \sigma) = \sum_{n=1}^N \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} - \log \sqrt{2\pi\sigma} \right\}$$

## Partial derivatives

$$\frac{\partial \ell}{\partial \mu} = \sum_n^N -\frac{2(x_n - \mu)}{2\sigma^2}$$

$$\frac{\partial \ell}{\partial \sigma} = \sum_n^N \left\{ \frac{(x_n - \mu)^2}{\sigma^3} - \frac{1}{\sigma} \right\}$$



# Stationary points defined by sets of equations

$$\frac{\partial \ell}{\partial \mu} = 0 \rightarrow \mu = \frac{1}{N} \sum_n x_n$$

$$\frac{\partial \ell}{\partial \sigma} = 0 \rightarrow \sigma^2 = \frac{1}{N} \sum_n (x_n - \mu)^2$$

**We can use the first one to solve the mean**

**and the second one to compute the standard deviation**

# *a loophole?*

**In both models, parameters are constrained**

$\theta$ : should be non-negative and be less 1

$\sigma$ : should be non-negative

**But the optimization did not enforce the constraint**

yes, we are lucky

# Constrained optimization

## Equality Constraints

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g(x) = 0 \end{array}$$

## Method of Lagrange multipliers

**Construct the following function (Lagrangian)**

$$L(x, \lambda) = f(x) + \lambda g(x)$$

# *More difficult situations*

## Inequality constraints

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & g(x) \leq 0 \end{array}$$

**generally are harder**

**We won't deal with these types of problems in its most general case**

**However, we will see some special instances.**

# Optimizing Convex functions

## Definition

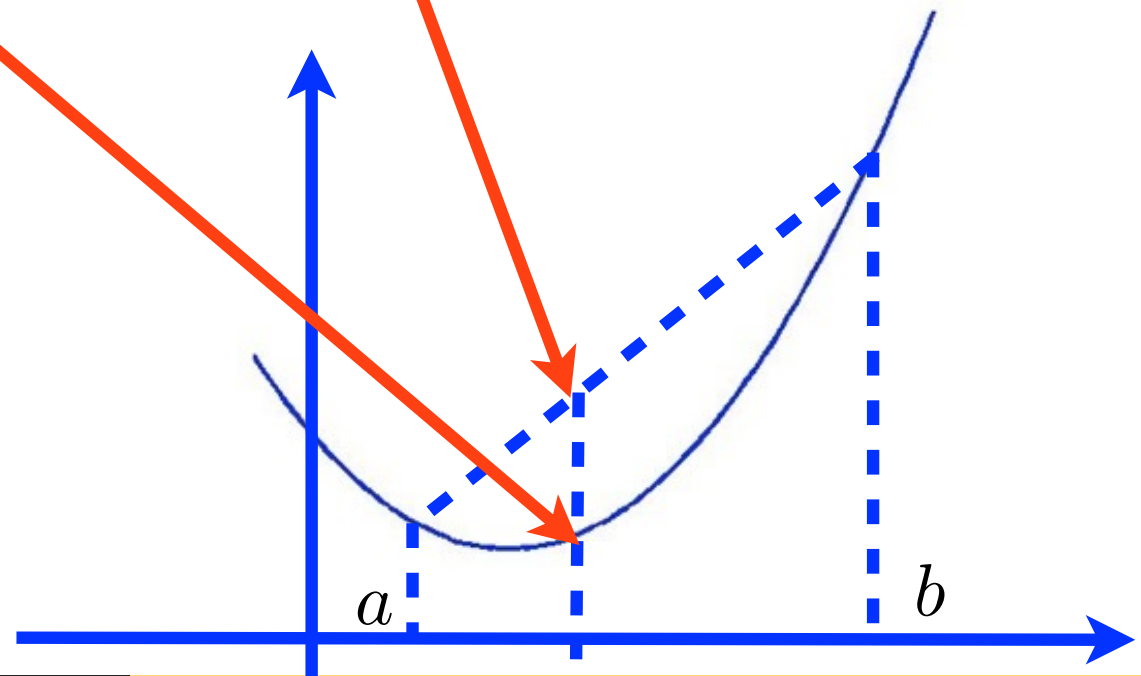
A function  $f(x)$  is convex if

$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b)$$

for

$$0 \leq \lambda \leq 1$$

Graphically,



# Local vs. global optimal

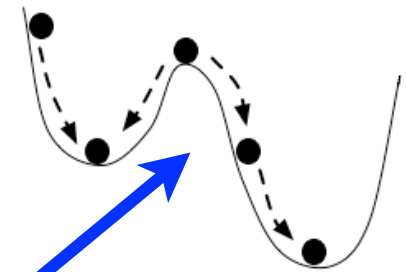
**For general objective functions  $f(x)$**

**We get local optimum**

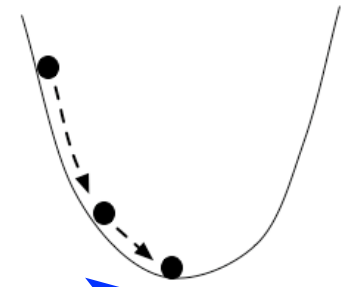
**For convex functions**

**the local optimum is the global optimum**

**Consider rolling a ball on a hill**



**depends on where you start**



**does not depend on where you start**

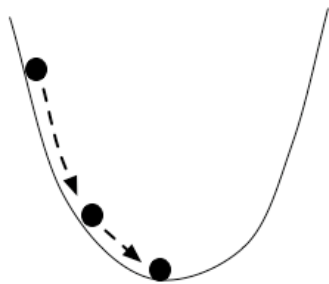
# Local vs. global optimal

In practice, convexity can be a very nice thing

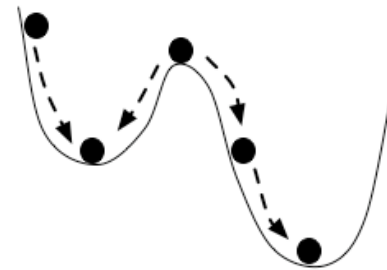
In general, convex problems -- minimizing a convex function over a convex set -- can be solved numerically very **efficiently**

This is advantageous especially if stationary points cannot be found analytically in closed-form

Convex: unique global optimum



nonconvex: local optimum



# Examples

## Convex functions

$$f(x) = x$$

$$f(x) = x^2$$

$$f(x) = e^x$$

$$f(x) = \frac{1}{x} \quad \text{when } x \geq 0$$



# Examples

## Nonconvex function

$$f(x) = \cos(x)$$

$$f(x) = e^x - x^2$$

**Difference in convex functions is not convex**



$$f(x) = \log x$$

**log (x) is called concave as its negation is convex**



# How to determine convexity?

**f(x) is convex if**

$$f''(x) \geq 0$$

## Examples

$$(-\log(x))'' = \frac{1}{x^2}$$

$$(\log(1 + e^x))'' = \left( \frac{e^x}{1 + e^x} \right)' = \frac{e^x}{(1 + e^x)^2}$$

# Multivariate functions

## Definition

$f(\mathbf{x})$  is convex if

$$f(\lambda \mathbf{a} + (1 - \lambda) \mathbf{b}) \leq \lambda f(\mathbf{a}) + (1 - \lambda) f(\mathbf{b})$$

## How to determine convexity in this case?

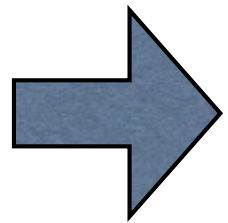
Second-order derivative becomes Hessian matrix

$$\mathbf{H} = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_D} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_D} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_D} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_D} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_D^2} \end{bmatrix}$$

# Convexity for multivariate function

If the Hessian is positive semidefinite, then the function is convex

**Ex:**  $f(\mathbf{x}) = \frac{x_1^2}{x_2}$



$$H = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} = \frac{2}{x_2^3} \begin{bmatrix} x_2^2 & -x_1x_2 \\ -x_1x_2 & x_1^2 \end{bmatrix}$$

**What does this function look like?**

