

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 head (or tail) of my next flip?



Model

Each flip is a Bernoulli random variable X .

X can take only two values: 1 (head), 0 (tail)



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



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

Parameter to be identified from data

Principles of MLE

5 (independent) trials

Observations



$$X_1 = 1$$



$$X_2 = 0$$



$$X_3 = 1$$



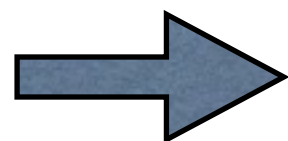
$$X_4 = 1$$



$$X_5 = 0$$

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

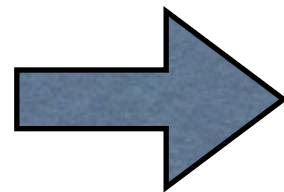
choose θ such that \mathcal{L} is maximized

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

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

$$\begin{aligned} \theta^{MLE} &= \arg \max_{\theta} \mathcal{L}(\mathcal{D}) \\ &= \arg \max_{\theta} \sum_{i=1}^N \log P(x_i; \theta) \end{aligned}$$

log-likelihood



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

Difficult to compute likelihood $\mathcal{L}(\mathcal{D})$ exactly

Need to approximate

Optimization

Given an objective function

$$f(x)$$

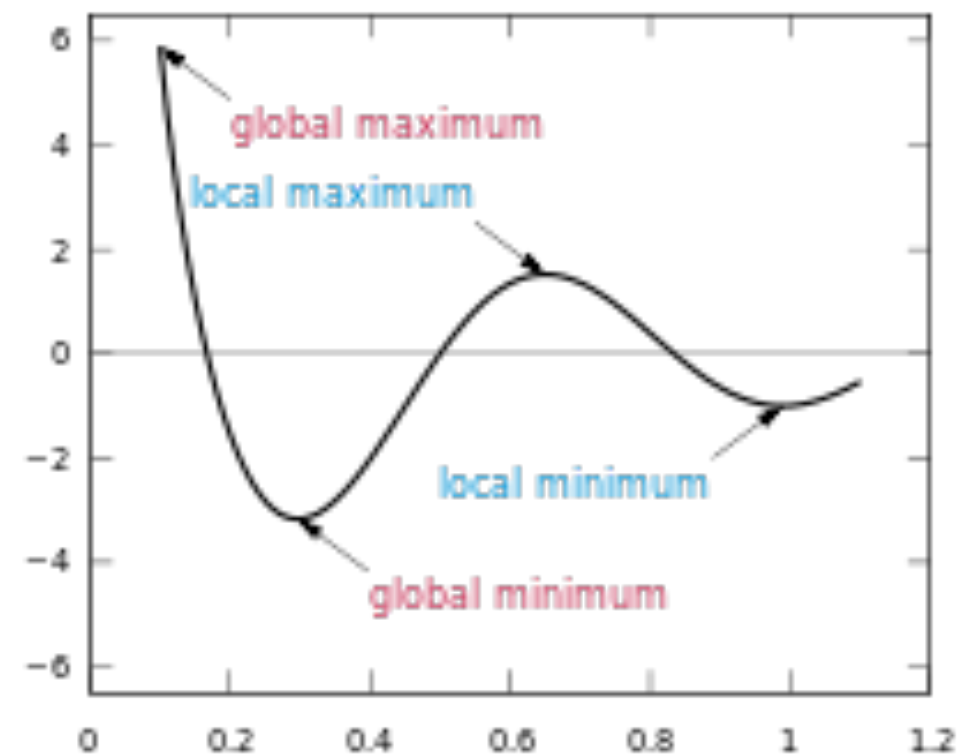
how do we find its minimum

$$\min f(x)$$

optionally, under constraints

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

difference between
global and local optimal

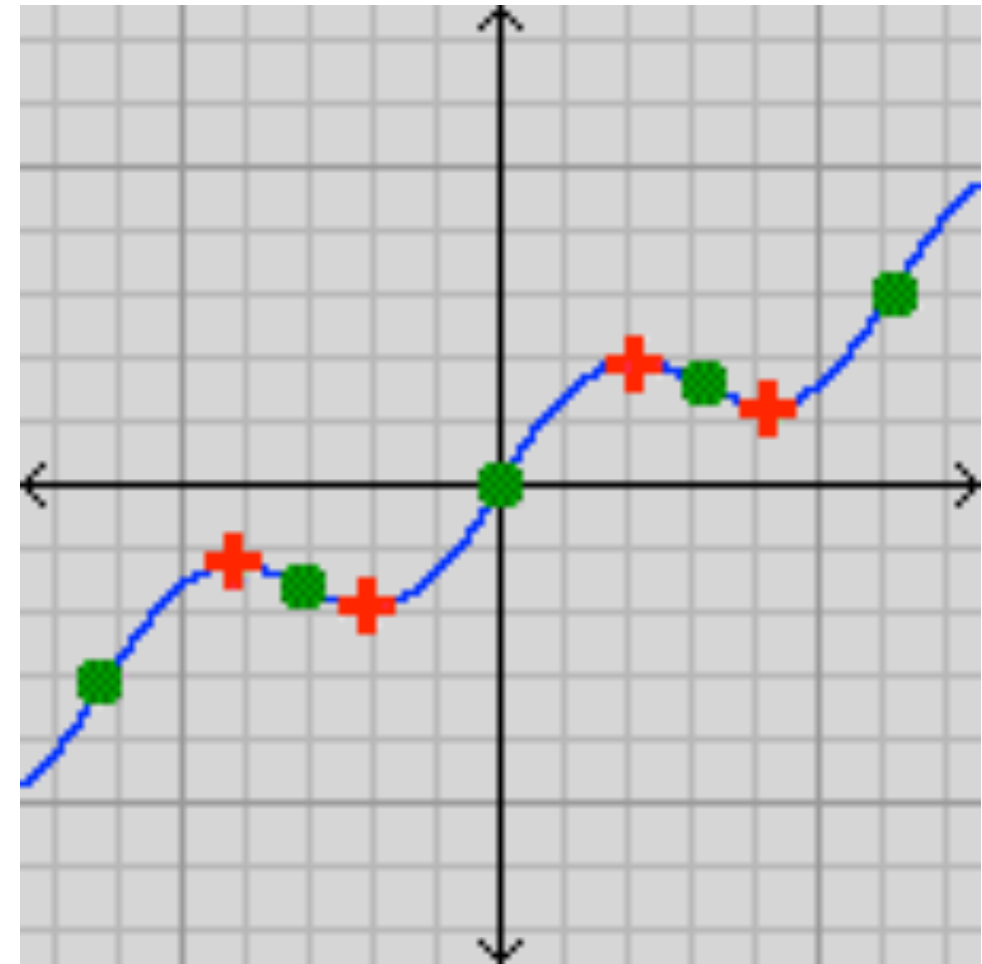


Unconstrained optimization

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



$$X_1 = 1$$



$$X_2 = 0$$



$$X_3 = 1$$



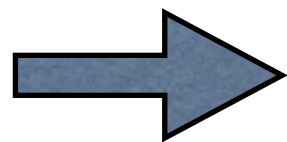
$$X_4 = 1$$



$$X_5 = 0$$

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 $L(\theta)$

Thus, be careful

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

We will discuss sufficient condition later.

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 will 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

θ : should be non-negative and be less 1

σ : should be non-negative

But the optimization did not enforce the constraint

yes, we are lucky

Constrained optimization

General case

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

Method of Lagrange multipliers

Construct the following function (Lagrangian)

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

Lagrange multiplier

Set derivative to zero

$$\frac{\partial L(x, \lambda)}{\partial x} = f'(x) + \lambda g'(x) = 0$$

Solve x in terms of λ

$$x = h(\lambda)$$

Substitute into constraint, solve λ , then x

$$g(h(\lambda)) = 0$$

Ex: roll a dice



Model

Probability of seeing the number k between 1 and 6

$$P(X = k) = \theta_k$$

Observations

$$\mathcal{D} = \{x_1, x_2, \dots, x_N\} \quad x_n \in \{1, 2, \dots, 6\}$$

Likelihood

$$L(\boldsymbol{\theta}) = \prod_{n=1}^N P(X = x_n) = \prod_{k=1}^6 \theta_k^{n_k}$$

of times k appear in observations

Optimization

Objective function (log-likelihood)

$$\max \sum_k n_k \log \theta_k$$

constraints

$$\sum_k \theta_k = 1 \quad \theta_k \geq 0$$

Lagrangian (ignoring the nonnegative constraint)

$$L(\boldsymbol{\theta}, \lambda) = \sum_k n_k \log \theta_k + \lambda \left(\sum_k \theta_k - 1 \right)$$

Finding both multiplier and the parameters

Derivatives

$$\frac{\partial L(\boldsymbol{\theta}, \lambda)}{\partial \theta_k} = \frac{n_k}{\theta_k} + \lambda$$

Setting them to zero

$$\theta_k = -\frac{1}{\lambda} n_k$$

Solving the multiplier by using the constraint

$$\sum_k \theta_k = -\frac{1}{\lambda} \sum_k n_k = 1 \rightarrow \lambda = -\sum_k n_k$$

Finally,

$$\theta_k = \frac{n_k}{\sum_k n_k}$$

Intuition:
proportional to #
of occurrences in
observations

Multiple constraints can be handled similarly

$$\begin{array}{ll}\min & f(x) \\ \text{s.t.} & g_1(x) = 0 \\ & g_2(x) = 0 \\ & g_3(x) = 0\end{array}$$

Each constraint gets a multiplier

$$L(\boldsymbol{\lambda}, x) = f(x) + \lambda_1 g_1(x) + \lambda_2 g_2(x) + \lambda_3 g_3(x)$$

and use the same stationary point condition

find all multipliers, then the variable 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.

Convex optimization

Popular tools in many areas, including machine learning

Computationally tractable: as efficient as “linear programming”

Global optimal: no worry of getting not-so-good solutions

Local vs. global optimal

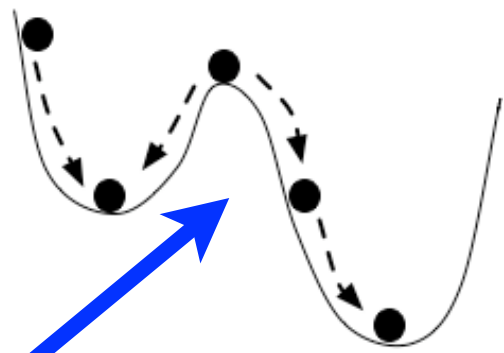
For general objective functions $f(x)$

We get local optimum

There are special types of functions

where 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

Convex functions

Definition

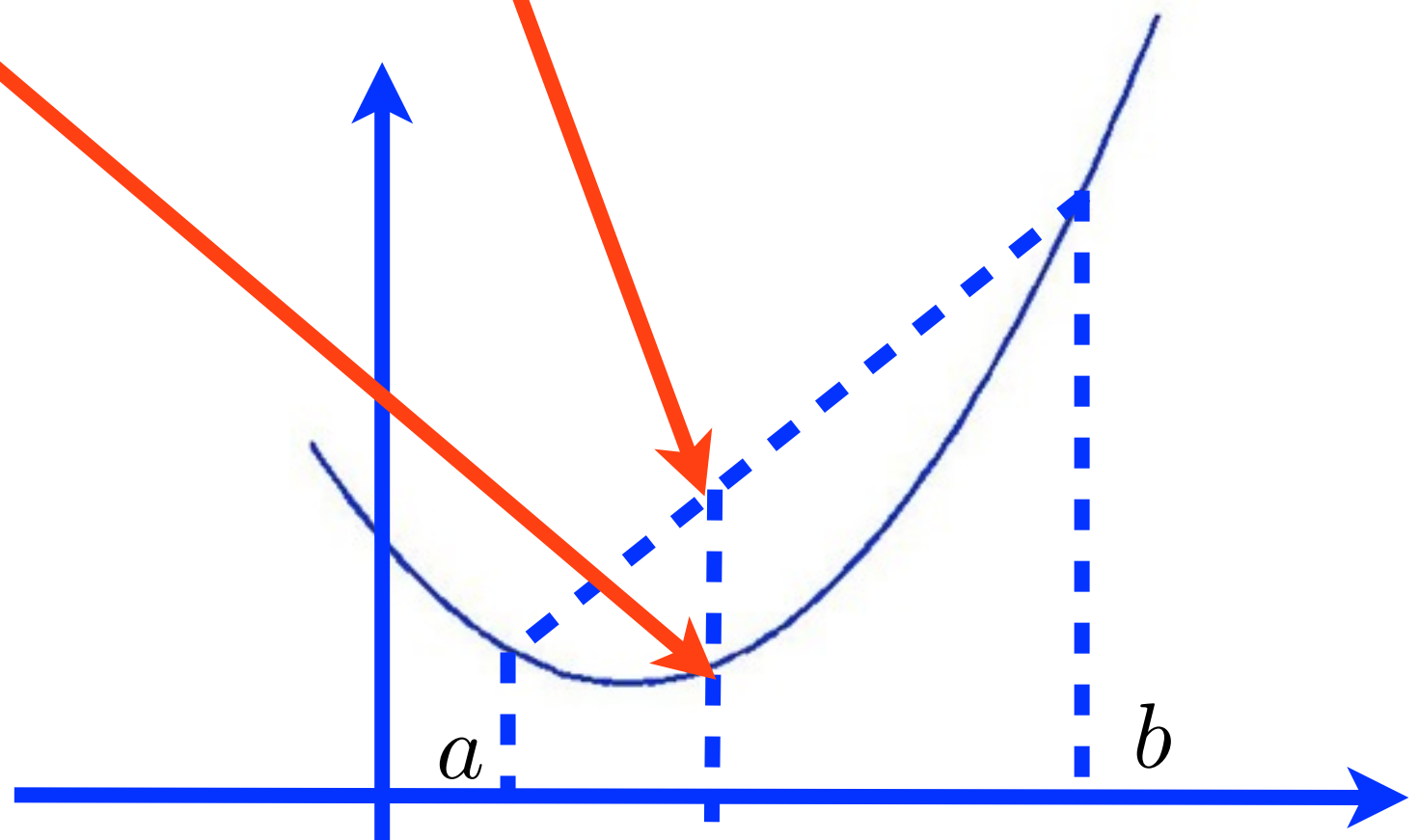
A function $f(x)$ is convex if

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

for

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

Graphically,



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

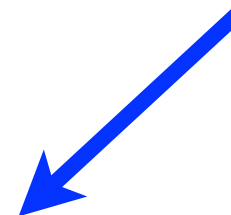
$$f''(x) \geq 0$$

Examples

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

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

**We will in future
lecture exploit this
property**



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?

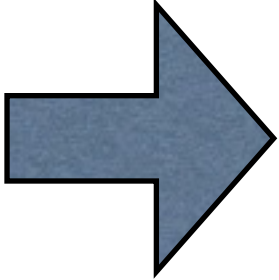
Second-order derivative becomes Hessian matrix

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

Convexity for multivariate function

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

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

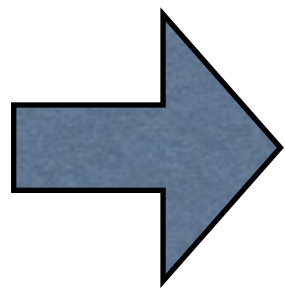

$$H = \begin{bmatrix} \frac{2}{x_2} & -\frac{2x_1}{x_2^2} \\ -\frac{2x_1}{x_2^2} & \frac{2x_1^2}{x_2^3} \end{bmatrix} = \frac{2}{x_2^3} \begin{bmatrix} x_2^2 & -x_1x_2 \\ -x_1x_2 & x_1^2 \end{bmatrix}$$

Verify that the Hessian is positive definite

Assume x_2 is positive, then

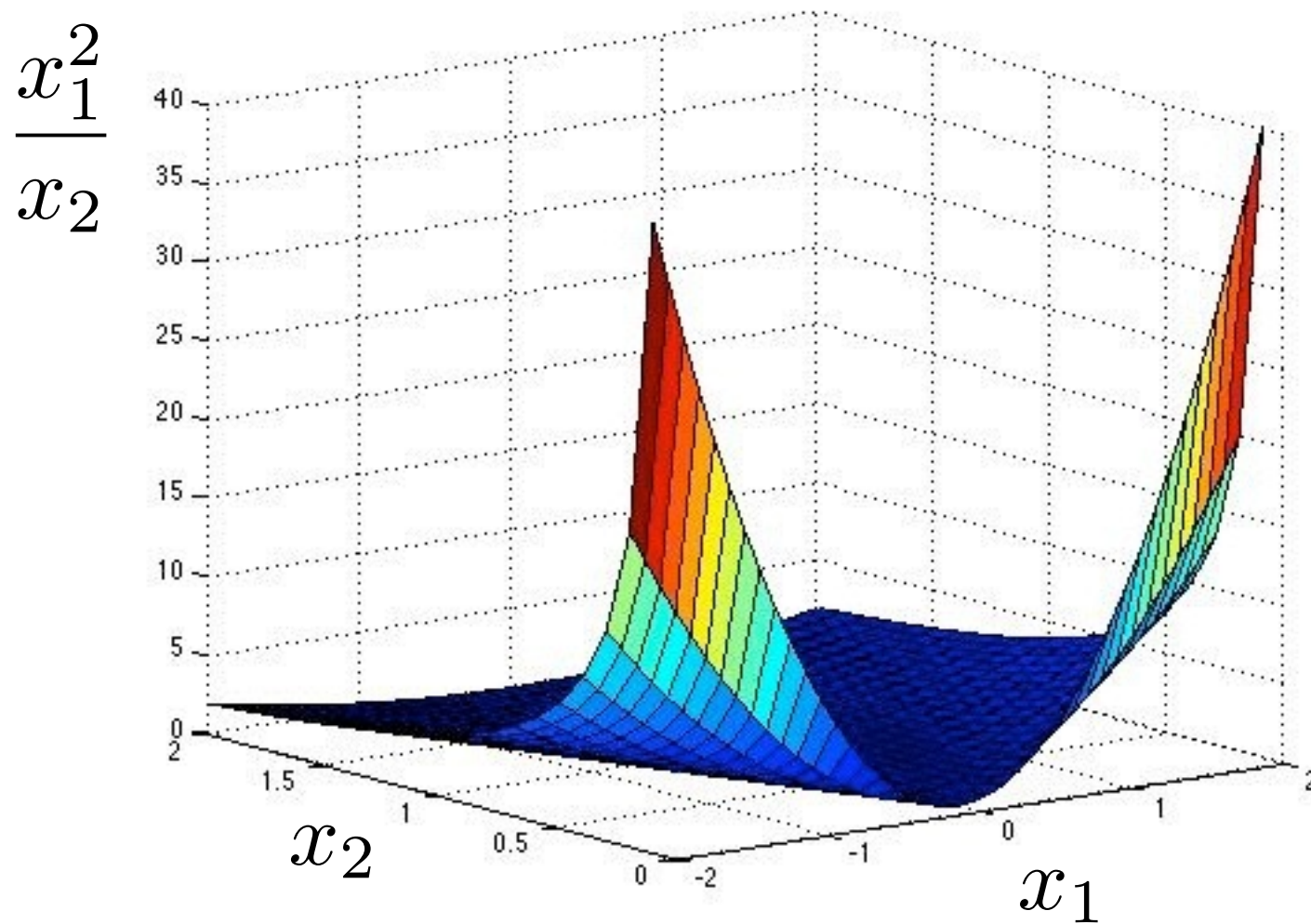
For any vector

$$\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}$$



$$\begin{aligned} \mathbf{v}^T \mathbf{H} \mathbf{v} &= \mathbf{v}^T \frac{2}{x_2^3} \begin{bmatrix} x_2^2 & -x_1 x_2 \\ -x_1 x_2 & x_1^2 \end{bmatrix} \mathbf{v} \\ &= \frac{2}{x_2^3} (a^2 x_2^2 - 2abx_1 x_2 + b^2 x_1^2) \\ &= \frac{2}{x_2^3} (ax_2 - bx_1)^2 \geq 0 \end{aligned}$$

What does this function look like?



Slightly complicated example

Take-home exercise

Verify the following function

$$f(\boldsymbol{w}) = \log \left(1 + e^{\sum_d w_d x_d} \right)$$

is convex in

$$\boldsymbol{w} = (w_1, w_2, \dots, w_D)^T$$

Why convex function?

if $f(x)$ is convex

then the local optimal

$$\min f(x)$$

is also global optimal

This generalizes to constrained optimization

if the constraint

$$g(x) \leq 0$$

define a convex set, namely, for $0 \leq \lambda \leq 1$

$$g(\mathbf{a}) \leq 0, g(\mathbf{b}) \leq 0 \rightarrow g(\lambda \mathbf{a} + (1 - \lambda) \mathbf{b}) \leq 0$$

Convex set

Take-home exercise

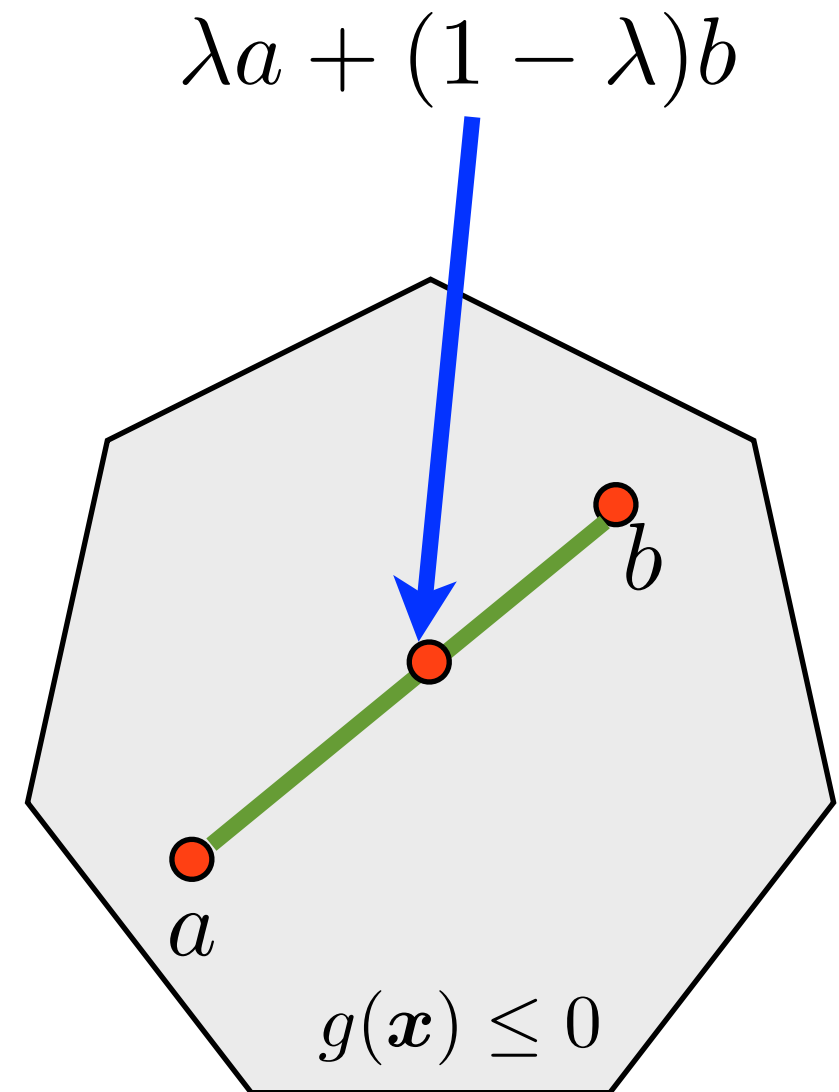
If $g(x)$ is convex

then

$$g(x) \leq 0$$

defines a convex set

graphically,



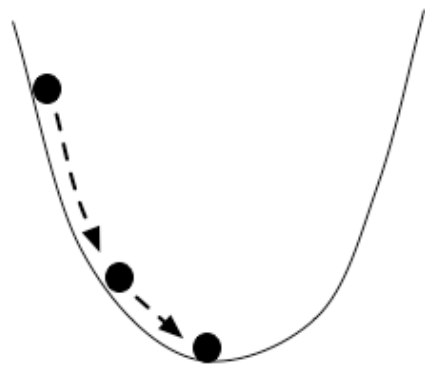
Local vs. global optimal

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

