Università
della
Svizzera
italiana

Faculty
of Informatics

Institute of
Computational
Science
ICS

A1: Introduction to Optimization theory

very short review

CECAM workshop, Mainz, 2019

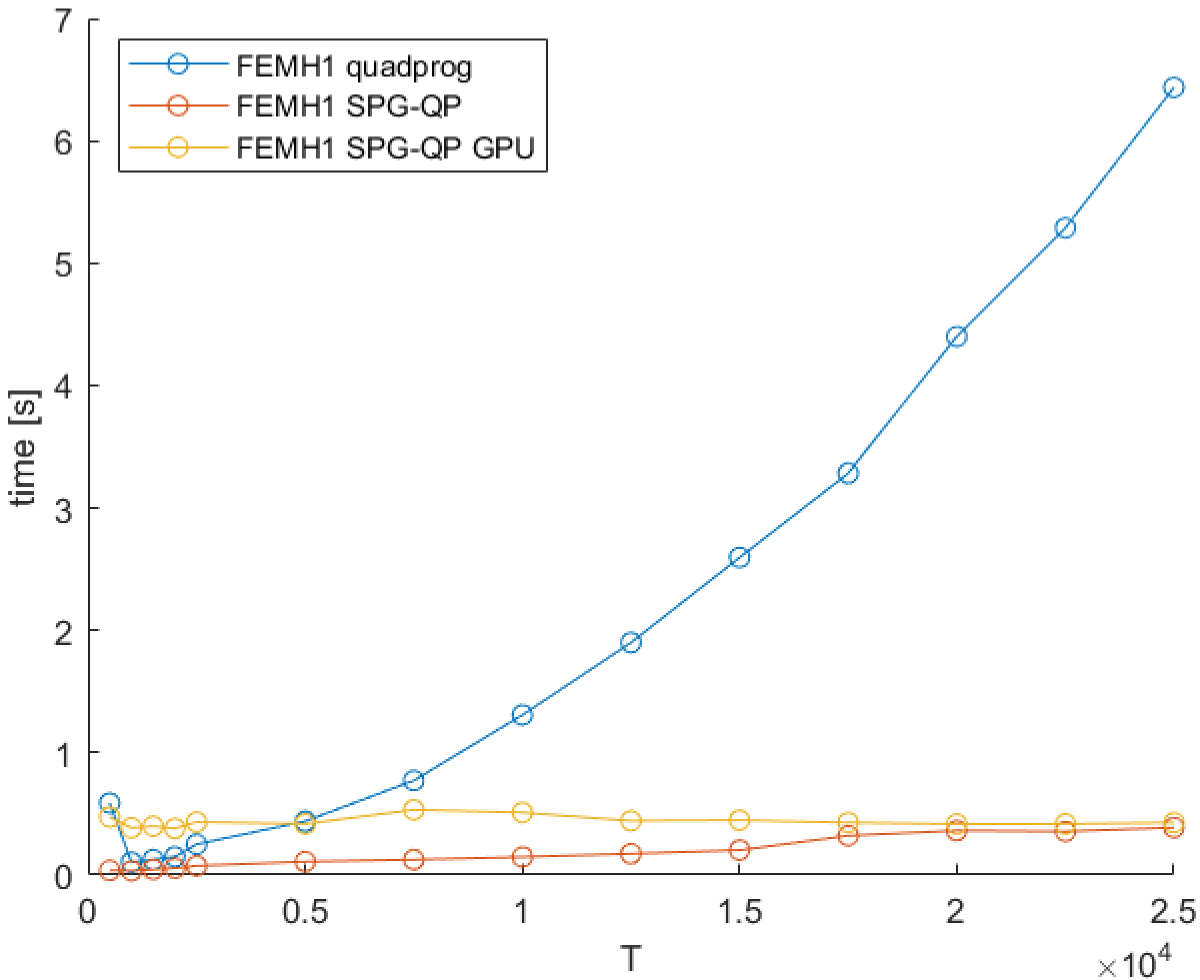
Illia Horenko, Lukas Pospisil

Algorithm matters

solve QP problem

$$\Gamma^* = \arg\min_{\gamma} \gamma^T H \gamma + g^T \gamma$$

subject to $B\gamma = c$ and $\gamma \geq 0$

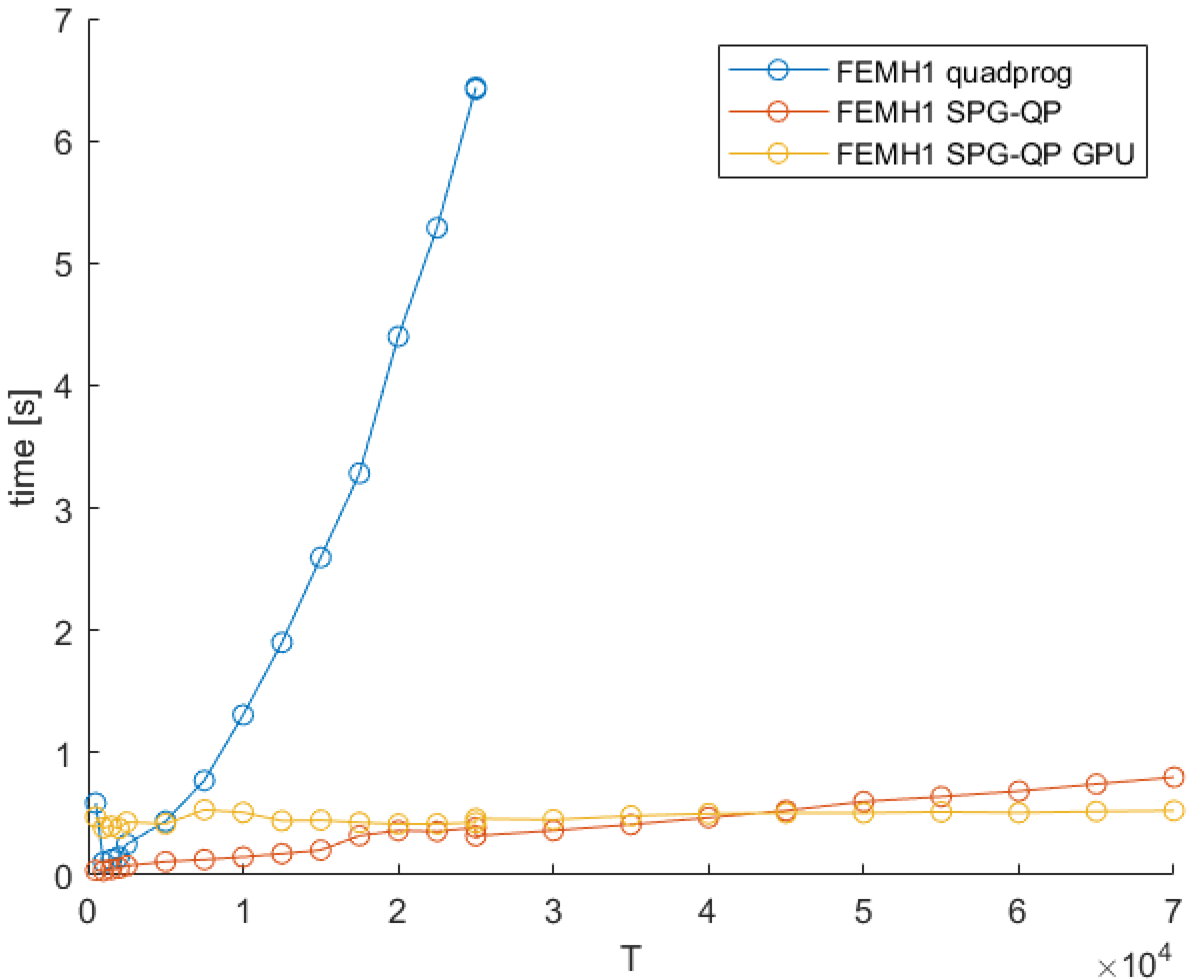


Algorithm matters

solve QP problem

$$\Gamma^* = \arg\min_{\gamma} \gamma^T H \gamma + g^T \gamma$$

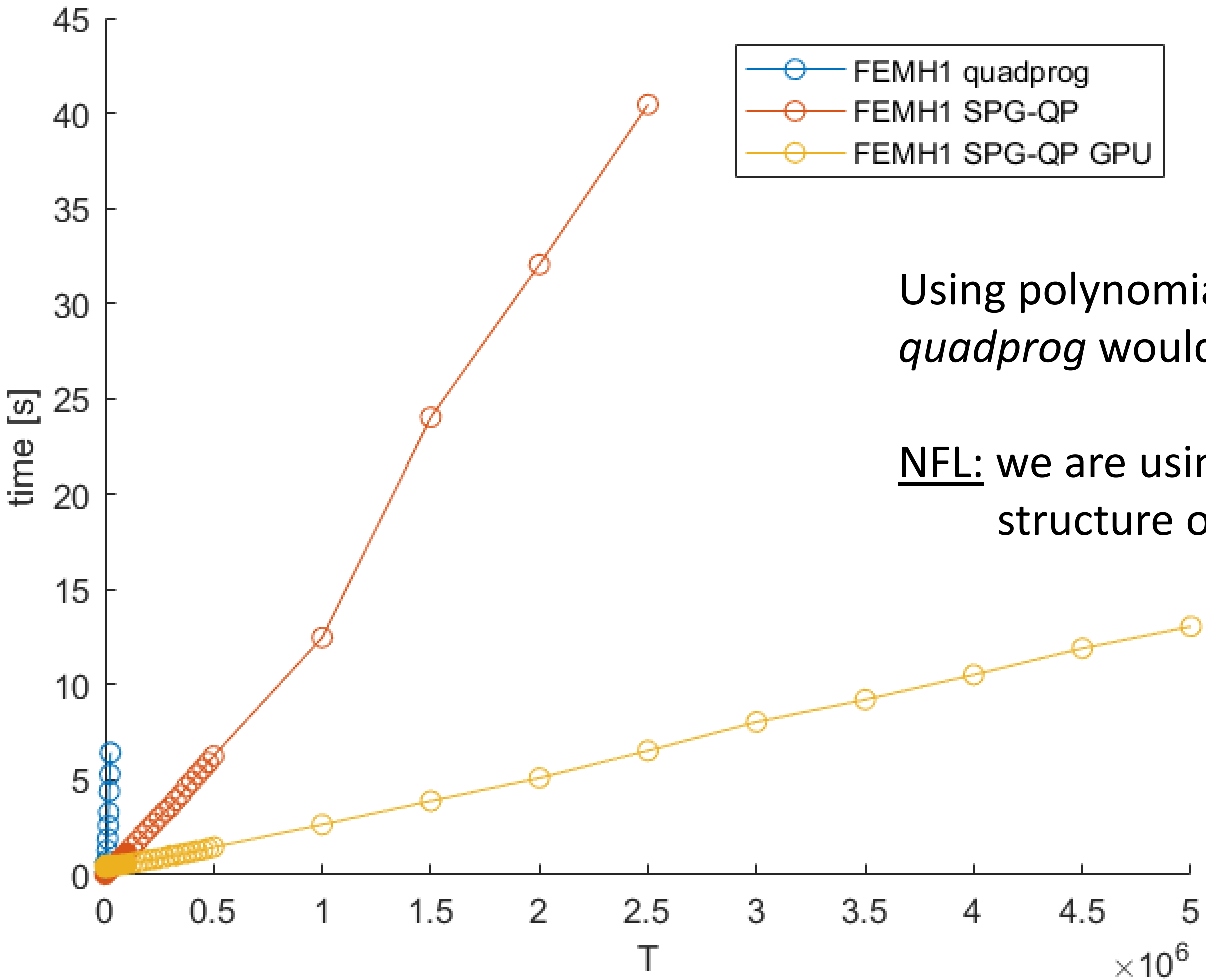
$$\text{subject to} \quad B\gamma = c \text{ and } \gamma \geq 0$$



Algorithm matters

solve QP problem

$$\Gamma^* = \arg\min_{\gamma} \gamma^T H \gamma + g^T \gamma$$
$$\text{subject to} \quad B\gamma = c \text{ and } \gamma \geq 0$$



Using polynomial regression (assuming O(T^2))
quadprog would solve T=5*10^6 in 2,5 days

NFL: we are using the specific structure of the problem

Outline

1.) Unconstrained optimization problem
(some theory and definitions)

2.) Equality constrained optimization problem
(Lagrange function, Lagrange multipliers)

3.) Inequality constrained optimization problem
(Karush-Kuhn-Tucker)

1.) Unconstrained optimization problems

Optimization problem

What is maximum on Ω ? $(\text{of } f : \Omega \to \mathbb{R}) \qquad \Omega \subset \mathcal{V}$

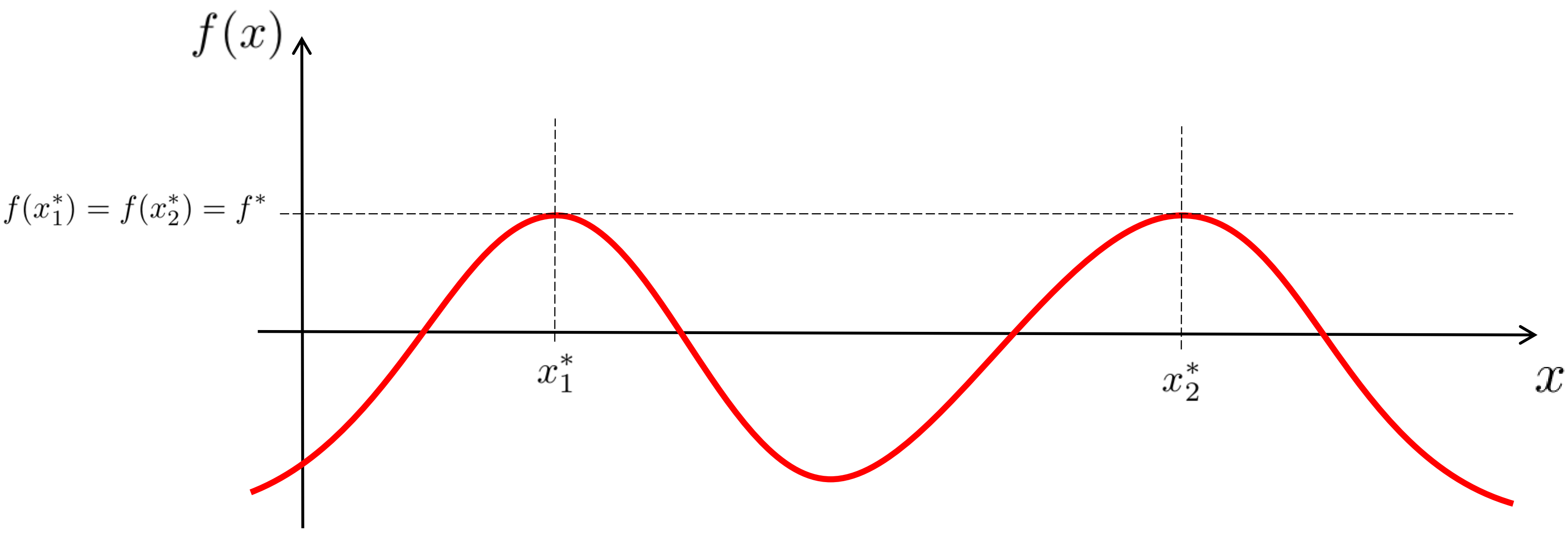
$$f^* = \max_{x \in \Omega} f(x)$$

$$\text{Find } f^* \in \{f(x) | x \in \Omega\} \text{ such that } \forall x \in \Omega : f(x) \leq f^*$$

$$x^* = \arg\max_{x \in \Omega} f(x)$$

$$\text{Find } x^* \in \Omega \text{ such that } \forall x \in \Omega : f(x) \leq f(x^*)$$

Example:



What is minimum?

What is maximum on Ω ? $\quad$ (of $f:\Omega\to\mathbb{R}$) $\quad \Omega\subset\mathcal{V}$

$$f^* = \max_{x\in\Omega} f(x)$$

Find $f^*\in\{f(x)|x\in\Omega\}$ such that $\forall x\in\Omega: f(x)\le f^*$

$$x^* = \arg\max_{x\in\Omega} f(x)$$

Find $x^*\in\Omega$ such that $\forall x\in\Omega: f(x)\le f(x^*)$

$$\updownarrow$$

What is minimum on Ω? $\quad$ (of $f:\Omega\to\mathbb{R}$) $\quad \Omega\subset\mathcal{V}$

$$f^* = \min_{x\in\Omega} f(x)$$

Find $f^*\in\{f(x)|x\in\Omega\}$ such that $\forall x\in\Omega: f(x)\ge f^*$

$$x^* = \arg\min_{x\in\Omega} f(x)$$

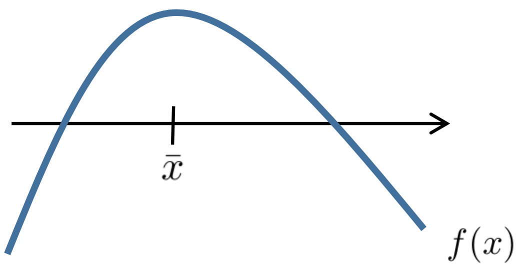
Find $x^*\in\Omega$ such that $\forall x\in\Omega: f(x)\ge f(x^*)$

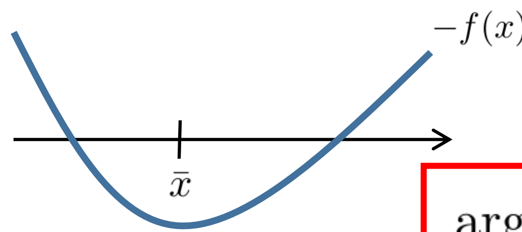
What is minimum?

What is maximum on Ω ? $(\text{of } f : \Omega \to \mathbb{R}) \qquad \Omega \subset \mathcal{V}$

$$f^* = \max_{x \in \Omega} f(x) \qquad \text{Find } f^* \in \{f(x) | x \in \Omega\} \text{ such that } \forall x \in \Omega : f(x) \leq f^*$$

$$x^* = \arg\max_{x \in \Omega} f(x) \qquad \text{Find } x^* \in \Omega \text{ such that } \forall x \in \Omega : f(x) \leq f(x^*)$$





$$\boxed{\arg\max f(x) = \arg\min -f(x)}$$

$$\max f(x) = -\min(-f(x))$$

What is minimum on Ω? $(\text{of } f : \Omega \to \mathbb{R}) \qquad \Omega \subset \mathcal{V}$

$$f^* = \min_{x \in \Omega} f(x) \qquad \text{Find } f^* \in \{f(x) | x \in \Omega\} \text{ such that } \forall x \in \Omega : f(x) \geq f^*$$

$$x^* = \arg\min_{x \in \Omega} f(x) \qquad \text{Find } x^* \in \Omega \text{ such that } \forall x \in \Omega : f(x) \geq f(x^*)$$

Using the definition

Exercise:

using the definition prove that:

$$\forall \alpha > 0 \forall \beta \in \mathbb{R} : \arg\min_{x \in \Omega} f(x) = \arg\min_{x \in \Omega} (\alpha f(x) + \beta)$$

$$\text{if } \forall x \in \Omega : f(x) \geq 0 \text{ then } \quad \arg\min_{x \in \Omega} f(x) = \arg\min_{x \in \Omega} f^2(x)$$

$$\text{if } \forall x \in \Omega : f(x) > 0 \text{ then } \quad \arg\max_{x \in \Omega} f(x) = \arg\max_{x \in \Omega} \log(f(x))$$

Reminder:

$$x^* = \arg\max_{x \in \Omega} f(x) \qquad \text{Find } x^* \in \Omega \text{ such that } \forall x \in \Omega : f(x) \leq f(x^*)$$

$$x^* = \arg\min_{x \in \Omega} f(x) \qquad \text{Find } x^* \in \Omega \text{ such that } \forall x \in \Omega : f(x) \geq f(x^*)$$

Function f is called *monotonically increasing* on Ω if $\forall x, y \in \Omega : x < y \Leftrightarrow f(x) < f(y)$

Solvability issues

$$\min_{x \in \Omega} f(x)$$

feasible set (defined by constraints)

- non-empty?
- convex?
- closed?
- bounded?
- integer?

...

objective function ("cost" function)

- continuous?
- defined everywhere in feasible set?
- convex? strictly/quasi convex?
- bounded from below?
- differentiable?

...

Properties define the solvability of the optimization problem.

... and many problems are still open!

What is optimization problem?

$$\min_{x \in \Omega} f(x)$$

<u>feasible set</u> (defined by constraints)

$$\Omega = \mathbb{R}^n$$

(unconstrained)

$$\Omega \subset \mathbb{R}^n$$

$$\begin{aligned}
\Omega &= \{x \in \mathbb{R}^n : x \geq 0\} \\
\Omega &= \{x \in \mathbb{R}^n : 0 \leq x \leq 1\} \\
\Omega &= \{x \in \mathbb{R}^n : Ax \leq b\} \\
\Omega &= \{x \in \mathbb{R}^n : Ax = b\} \\
\Omega &= \{x \in \mathbb{R}^n : g(x) \leq 0\}
\end{aligned}$$

equality constraints

inequality constraints

<u>objective function</u> ("cost" function)

$$f : \mathbb{R}^n \to \mathbb{R}$$

Basic theorem of solvability

$$\min_{x \in \Omega} f(x)$$

(Weierstrass extreme value theorem.)

If f is a real-valued continuous function on a non-empty compact (i.e. bounded and closed) domain Ω, then there exists $x \in \Omega$ such that $f(x) \geq f(y)$ for all $y \in \Omega$.

(Bounded set.)

The set Ω is bounded, if there exists a real constant $M > 0$ such that

$$\forall x \in \Omega : \|x\| \leq M \ .$$

(Closed set.)

The set Ω is closed if for any sequence of points $\{x^k\}$ in Ω, all limit points of this sequence belong to Ω.

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x) \quad (P)$$

First order necessary condition for unconstrained problem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at point $x^* \in \mathbb{R}^n$.
If x^* is a solution of (P), then $\nabla f(x^*) = 0$.

Second order necessary condition for unconstrained problem

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable.
If x^* is a solution of (P), then $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice differentiable, $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*) \succ 0$.
Then x^* is *local* solution of (P).

Let $f : \mathbb{R}^n \to \mathbb{R}$ is strictly convex function.
Then if (P) has solution, then this solution is unique.

Example: (Linear) Regression

Input

$t = 0, 1, 2, \ldots$

Real system

$x(t) = \alpha t + \beta,$
$\theta := [\alpha, \beta] \in \mathbb{R}^2$ given

Output

$x(0), x(1), x(2), \ldots$

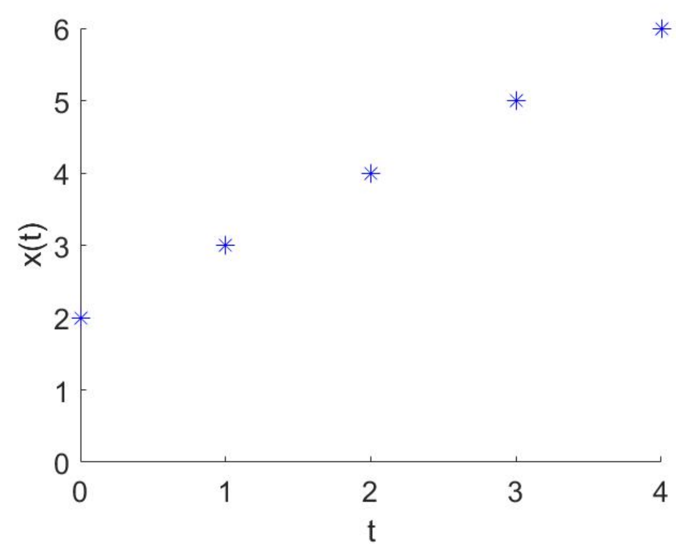
Example:

Uniform motion describes an object that is moving in a specific direction at a constant speed.

$$s = v \cdot t + s_0$$

$$v = 1, s_0 = 2$$

$$x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6$$



Example: (Linear) Regression

Input

$t = 0, 1, 2, \dots$

Real system

$x(t) = \alpha t + \beta,$

$\theta := [\alpha, \beta] \in \mathbb{R}^2$ ~~given~~

unknown

Output

$x(0), x(1), x(2), \dots$

given

Example:

$$x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6$$

Example: (Linear) Regression

Input

$t = 0, 1, 2, \dots$

$\rightarrow$

Real system

$x(t) = \alpha t + \beta,$

$\theta := [\alpha, \beta] \in \mathbb{R}^2$ ~~given~~

unknown

$\rightarrow$

Output

$x(0), x(1), x(2), \dots$

given

Example:

$$x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6$$

$$\left.\begin{array}{rclcrcl}
x(0) & = & \alpha \cdot 0 + \beta & \Rightarrow & 2 & = & \beta \\
x(1) & = & \alpha \cdot 1 + \beta & \Rightarrow & 3 & = & 1\alpha + \beta \\
x(2) & = & \alpha \cdot 2 + \beta & \Rightarrow & 4 & = & 2\alpha + \beta \\
x(3) & = & \alpha \cdot 3 + \beta & \Rightarrow & 5 & = & 3\alpha + \beta \\
x(4) & = & \alpha \cdot 4 + \beta & \Rightarrow & 6 & = & 4\alpha + \beta
\end{array}\right\} \quad B\theta = c$$

(system of linear equations)

Example: (Linear) Regression

Input

$t = 0, 1, 2, \ldots$

Real system

$x(t) = \alpha t + \beta,$

$\theta := [\alpha, \beta] \in \mathbb{R}^2$ ~~given~~

unknown

Output

$x(0), x(1), x(2), \ldots$

given

Example:

$$x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6$$

$$\alpha = 1, \beta = 2$$

satisfies all equations

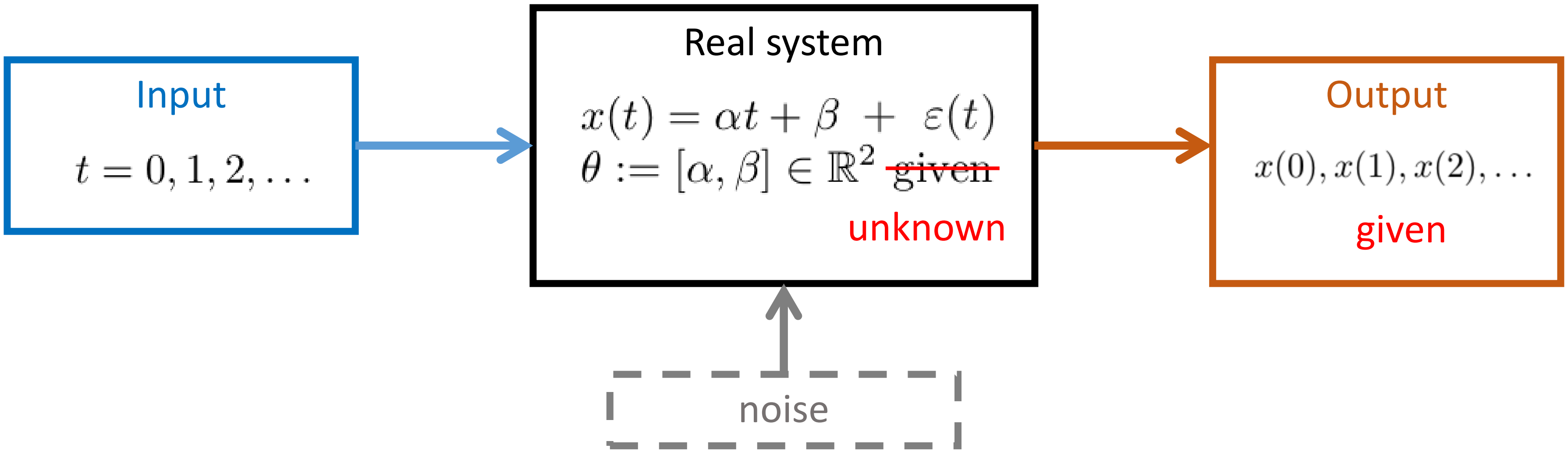
$$\begin{array}{rcccrcl}
x(0) & = & \alpha \cdot 0 + \beta & \Rightarrow & 2 & = & \beta \\
x(1) & = & \alpha \cdot 1 + \beta & \Rightarrow & 3 & = & 1\alpha + \beta \\
x(2) & = & \alpha \cdot 2 + \beta & \Rightarrow & 4 & = & 2\alpha + \beta \\
x(3) & = & \alpha \cdot 3 + \beta & \Rightarrow & 5 & = & 3\alpha + \beta \\
x(4) & = & \alpha \cdot 4 + \beta & \Rightarrow & 6 & = & 4\alpha + \beta
\end{array}$$

redundant

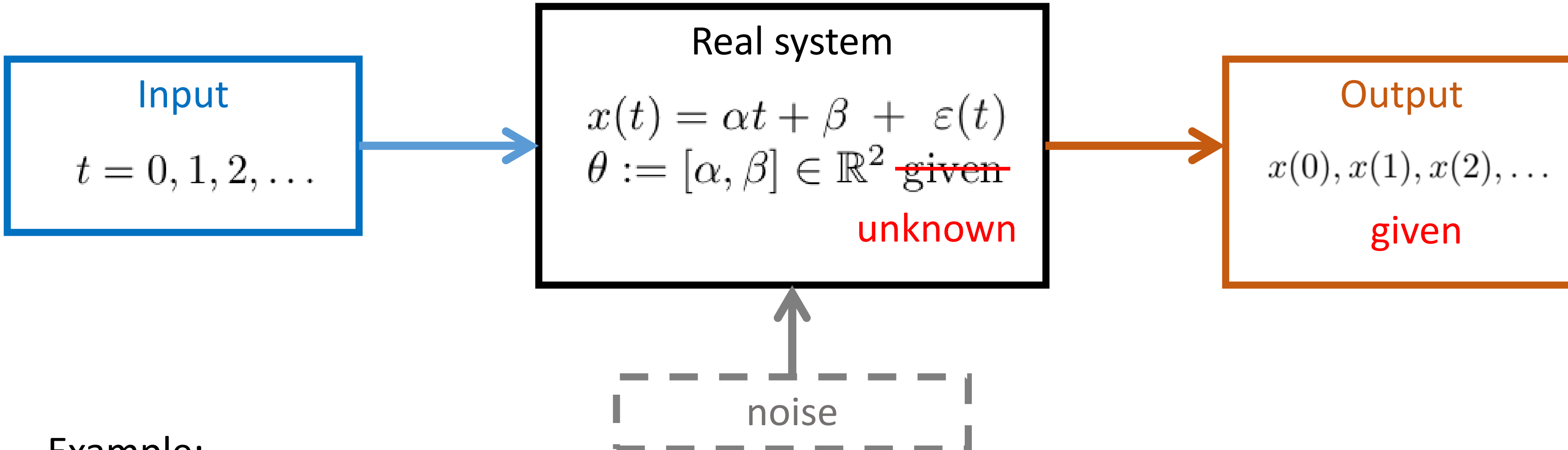
$$B\theta = c$$

(system of linear equations)

(Linear) Regression



(Linear) Regression



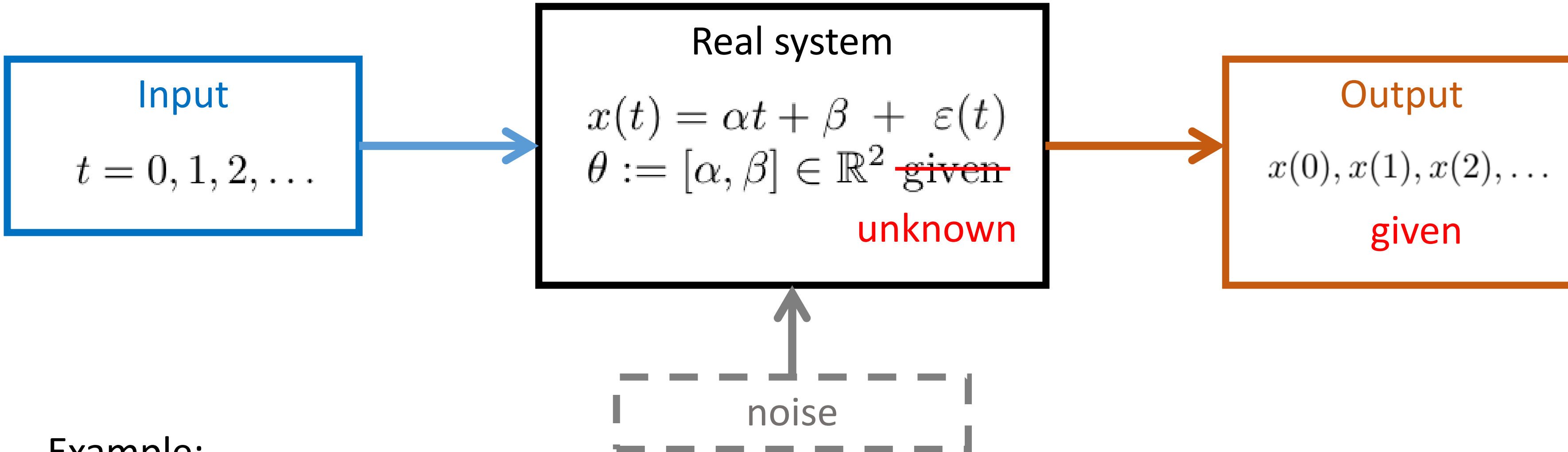
Example:

$$\cancel{x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6}$$

$$x(0) = 1.9865, x(1) = 3.0303, x(2) = 4.0073, x(3) = 4.9994, x(4) = 6.0071$$

$$\begin{aligned}
x(0) &= \alpha \cdot 0 + \beta & \Rightarrow \quad 1.9865 &= \beta \\
x(1) &= \alpha \cdot 1 + \beta & \Rightarrow \quad 3.0303 &= 1\alpha + \beta \\
x(2) &= \alpha \cdot 2 + \beta & \Rightarrow \quad 4.0073 &= 2\alpha + \beta \\
x(3) &= \alpha \cdot 3 + \beta & \Rightarrow \quad 4.9994 &= 3\alpha + \beta \\
x(4) &= \alpha \cdot 4 + \beta & \Rightarrow \quad 6.0071 &= 4\alpha + \beta
\end{aligned}$$

(Linear) Regression



Example:

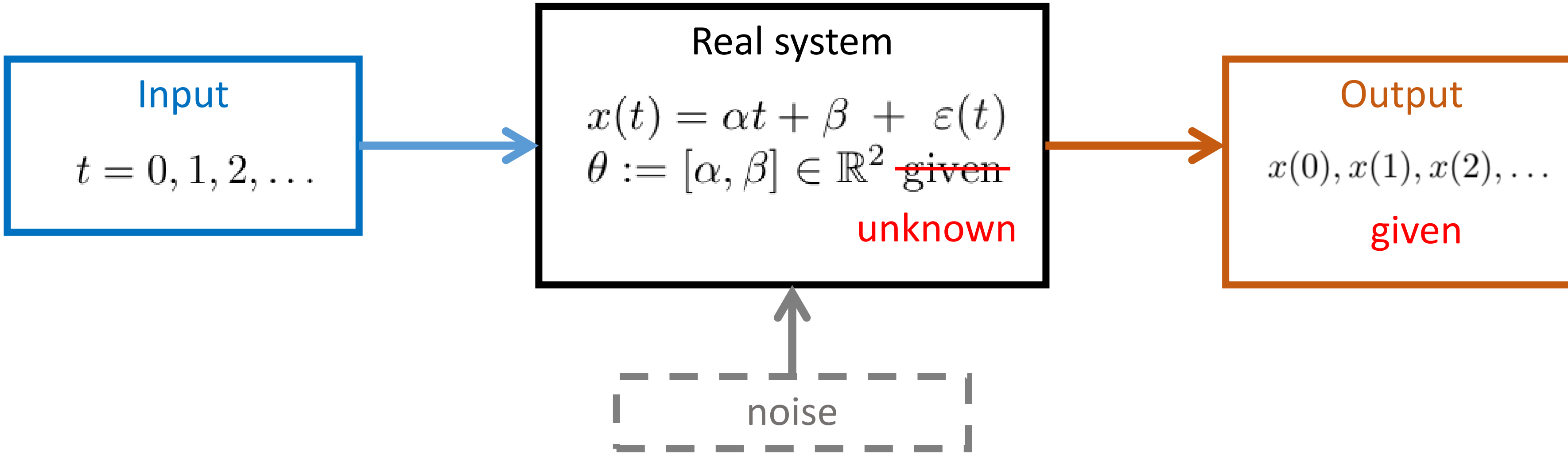
$\cancel{x(0) = 2, x(1) = 3, x(2) = 4, x(3) = 5, x(4) = 6}$

$$x(0) = 1.9865, x(1) = 3.0303, x(2) = 4.0073, x(3) = 4.9994, x(4) = 6.0071$$

$$\begin{array}{lclcrcl}
x(0) & = & \alpha \cdot 0 + \beta & \Rightarrow & 1.9865 & = & \beta \\
x(1) & = & \alpha \cdot 1 + \beta & \Rightarrow & 3.0303 & = & 1\alpha + \beta \\
x(2) & = & \alpha \cdot 2 + \beta & \Rightarrow & 4.0073 & = & 2\alpha + \beta \\
x(3) & = & \alpha \cdot 3 + \beta & \Rightarrow & 4.9994 & = & 3\alpha + \beta \\
x(4) & = & \alpha \cdot 4 + \beta & \Rightarrow & 6.0071 & = & 4\alpha + \beta
\end{array}$$

does not have solution :(

(Linear) Regression

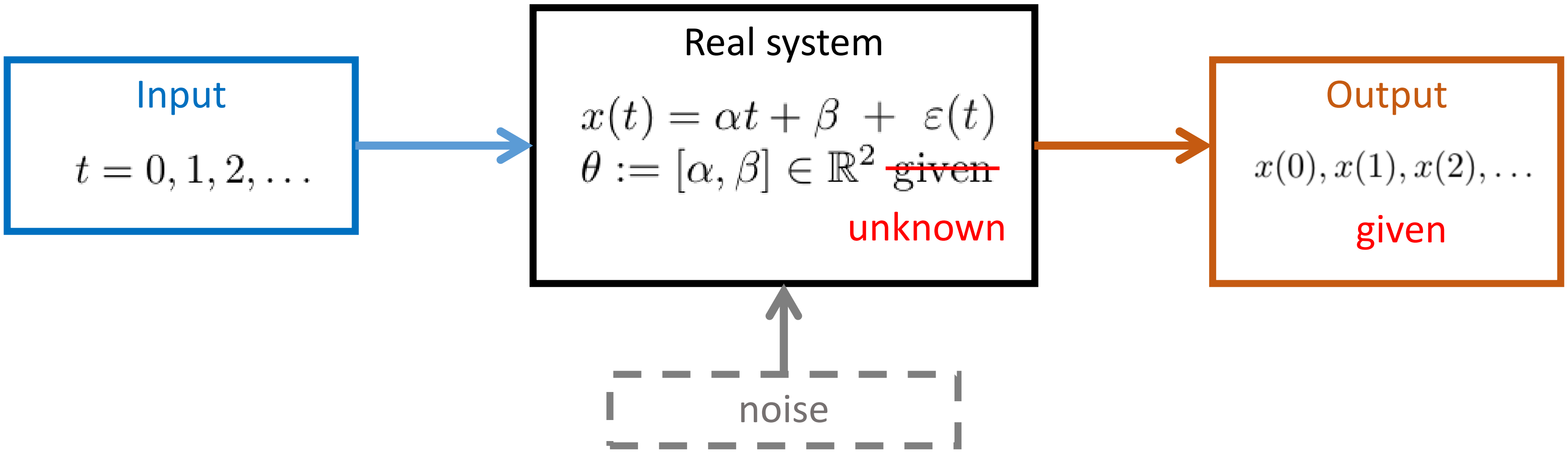


$$\theta^* = \arg\min_{\theta} \ \text{distance}\,(\ \text{data(input)}\ ,\ \text{model(input,}\theta)\ )$$

what we observed

what we "should" observe

(Linear) Regression



$$\theta^* = \arg\min_{\theta} \ \text{distance} (\ \text{data}(\text{input}) \ , \ \text{model}(\text{input},\theta) \ )$$

Example:

$t = 0, 1, \dots$	input
$x(0), x(1), x(2), \dots$	data
$m(t, \theta) := \alpha t + \beta$	model (linear)

$$\rho(x(.), m(., \theta)) := \sum_{t=0}^{T-1} \|x(t) - m(t, \theta)\|_2^2 \qquad \text{(least square error)}$$

(Linear) Regression

$$\theta^* = \arg\min \sum_{t=0}^{T-1} \|x(t) - m(t,\theta)\|_2^2 = \arg\min \sum_{t=0}^{T-1} (x(t) - m(t,\theta))^2 = \arg\min \underbrace{\|B\theta - c\|_2^2}_{=:\, f(\theta)}$$

$x_t \in \mathbb{R}$

$m(t,\theta) := \alpha t + \beta$

<u>Example:</u>

$t = 0, 1, \ldots$	input
$x(0), x(1), x(2), \ldots$	data
$m(t,\theta) := \alpha t + \beta$	model (linear)

$$\rho(x(.), m(.,\theta)) := \sum_{t=0}^{T-1} \|x(t) - m(t,\theta)\|_2^2 \qquad \text{(least square error)}$$

(Linear) Regression

$$\theta^* = \arg\min \sum_{t=0}^{T-1} \|x(t) - m(t,\theta)\|_2^2 = \arg\min \sum_{t=0}^{T-1} (x(t) - m(t,\theta))^2 = \arg\min \underbrace{\|B\theta - c\|_2^2}_{=:\, f(\theta)}$$

$x_t \in \mathbb{R}$

$m(t,\theta) := \alpha t + \beta$

Reminder:

$$\begin{aligned} x(0) &= \alpha \cdot 0 + \beta \\ x(1) &= \alpha \cdot 1 + \beta \\ x(2) &= \alpha \cdot 2 + \beta \\ x(3) &= \alpha \cdot 3 + \beta \\ x(4) &= \alpha \cdot 4 + \beta \end{aligned}$$

$$\underbrace{\begin{array}{cc} t^1 & t^0 \\ \end{array}}_{} \quad \underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} \alpha \\ \beta \end{bmatrix}}_{\theta} = \underbrace{\begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ x(4) \end{bmatrix}}_{c}$$

(Linear) Regression

$$\theta^* = \arg\min \sum_{t=0}^{T-1} \|x(t) - m(t,\theta)\|_2^2 = \arg\min \sum_{t=0}^{T-1} (x(t) - m(t,\theta))^2 = \arg\min \underbrace{\|B\theta - c\|_2^2}_{=:\, f(\theta)}$$

$x_t \in \mathbb{R}$ $\qquad$ $m(t,\theta) := \alpha t + \beta$

$$\begin{aligned} f(\theta) &= \|B\theta - c\|_2^2 = \langle B\theta - c, B\theta - c\rangle \\ &= \theta^T B^T B\theta - 2c^T B\theta + c^T c \end{aligned}$$

$$\begin{aligned} \nabla f(\theta) &= 2B^T B\theta - 2B^T c \\ \nabla^2 f(\theta) &= 2B^T B \end{aligned}$$

$$\nabla f(\theta) = 0 \quad \Leftrightarrow \quad (B^T B)\theta = B^T c$$

$$\forall y \in \mathbb{R}^2 : y^T B^T B y = \|By\|^2 \geq 0 \quad \Rightarrow \quad f(\theta) \text{ is convex.}$$

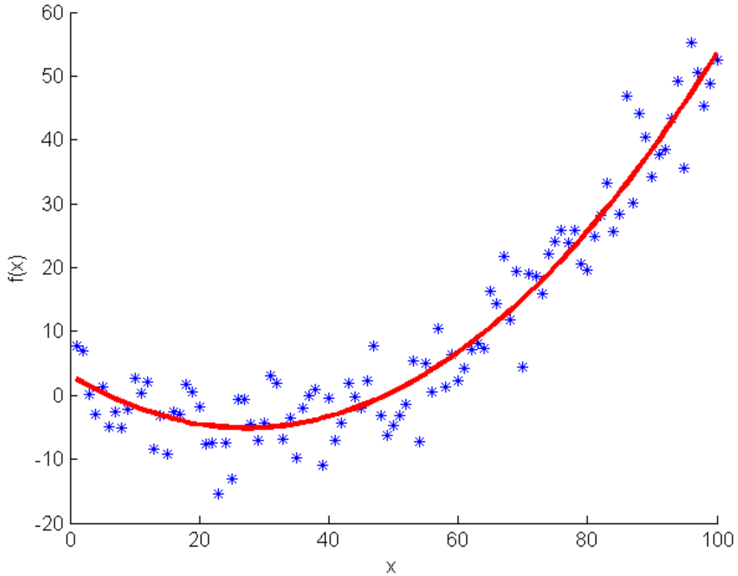
this system has always solution...

(Polynomial) Regression

degree of polynomial model

$$m(t,\theta) := \sum_{k=0}^{P} \theta_k t^k, \quad t = 0, 1, \ldots, T-1$$

$$\theta := \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \vdots \\ \theta_P \end{bmatrix} \in \mathbb{R}^{P+1} \qquad c := \begin{bmatrix} x(0) \\ x(1) \\ x(2) \\ x(3) \\ \vdots \\ x(T-1) \end{bmatrix} \in \mathbb{R}^{T-1}$$

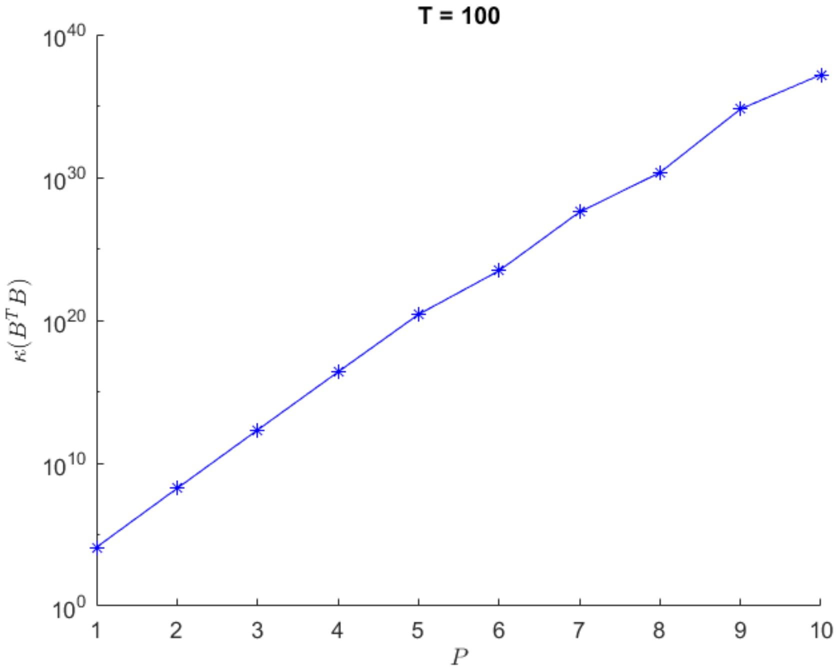


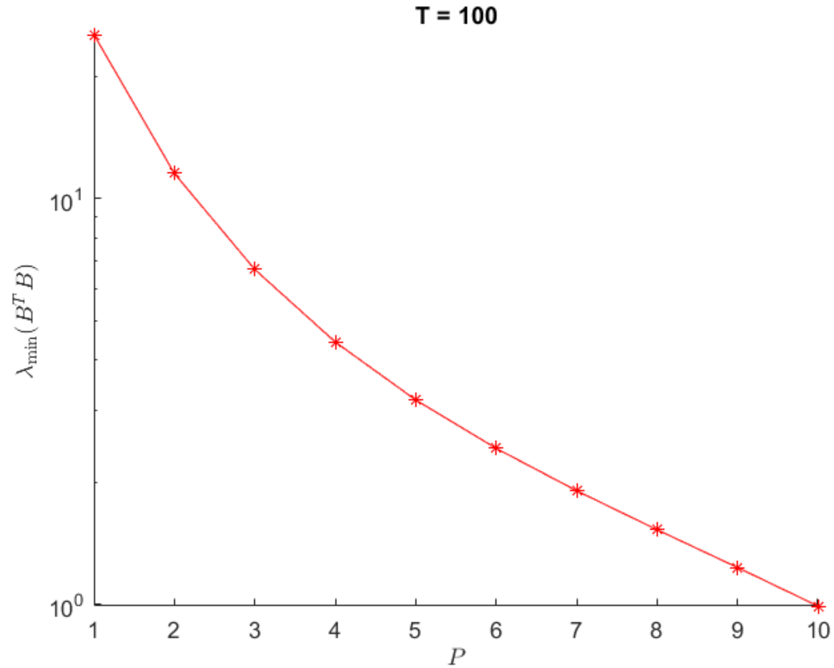
$$B := \begin{bmatrix} 1 & 0^1 & 0^2 & \ldots & 0^P \\ 1 & 1^1 & 1^2 & \ldots & 1^P \\ 1 & 2^1 & 2^2 & \ldots & 2^P \\ 1 & 3^1 & 3^2 & \ldots & 3^P \\ \vdots & & & & \\ 1 & (T-1)^1 & (T-1)^2 & \ldots & (T-1)^P \end{bmatrix} \in \mathbb{R}^{T-1,P+1}$$

solve:

$$(B^T B)\theta = B^T c$$

(Polynomial) Regression





$$B := \begin{bmatrix} 1 & 0^1 & 0^2 & \dots & 0^P \\ 1 & 1^1 & 1^2 & \dots & 1^P \\ 1 & 2^1 & 2^2 & \dots & 2^P \\ 1 & 3^1 & 3^2 & \dots & 3^P \\ \vdots & & & & \\ 1 & (T-1)^1 & (T-1)^2 & \dots & (T-1)^P \end{bmatrix} \in \mathbb{R}^{T-1,P+1}$$

solve:

$$(B^T B)\theta = B^T c$$

2.) Equality constrained optimization problem

Optimization with equality constraints

$$x^* = \arg\min_{x \in \Omega} f(x), \qquad \Omega := \{x \in \mathbb{R}^n : h_i(x) = 0, i = 1, \ldots, m\}$$

$f : \mathbb{R}^n \to \mathbb{R}$	is objective function,
$\Omega \subset \mathbb{R}^n$	is feasible set,
$h_i : \mathbb{R}^n \to \mathbb{R}$	is i-th equality constraint (m constraints).

Necessary optimality conditions:

Let x be an optimality point. Then there exist $\lambda \in \mathbb{R}^m$ such that

$$\begin{aligned} \nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) &&= 0, \\ \nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &&= 0, \quad i = 1, \ldots, m, \end{aligned}$$

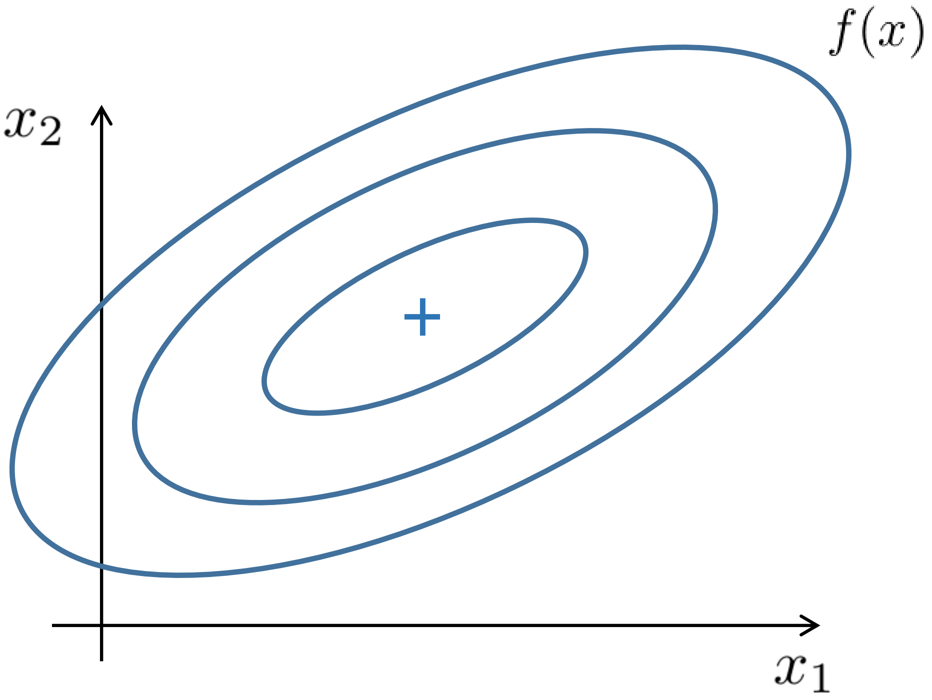
where $L : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is Lagrange function defined as

$$L(x,\lambda) := f(x) + \sum_{i=1}^m \lambda_i h_i(x).$$

Optimization with equality constraints

"proof"

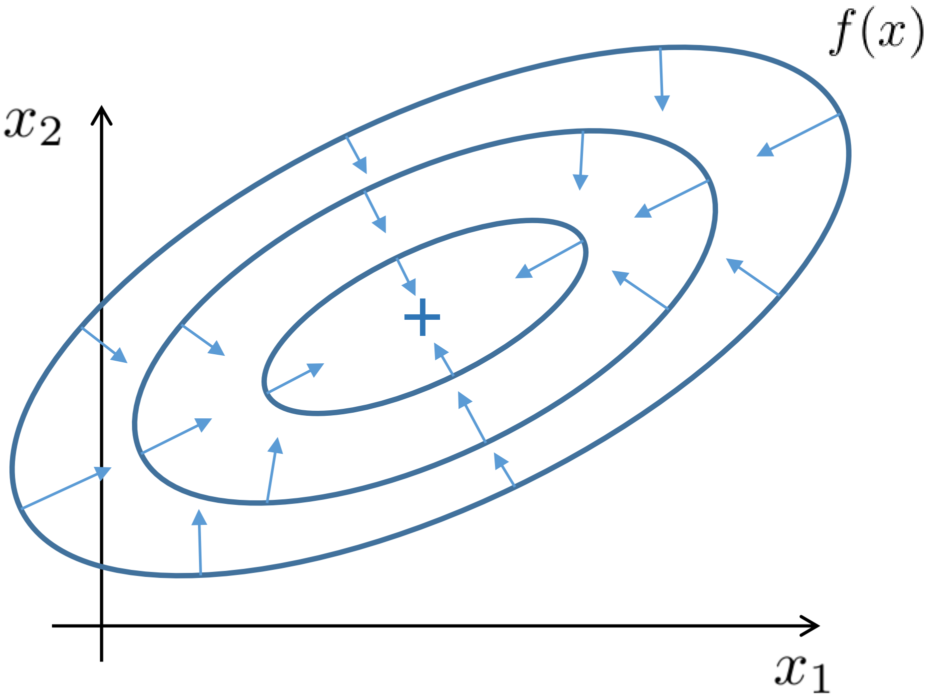
$$
\begin{aligned}
\nabla f(x) + \lambda \nabla h(x) &= 0 \\
h(x) &= 0
\end{aligned}
$$



Optimization with equality constraints

"proof"

$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$

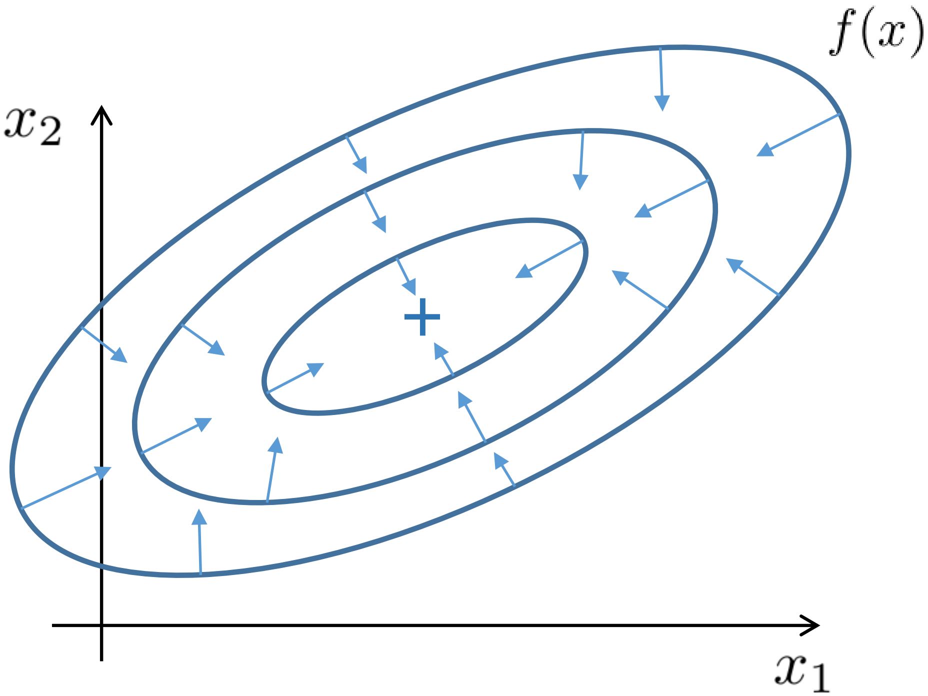


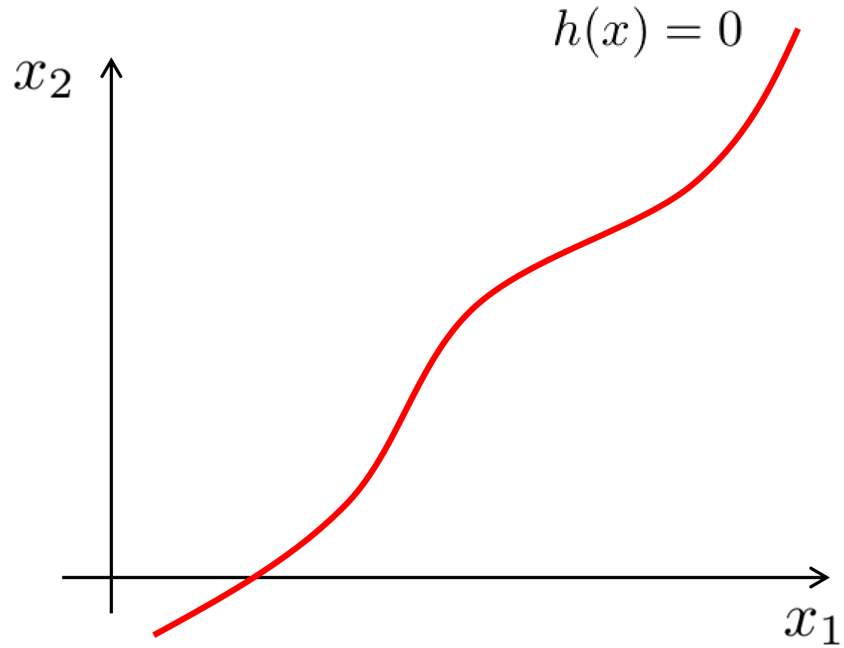
$-\nabla f(x)$ is vector of steepest descent

Optimization with equality constraints

"proof"

$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



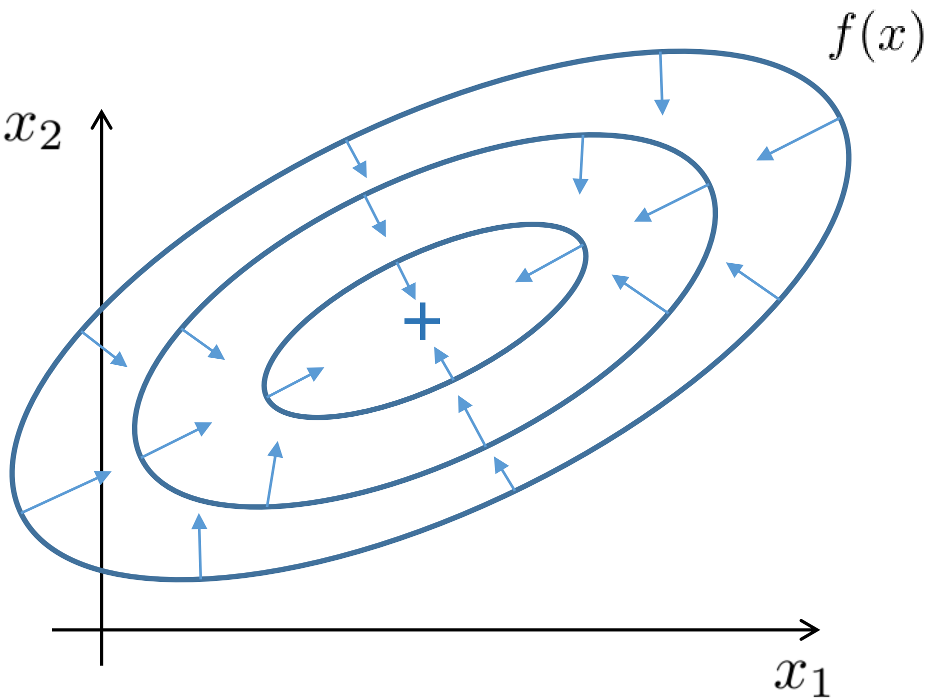


$-\nabla f(x)$ is vector of steepest descent

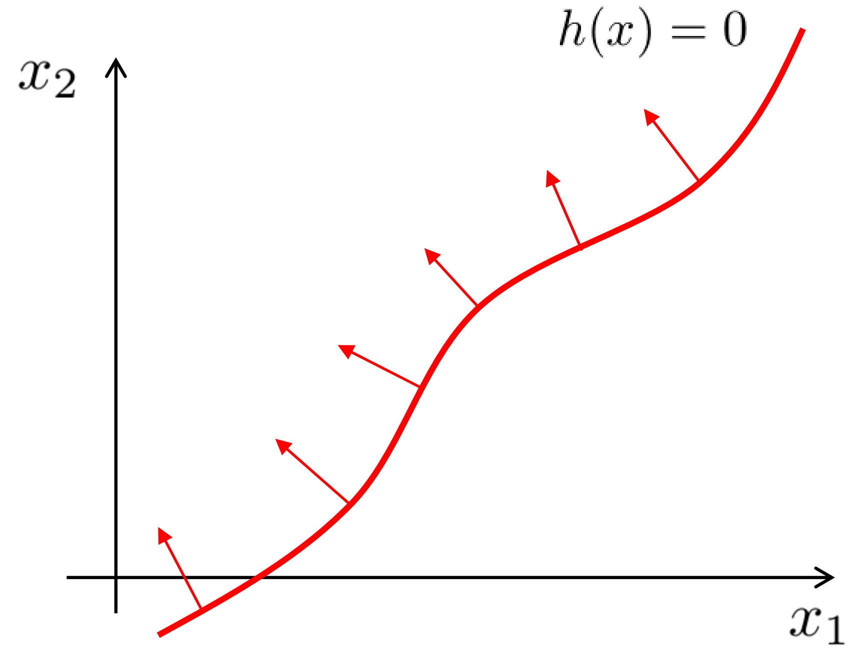
Optimization with equality constraints

"proof"

$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



$-\nabla f(x)$ is vector of steepest descent

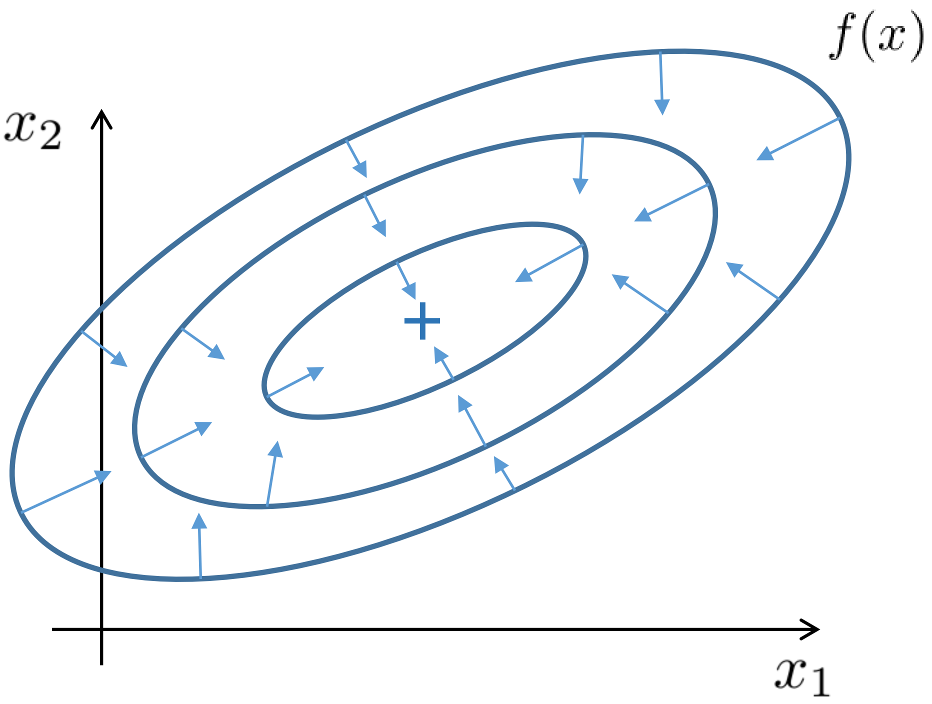


$\nabla h(x)$ is normal vector

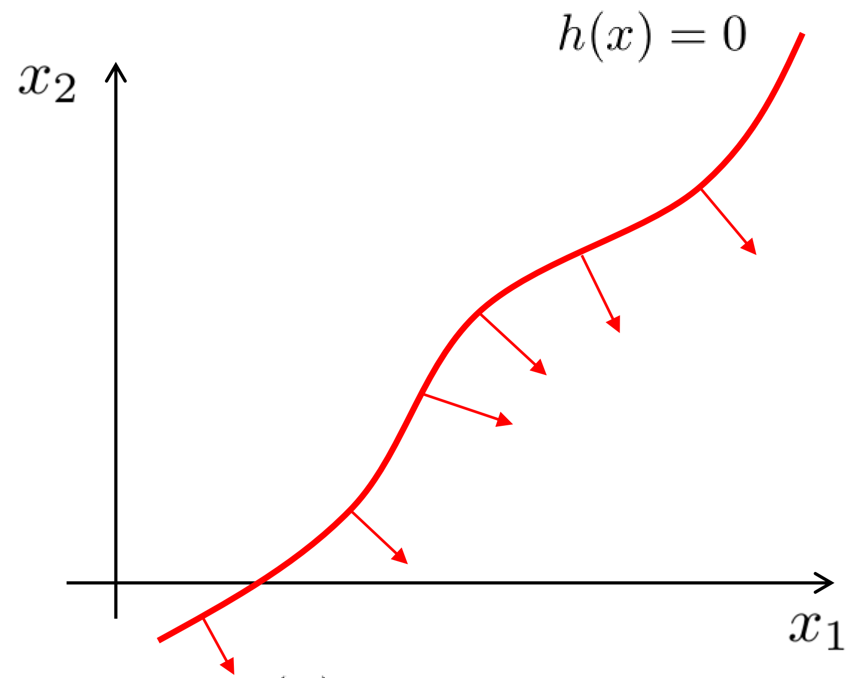
Optimization with equality constraints

"proof"

$$\begin{aligned}\nabla f(x) + \lambda \nabla h(x) &= 0\\ h(x) &= 0\end{aligned}$$



$-\nabla f(x)$ is vector of steepest descent



$\nabla h(x)$ is normal vector

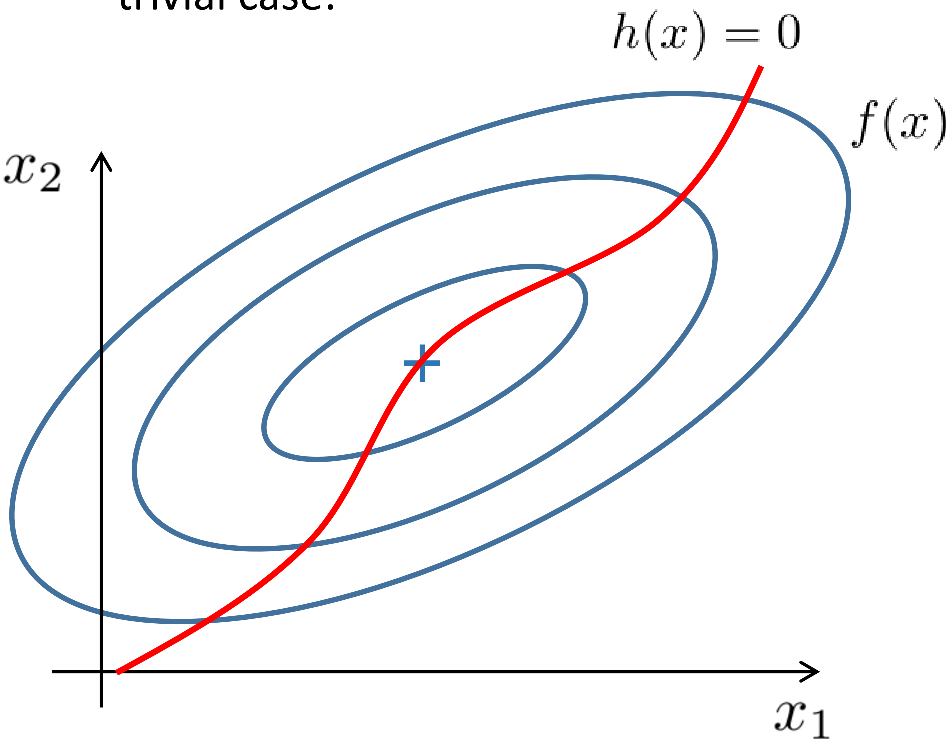
$$h(x) = 0 \quad \Leftrightarrow \quad -h(x) = 0$$

$$\nabla h(x) \qquad\qquad -\nabla h(x)$$

Optimization with equality constraints

"proof"

- trivial case:



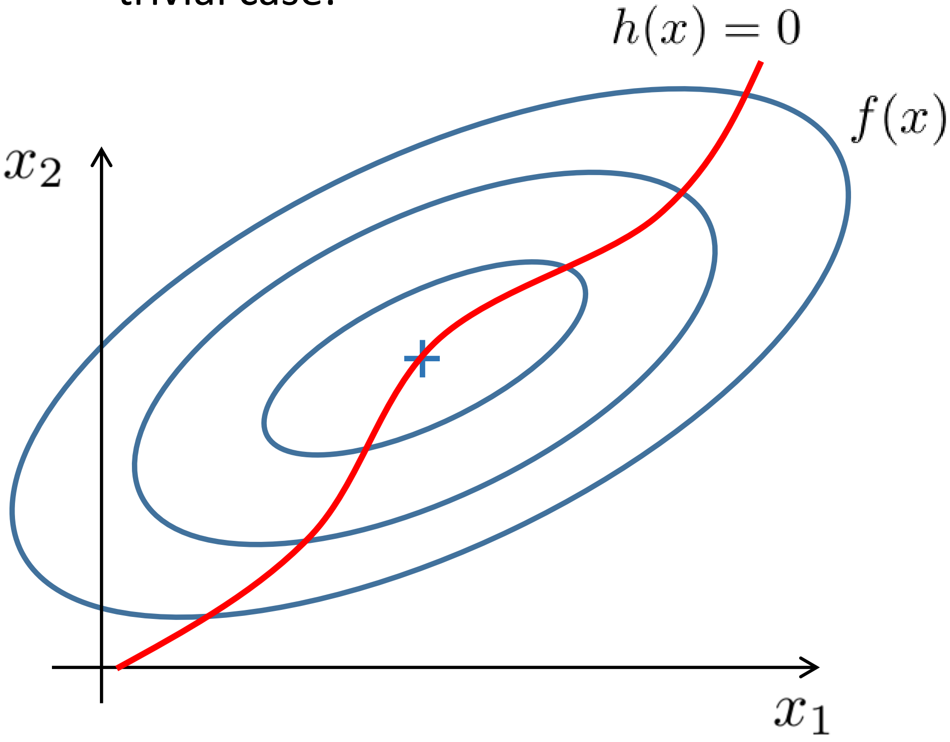
$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$

$$\arg\min_{x \in \mathbb{R}^n} f(x) \in \Omega$$

Optimization with equality constraints

"proof"

- trivial case:



condition for unconstrained problem

$$\begin{aligned} \nabla f(x) + \cancel{\lambda \nabla h(x)} &= 0 \\ h(x) &= 0 \end{aligned}$$

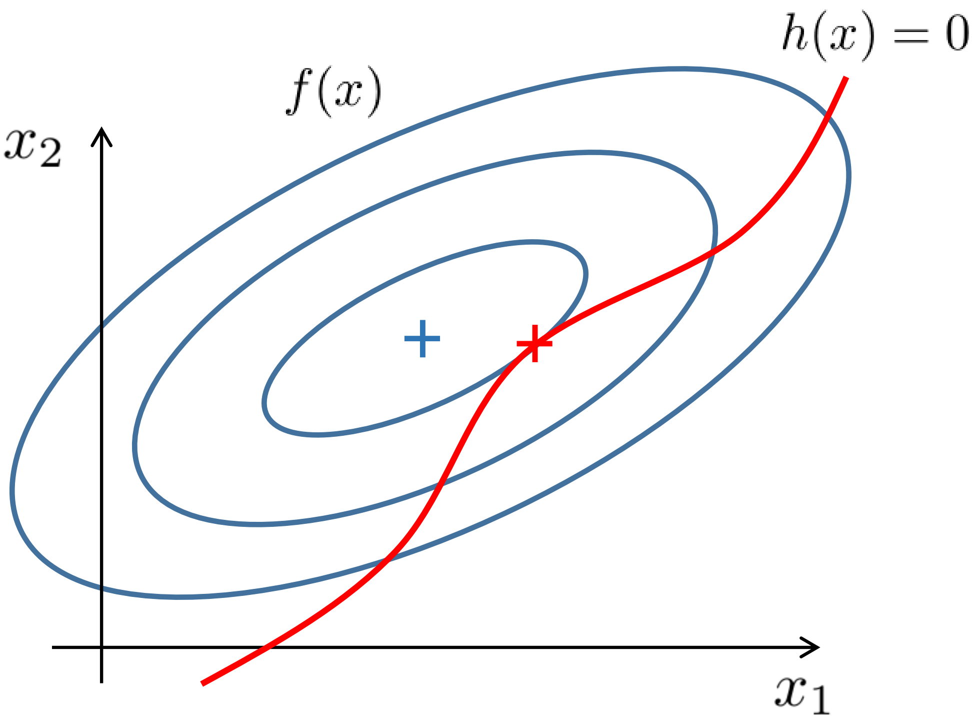
$$\arg\min_{x \in \mathbb{R}^n} f(x) \;\in \Omega$$

$$\lambda = 0$$

Optimization with equality constraints

"proof"

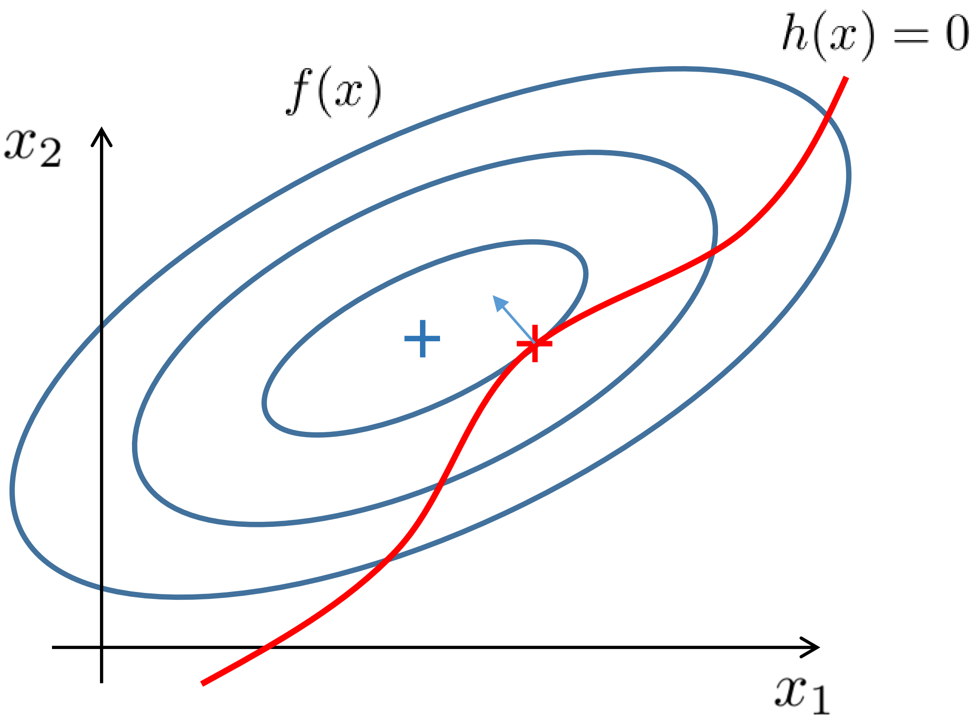
$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



Optimization with equality constraints

"proof"

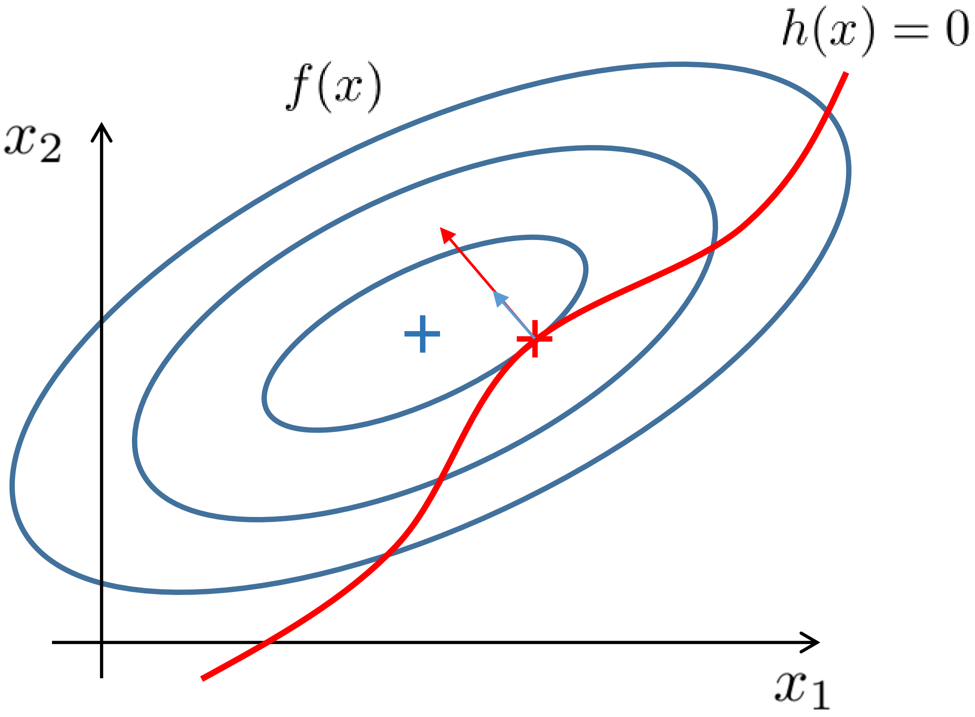
$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



Optimization with equality constraints

"proof"

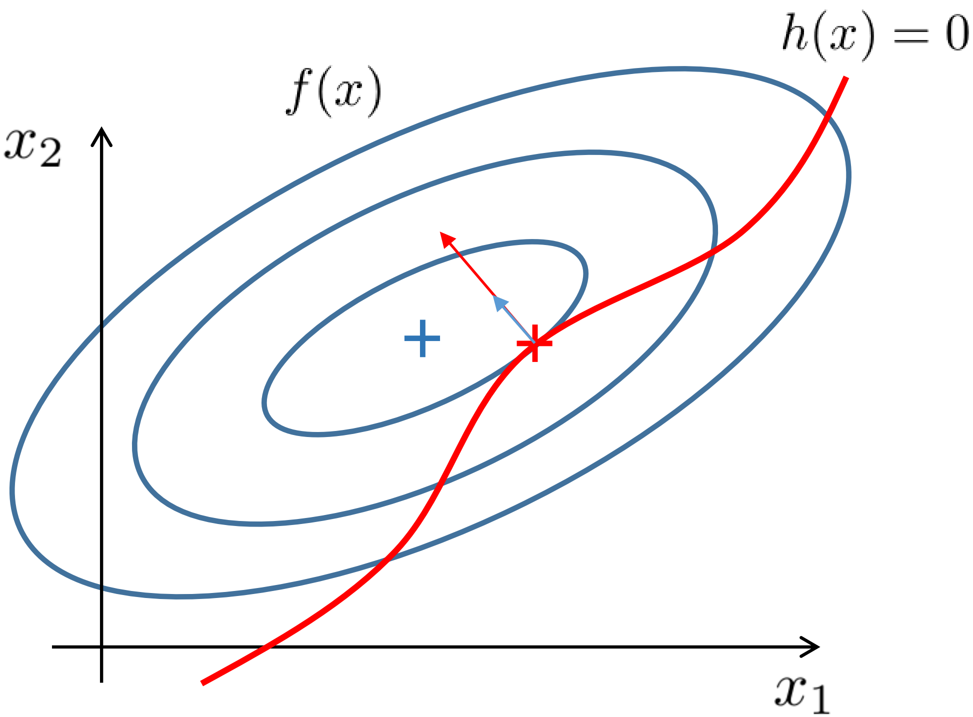
$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



Optimization with equality constraints

"proof"

$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$

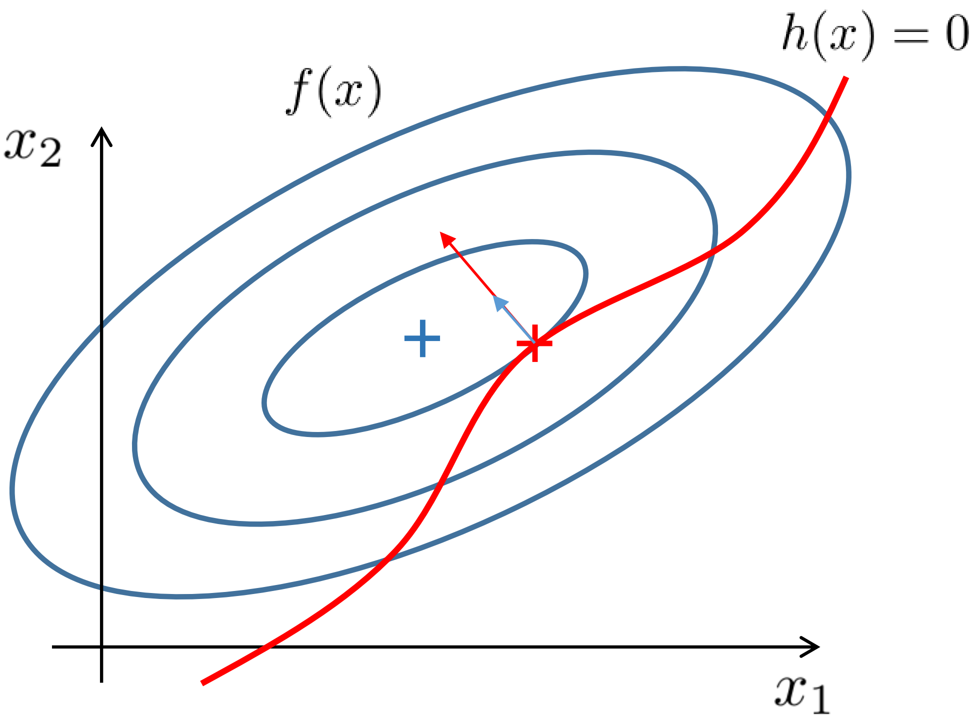


$\Rightarrow$ $\nabla f(x)$ and $\nabla h(x)$ have "same" direction (or opposite)
$\Rightarrow$ $\nabla f(x)$ and $\nabla h(x)$ are linearly dependent
$\Rightarrow$ there exists coefficient $\alpha \in \mathbb{R}$ such that $\nabla f(x) = \alpha \nabla h(x)$

Optimization with equality constraints

"proof"

$$\begin{aligned} \nabla f(x) + \lambda \nabla h(x) &= 0 \\ h(x) &= 0 \end{aligned}$$



$\Rightarrow$ $\nabla f(x)$ and $\nabla h(x)$ have "same" direction (or opposite)
$\Rightarrow$ $\nabla f(x)$ and $\nabla h(x)$ are linearly dependent
$\Rightarrow$ there exists coefficient $\underset{-\lambda}{\cancel{\alpha}} \in \mathbb{R}$ such that $\nabla f(x) = \underset{-\lambda}{\cancel{\alpha}} \nabla h(x)$

Equality constrained problems

Example:

$$\arg \min_{x_1+x_2=2} x_1^2 + x_2^2 + x_1 x_2 + 1000$$

$$L(x_1, x_2, \lambda) = x_1^2 + x_2^2 + x_1 x_2 + 1000 + \lambda(x_1 + x_2 - 2)$$

$$\begin{aligned}
\frac{\partial L}{\partial x_1} &= 2x_1 + x_2 + \lambda &= 0 \\
\frac{\partial L}{\partial x_2} &= 2x_2 + x_1 + \lambda &= 0 \\
\frac{\partial L}{\partial \lambda} &= x_1 + x_2 - 2 &= 0
\end{aligned}$$

$$\bar{x}_1 = \bar{x}_2 = 1, \quad \lambda = -3$$

3.) Inequality constrained optimization problem

Inequality constrained problems

$$\arg\min_{x\in\Omega} f(x), \quad \Omega = \{x \in \mathbb{R}^n : h_i(x) \leq 0, i = 1, \ldots, m\}$$

Inequality constrained problems

$$\arg\min_{x\in\Omega} f(x), \quad \Omega = \{x \in \mathbb{R}^n : h_i(x) \leq 0, i = 1, \ldots, m\}$$

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$$

Lagrange function

$$\begin{array}{rcrcl}
\nabla_x L(x,\lambda) & = & \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) & = & 0 \\
\nabla_{\lambda_i} L(x,\lambda) & = & h_i(x) & \leq & 0, \quad i = 1, \ldots, m \\
 & & \lambda_i & \geq & 0 \\
 & & \lambda_i h_i(x) & = & 0
\end{array}$$

Karush-Kuhn-Tucker optimality conditions

Inequality constrained problems

$$\arg\min_{x\in\Omega} f(x), \quad \Omega = \{x \in \mathbb{R}^n : h_i(x) \leq 0, i = 1, \ldots, m\}$$

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$$

Lagrange function

!!!

$$\begin{array}{rcll}
\nabla_x L(x,\lambda) & = & \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) & = \quad 0 \\
\nabla_{\lambda_i} L(x,\lambda) & = & h_i(x) & \leq \quad 0, \quad i = 1, \ldots, m \\
& & \lambda_i & \geq \quad 0 \\
& & \lambda_i h_i(x) & = \quad 0
\end{array}$$

Karush-Kuhn-Tucker optimality conditions

Inequality constrained problems

$$\arg\min_{x\in\Omega} f(x), \quad \Omega = \{x \in \mathbb{R}^n : h_i(x) \leq 0, i = 1, \ldots, m\}$$

$$L(x,\lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x)$$ Lagrange function

$$\begin{aligned}
\nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) &= 0 \\
\nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &\leq 0, \quad i = 1, \ldots, m \\
& \lambda_i &\geq 0 \\
& \boxed{\lambda_i h_i(x) = 0}
\end{aligned}$$

(complementarity condition)

Karush-Kuhn-Tucker optimality conditions

<u>"proof":</u>

1.) $\lambda = 0 \text{ and } h(x) \leq 0$

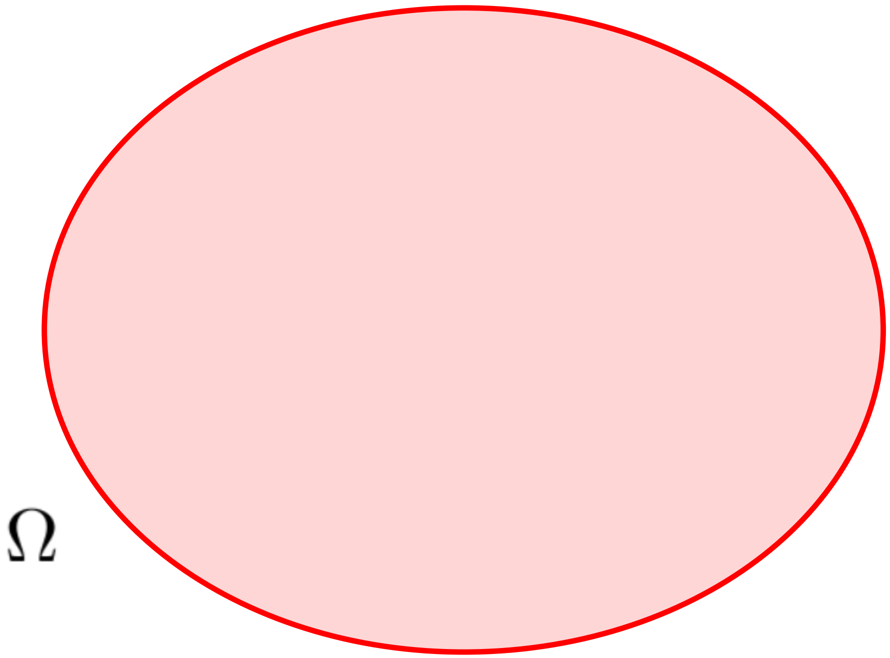
2.) $\lambda \geq 0 \text{ and } h(x) = 0$

3.) $\lambda = 0 \text{ and } h(x) = 0$

Inequality constrained problems

$$
\begin{aligned}
\nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) &&= 0 \\
\nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &&\leq 0, \quad i = 1,\ldots,m \\
& \lambda_i &&\geq 0 \\
& \lambda_i h_i(x) &&= 0
\end{aligned}
$$

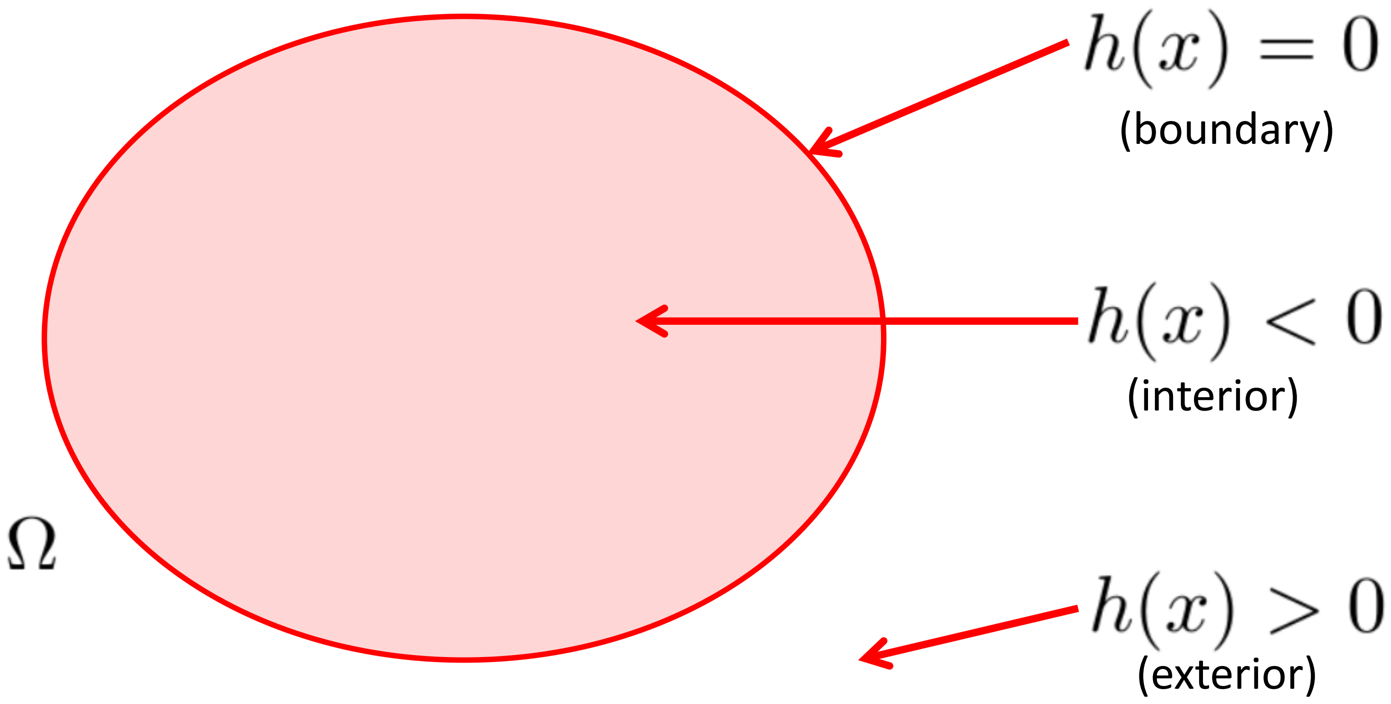
"proof":



Inequality constrained problems

$$
\begin{array}{rclcl}
\nabla_x L(x,\lambda) & = & \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) & = & 0 \\
\nabla_{\lambda_i} L(x,\lambda) & = & h_i(x) & \leq & 0, \quad i = 1,\ldots,m \\
 & & \lambda_i & \geq & 0 \\
 & & \lambda_i h_i(x) & = & 0
\end{array}
$$

"proof":

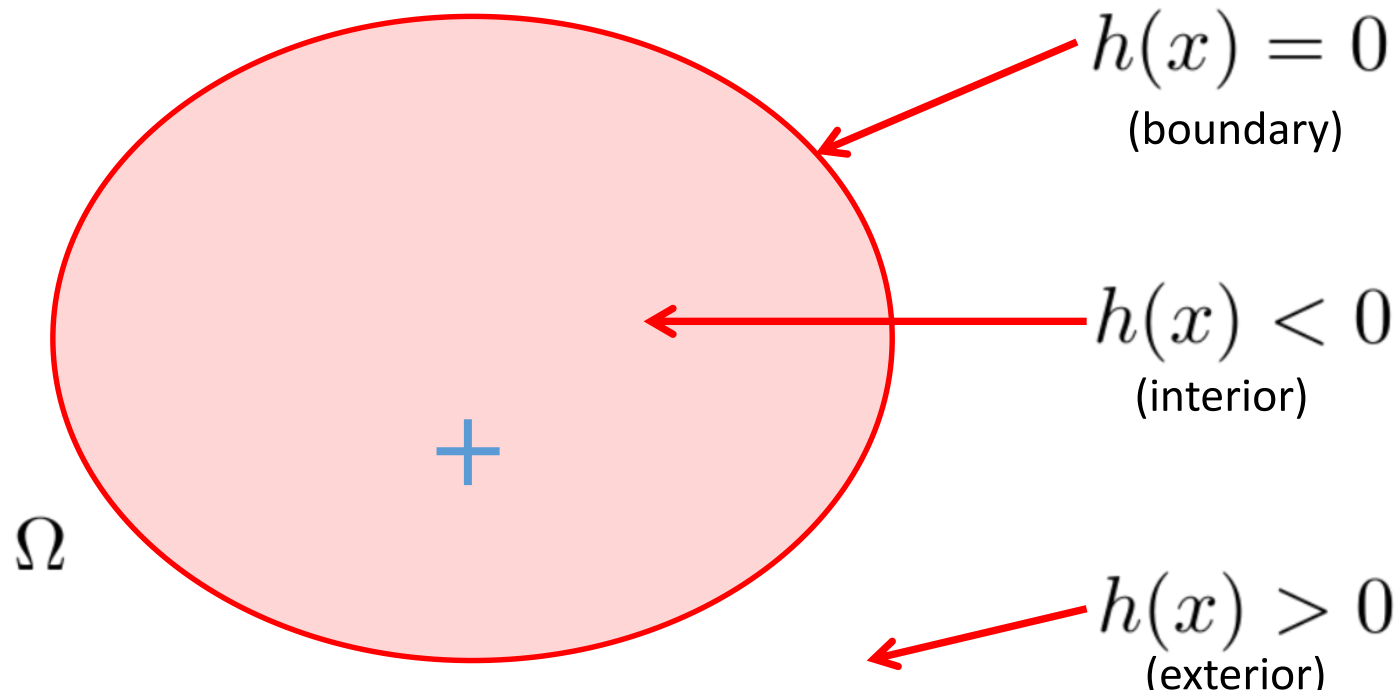


Inequality constrained problems

$$
\begin{aligned}
\nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) &&= 0 \\
\nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &&\leq 0, \quad i = 1,\dots,m \\
& \quad \lambda_i &&\geq 0 \\
& \quad \lambda_i h_i(x) &&= 0
\end{aligned}
$$

"proof":

1.) $\lambda = 0 \ \text{ and } \ h(x) \leq 0$



- Constraint does not have impact on solution
- After substitution of lambda into KKT we get unconstrained condition

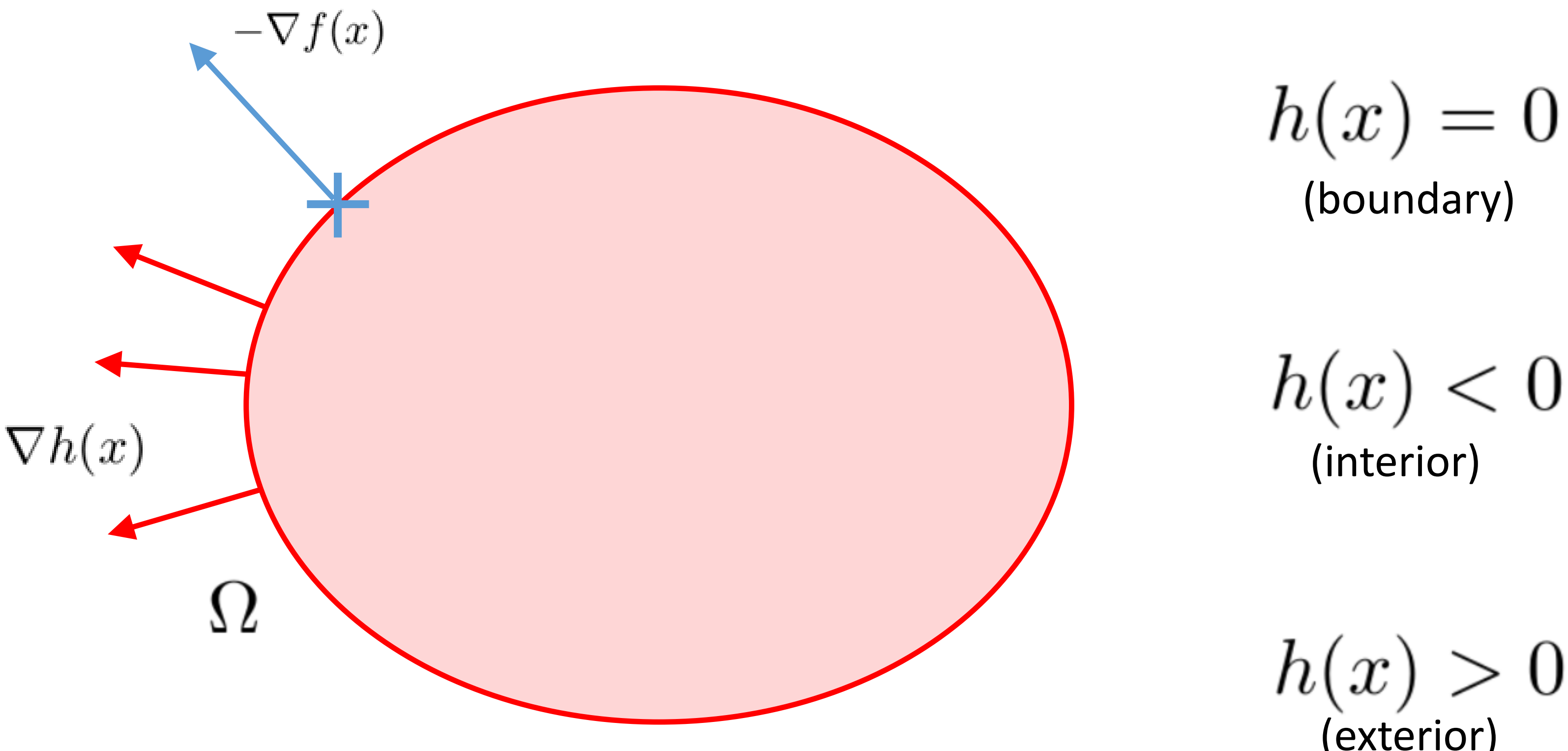
Inequality constrained = Unconstrained

Inequality constrained problems

$$\begin{aligned}
\nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) &&= 0 \\
\nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &&\leq 0, \quad i = 1,\dots,m \\
& \lambda_i &&\geq 0 \\
& \lambda_i h_i(x) &&= 0
\end{aligned}$$

"proof":

2.) $\lambda \geq 0 \;\text{ and }\; h(x) = 0$



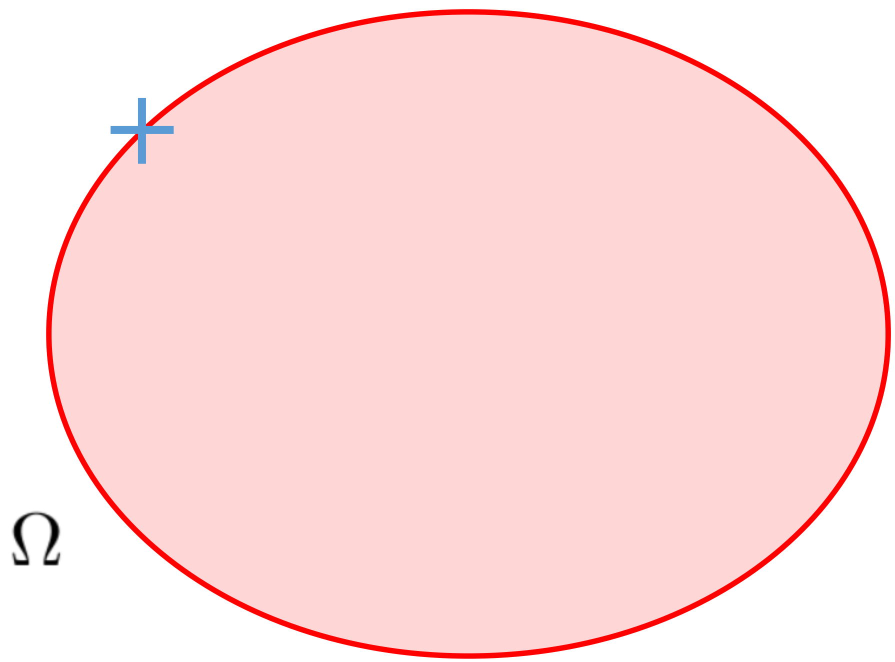
- Solution is on boundary (= equality constrained problem)
- Outer normal of feasible set and –gradient have the same direction ("function is decreasing out of the feasible set")

Inequality constrained problems

$$
\begin{aligned}
\nabla_x L(x,\lambda) &= \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla_x h_i(x) &&= 0 \\
\nabla_{\lambda_i} L(x,\lambda) &= h_i(x) &&\leq 0, \quad i = 1,\ldots,m \\
& \lambda_i &&\geq 0 \\
& \lambda_i h_i(x) &&= 0
\end{aligned}
$$

<u>"proof":</u>

3.) $\lambda = 0 \;\text{ and }\; h(x) = 0$



$h(x) = 0$
(boundary)

$h(x) < 0$
(interior)

$h(x) > 0$
(exterior)

- Solution is on boundary without impact of the constraint, gradient is equal to zero

Example: Kullback-Leibler estimation of Markov transitional matrix

- consider categorical data:

$$x_t \in \{s_1, \ldots, s_n\}, \quad t = 1, \ldots, T$$

Example: Kullback-Leibler estimation of Markov transitional matrix

- consider categorical data:

$$x_t \in \{s_1, \ldots, s_n\}, \quad t = 1, \ldots, T$$

- discrete probability density vectors:

$$\pi_t \in [0,1]^n, \quad \sum_{i=1}^{n} \{\pi_t\}_i = 1 \qquad \pi_t = \begin{bmatrix} P(x_t = s_1) \\ \vdots \\ P(x_t = s_n) \end{bmatrix}$$

Example: Kullback-Leibler estimation of Markov transitional matrix

- consider categorical data:

$$x_t \in \{s_1, \ldots, s_n\}, \quad t = 1, \ldots, T$$

- discrete probability density vectors:

$$\pi_t \in [0,1]^n, \quad \sum_{i=1}^{n} \{\pi_t\}_i = 1 \qquad \pi_t = \begin{bmatrix} P(x_t = s_1) \\ \vdots \\ P(x_t = s_n) \end{bmatrix}$$

- assume Markov process:

$$\pi_{t+1} = \Lambda \pi_t, \quad \Lambda \in [0,1]^{n,n}, \quad \forall j : \sum_{i=1}^{n} \{\Lambda\}_{i,j} = 1$$

$$\Lambda = \begin{bmatrix} P(x_{t+1} = s_1 \mid x_t = s_1) & \ldots & P(x_{t+1} = s_1 \mid x_t = s_n) \\ \vdots & \ddots & \vdots \\ P(x_{t+1} = s_n \mid x_t = s_1) & \ldots & P(x_{t+1} = s_n \mid x_t = s_n) \end{bmatrix}$$

Example: Kullback-Leibler estimation of Markov transitional matrix

- there is a "noise" in the data, therefore:

$$\pi_{t+1} \approx \Lambda \pi_t$$

In mathematical statistics, the **Kullback–Leibler divergence** (also called **relative entropy**) is a measure of how one probability distribution is different from a second, reference probability distribution.[1][2]

For discrete probability distributions P and Q defined on the same probability space, the Kullback–Leibler divergence between P and Q is defined[4] to be

$$D_{\mathrm{KL}}(P \parallel Q) = -\sum_{x \in \mathcal{X}} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$$

which is equivalent to

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log\left(\frac{P(x)}{Q(x)}\right).$$

$$\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda \pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda \pi_t\}_i}
\end{aligned}$$

...

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \arg\min \underbrace{- \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i \{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:=f(\Lambda), \quad f:\mathbb{R}^{n,n}\to\mathbb{R}}
\end{aligned}
$$

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \arg\min \underbrace{- \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i \{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:=f(\Lambda), \quad f:\mathbb{R}^{n,n}\rightarrow\mathbb{R}}
\end{aligned}
$$

$$
L(\Lambda, \lambda) := f(\Lambda) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{n} \{\Lambda\}_{i,j} - 1 \right)
$$

$$
\begin{aligned}
\frac{\partial L}{\partial\{\Lambda\}_{\hat{i},\hat{j}}} &= -\frac{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}} \{\pi_t\}_{\hat{j}}}{\{\Lambda\}_{\hat{i},\hat{j}}} + \lambda_{\hat{j}} &&= 0 \\
\frac{\partial L}{\partial\{\lambda\}_{\hat{j}}} &= \sum_{i=1}^{n} \{\Lambda\}_{i,\hat{j}} - 1 &&= 0
\end{aligned}
$$

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \arg\min \underbrace{- \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i \{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:= f(\Lambda), \quad f: \mathbb{R}^{n,n} \to \mathbb{R}}
\end{aligned}
$$

$$
L(\Lambda, \lambda) := f(\Lambda) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{n} \{\Lambda\}_{i,j} - 1 \right)
$$

$$
\{\Lambda\}_{\hat{i},\hat{j}} = \frac{c_{\hat{i},\hat{j}}}{\lambda_{\hat{j}}}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \{\Lambda\}_{\hat{i},\hat{j}}} &= -\frac{\boxed{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}} \{\pi_t\}_{\hat{j}}} =: c_{\hat{i},\hat{j}}}{\{\Lambda\}_{\hat{i},\hat{j}}} + \lambda_{\hat{j}} &= 0 \\
\frac{\partial L}{\partial \{\lambda\}_{\hat{j}}} &= \sum_{i=1}^{n} \{\Lambda\}_{i,\hat{j}} - 1 &= 0
\end{aligned}
$$

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \underbrace{\arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i\{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:=f(\Lambda), \quad f:\mathbb{R}^{n,n}\to\mathbb{R}}
\end{aligned}
$$

$$
\sum_{i=1}^{n} \frac{c_{i,\hat{j}}}{\lambda_{\hat{j}}} = 1
$$

$$
L(\Lambda, \lambda) := f(\Lambda) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{n} \{\Lambda\}_{i,j} - 1 \right)
$$

$$
\{\Lambda\}_{\hat{i},\hat{j}} = \frac{c_{\hat{i},\hat{j}}}{\lambda_{\hat{j}}}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial\{\Lambda\}_{\hat{i},\hat{j}}} &= -\frac{\overbrace{\boxed{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}}\{\pi_t\}_{\hat{j}}}}^{=:\, c_{\hat{i},\hat{j}}}}{\{\Lambda\}_{\hat{i},\hat{j}}} + \lambda_{\hat{j}} &= 0 \\
\frac{\partial L}{\partial\{\lambda\}_{\hat{j}}} &= \sum_{i=1}^{n} \{\Lambda\}_{i,\hat{j}} - 1 &= 0
\end{aligned}
$$

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \arg\min \underbrace{- \sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i\{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:=f(\Lambda), \quad f:\mathbb{R}^{n,n}\to\mathbb{R}}
\end{aligned}
$$

$$
\lambda_{\hat{j}} = \sum_{i=1}^{n} c_{i,\hat{j}} = \sum_{t=1}^{T-1} \underbrace{\sum_{i=1}^{n} \{\pi_{t+1}\}_i}_{=1} \{\pi_t\}_{\hat{j}} = \sum_{t=1}^{T-1} \{\pi_t\}_{\hat{j}}
$$

$$
\sum_{i=1}^{n} \frac{c_{i,\hat{j}}}{\lambda_{\hat{j}}} = 1
$$

$$
L(\Lambda, \lambda) := f(\Lambda) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{n} \{\Lambda\}_{i,j} - 1 \right)
$$

$$
\{\Lambda\}_{\hat{i},\hat{j}} = \frac{c_{\hat{i},\hat{j}}}{\lambda_{\hat{j}}}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial \{\Lambda\}_{\hat{i},\hat{j}}} &= -\frac{\boxed{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}}\{\pi_t\}_{\hat{j}}} =: c_{\hat{i},\hat{j}}}{\{\Lambda\}_{\hat{i},\hat{j}}} + \lambda_{\hat{j}} &= 0 \\
\frac{\partial L}{\partial \{\lambda\}_{\hat{j}}} &= \sum_{i=1}^{n} \{\Lambda\}_{i,\hat{j}} - 1 &= 0
\end{aligned}
$$

Example: Kullback-Leibler estimation of Markov transitional matrix

$$
\begin{aligned}
\Lambda^* &= \arg\min \sum_{t=1}^{T-1} D_{KL}(\pi_{t+1} \mid \Lambda\pi_t) \\
&= \arg\min \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \frac{\{\pi_{t+1}\}_i}{\{\Lambda\pi_t\}_i} \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log\{\Lambda\pi_t\}_i \\
&= \arg\min - \sum_{t=1}^{T-1} \sum_{i=1}^{n} \{\pi_{t+1}\}_i \log \sum_{j=1}^{n} \{\Lambda\}_{i,j}\{\pi_t\}_j \\
&\leq \arg\min - \underbrace{\sum_{t=1}^{T-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \{\pi_{t+1}\}_i \{\pi_t\}_j \log\{\Lambda\}_{i,j}}_{:=f(\Lambda), \quad f:\mathbb{R}^{n,n}\to\mathbb{R}}
\end{aligned}
$$

$$
L(\Lambda, \lambda) := f(\Lambda) + \sum_{j=1}^{n} \lambda_j \left(\sum_{i=1}^{n} \{\Lambda\}_{i,j} - 1 \right)
$$

$$
\begin{aligned}
\frac{\partial L}{\partial\{\Lambda\}_{\hat{i},\hat{j}}} &= -\frac{\boxed{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}}\{\pi_t\}_{\hat{j}}} =: c_{\hat{i},\hat{j}}}{\{\Lambda\}_{\hat{i},\hat{j}}} + \lambda_{\hat{j}} &= 0 \\
\frac{\partial L}{\partial\{\lambda\}_{\hat{j}}} &= \sum_{i=1}^{n} \{\Lambda\}_{i,\hat{j}} - 1 &= 0
\end{aligned}
$$

$$
\{\Lambda\}_{\hat{i},\hat{j}} = \frac{c_{\hat{i},\hat{j}}}{\lambda_{\hat{j}}}
$$

$$
\sum_{i=1}^{n} \frac{c_{i,\hat{j}}}{\lambda_{\hat{j}}} = 1
$$

$$
\lambda_{\hat{j}} = \sum_{i=1}^{n} c_{i,\hat{j}} = \sum_{t=1}^{T-1} \underbrace{\sum_{i=1}^{n} \{\pi_{t+1}\}_i}_{=1} \{\pi_t\}_{\hat{j}} = \sum_{t=1}^{T-1} \{\pi_t\}_{\hat{j}}
$$

$$
\boxed{\{\Lambda\}_{\hat{i},\hat{j}} = \frac{\sum_{t=1}^{T-1} \{\pi_{t+1}\}_{\hat{i}}\{\pi_t\}_{\hat{j}}}{\sum_{t=1}^{T-1} \{\pi_t\}_{\hat{j}}}}
$$