Faculty of Science

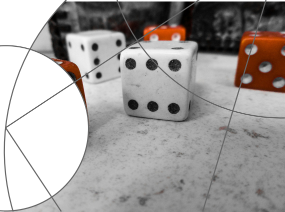
L2 – Linear Regression II (Excursus)

Modelling and Analysis of Data

Fabian Gieseke

Machine Learning Section
Department of Computer Science
University of Copenhagen

Universitetsparken 1, 1-1-N110
fabian.gieseke@di.ku.dk



21 November 2019

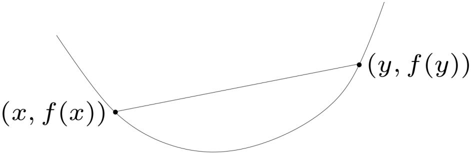
Excursus: Convex Optimization

Convex Functions (Boyd & Vandenberghe, 2009)

A function $f : \mathbb{R}^D \to \mathbb{R}$ is convex if

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ and $0 \leq \theta \leq 1$.



https://web.stanford.edu/~boyd/cvxbook/

Excursus: Convex Optimization

Convex Functions (Boyd & Vandenberghe, 2009)

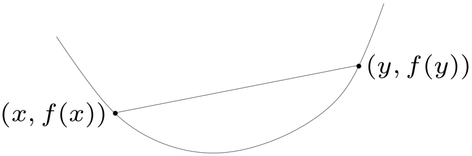
A function $f : \mathbb{R}^D \to \mathbb{R}$ is convex if

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$ and $0 \leq \theta \leq 1$. The function f is strictly convex if

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^D$, $\mathbf{x} \neq \mathbf{y}$, and $0 < \theta < 1$.



https://web.stanford.edu/~boyd/cvxbook/

Excursus: Convex Optimization

Hessian Matrix & Convexity (Boyd & Vandenberghe, 2009)

The Hessian matrix is a square matrix containing all the second-order partial derivatives $f_{x_i x_j} = \frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}$ of a function $f : \mathbb{R}^D \to \mathbb{R}$, i.e.:

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

https://web.stanford.edu/~boyd/cvxbook/

Excursus: Convex Optimization

Hessian Matrix & Convexity (Boyd & Vandenberghe, 2009)

The Hessian matrix is a square matrix containing all the second-order partial derivatives $f_{x_i x_j} = \frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}$ of a function $f : \mathbb{R}^D \rightarrow \mathbb{R}$, i.e.:

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

1. $\mathbf{H} = \nabla^2 f(\mathbf{x})$ positive semidefinite for all $\mathbf{x} \in \mathbb{R}^D \Leftrightarrow f$ convex.
   Here, $\mathbf{H}$ is positive semidefinite if $\mathbf{z}^T \boldsymbol{H}\boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^D$ with $\boldsymbol{z} \neq \mathbf{0}$.

https://web.stanford.edu/~boyd/cvxbook/

Excursus: Convex Optimization

Hessian Matrix & Convexity (Boyd & Vandenberghe, 2009)

The Hessian matrix is a square matrix containing all the second-order partial derivatives $f_{x_i x_j} = \frac{\partial^2}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\frac{\partial f}{\partial x_j}$ of a function $f : \mathbb{R}^D \to \mathbb{R}$, i.e.:

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

1. $\mathbf{H} = \nabla^2 f(\mathbf{x})$ positive semidefinite for all $\mathbf{x} \in \mathbb{R}^D \Leftrightarrow f$ convex.
   Here, $\mathbf{H}$ is positive semidefinite if $\mathbf{z}^T \boldsymbol{H} \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^D$ with $\boldsymbol{z} \neq \mathbf{0}$.
2. $\mathbf{H} = \nabla^2 f(\mathbf{x})$ positive definite for all $\mathbf{x} \in \mathbb{R}^D \Rightarrow f$ strictly convex.
   Here, $\mathbf{H}$ is positive definite if $\mathbf{z}^T \boldsymbol{H} \boldsymbol{z} > 0$ for all $\boldsymbol{z} \in \mathbb{R}^D$ with $\boldsymbol{z} \neq \mathbf{0}$.

https://web.stanford.edu/~boyd/cvxbook/

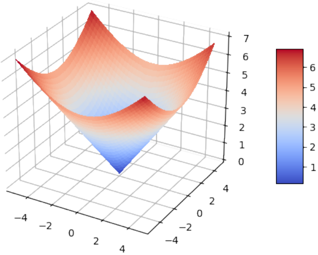
Excursus: Convex Optimization

Example I

Let $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1, x_2) = x_1^2 + x_2^2$. We have $\frac{\partial f}{\partial x_1} = 2x_1$, $\frac{\partial f}{\partial x_2} = 2x_2$, and:

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Is **H** positive (semi-)definite?



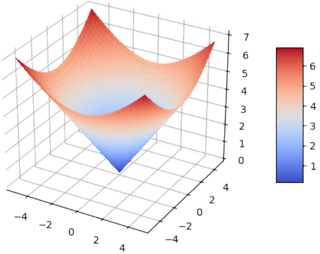
Excursus: Convex Optimization

Example I

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $f(x_1, x_2) = x_1^2 + x_2^2$. We have $\frac{\partial f}{\partial x_1} = 2x_1$, $\frac{\partial f}{\partial x_2} = 2x_2$, and:

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

Is $\mathbf{H}$ positive (semi-)definite? Since $\mathbf{z}^T \boldsymbol{H} \boldsymbol{z} = 2(z_1^2 + z_2^2) > 0$ for all $\boldsymbol{z} \neq \mathbf{0}$, the Hessian $\mathbf{H}$ is positive definite (for all $\mathbf{x}$). Hence, f is strictly convex.



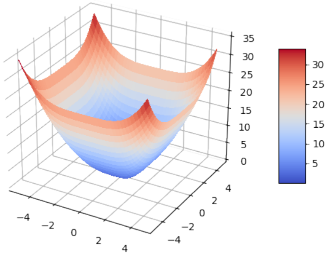
Excursus: Convex Optimization

Example II

Let $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1, x_2) = x_1^4 + x_2^4$. We have $\frac{\partial f}{\partial x_1} = 4x_1^3$, $\frac{\partial f}{\partial x_2} = 4x_2^3$, and:

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 & 0 \\ 0 & 12x_2^2 \end{bmatrix}$$

Is $\mathbf{H}$ positive (semi-)definite?



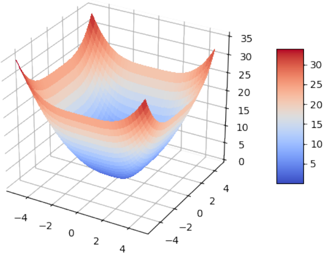
Excursus: Convex Optimization

Example II

Let $f : \mathbb{R}^2 \to \mathbb{R}$ with $f(x_1, x_2) = x_1^4 + x_2^4$. We have $\frac{\partial f}{\partial x_1} = 4x_1^3$, $\frac{\partial f}{\partial x_2} = 4x_2^3$, and:

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 & 0 \\ 0 & 12x_2^2 \end{bmatrix}$$

Is $\mathbf{H}$ positive (semi-)definite? Since $\mathbf{z}^T \boldsymbol{H} \boldsymbol{z} = 12(z_1^2 x_1^2 + z_2^2 x_2^2) \geq 0$ for all $\boldsymbol{z} \neq \mathbf{0}$, the Hessian $\mathbf{H}$ is positive semidefinite for all $\mathbf{x}$. Hence, f is convex (at least).



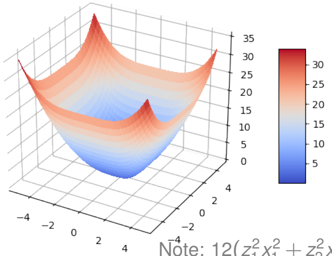
Excursus: Convex Optimization

Example II

Let $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ with $f(x_1, x_2) = x_1^4 + x_2^4$. We have $\frac{\partial f}{\partial x_1} = 4x_1^3$, $\frac{\partial f}{\partial x_2} = 4x_2^3$, and:

$$\mathbf{H} = \nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2 \partial x_2} \end{bmatrix} = \begin{bmatrix} 12x_1^2 & 0 \\ 0 & 12x_2^2 \end{bmatrix}$$

Is $\mathbf{H}$ positive (semi-)definite? Since $\mathbf{z}^T \boldsymbol{H}\boldsymbol{z} = 12(z_1^2 x_1^2 + z_2^2 x_2^2) \geq 0$ for all $\boldsymbol{z} \neq \mathbf{0}$, the Hessian $\mathbf{H}$ is positive semidefinite for all $\mathbf{x}$. Hence, f is convex (at least).

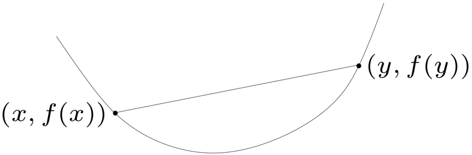


Note: $12(z_1^2 x_1^2 + z_2^2 x_2^2)$ becomes zero for $\mathbf{x} = (0, 0)^T$.

Excursus: Convex Optimization

Local and Global Optima (Boyd & Vandenberghe, 2009)

Any local minimum of a convex function $f : \mathbb{R}^D \to \mathbb{R}$ is a global minimum.



https://web.stanford.edu/~boyd/cvxbook/

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$
- Enforcing $\nabla\mathcal{L}(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^T\boldsymbol{t} \stackrel{!}{=} \mathbf{0}$ leads to

$$\hat{\mathbf{w}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$
- Enforcing $\nabla\mathcal{L}(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^T\boldsymbol{t} \stackrel{!}{=} \mathbf{0}$ leads to

$$\hat{\mathbf{w}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

- Now, the Hessian is given by $\mathbf{H} = \nabla^2 f(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}$
  Check on your own. Or have a look at formula (98) of
  https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$
- Enforcing $\nabla\mathcal{L}(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^T\boldsymbol{t} \stackrel{!}{=} \mathbf{0}$ leads to

$$\hat{\mathbf{w}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

- Now, the Hessian is given by $\mathbf{H} = \nabla^2 f(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}$
  Check on your own. Or have a look at formula (98) of
  `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- For any $\boldsymbol{z} \in \mathbb{R}^{D+1}$, we have $\boldsymbol{z}^T\left(\frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\right)\mathbf{z} \geq 0$. Why?

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$
- Enforcing $\nabla\mathcal{L}(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^T\boldsymbol{t} \stackrel{!}{=} \mathbf{0}$ leads to

$$\hat{\mathbf{w}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

- Now, the Hessian is given by $\mathbf{H} = \nabla^2 f(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}$
  Check on your own. Or have a look at formula (98) of
  `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- For any $\boldsymbol{z} \in \mathbb{R}^{D+1}$, we have $\boldsymbol{z}^T\left(\frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\right)\mathbf{z} \geq 0$. Why?
- A little trick: We can rewrite this in the following form

$$\boldsymbol{z}^T\left(\frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\right)\mathbf{z} = \frac{2}{N}(\mathbf{X}\boldsymbol{z})^T(\boldsymbol{X}\mathbf{z}) = \frac{2}{N}\mathbf{v}^T\mathbf{v} = \frac{2}{N}\sum_{j=1}^{N} v_j^2 \geq 0$$

Gradient & Global Minimum

- We have $\mathcal{L}(\mathbf{w}) = \frac{1}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{t} + \frac{1}{N}\boldsymbol{t}^T\boldsymbol{t}$
- Enforcing $\nabla\mathcal{L}(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - \frac{2}{N}\boldsymbol{X}^T\boldsymbol{t} \stackrel{!}{=} \mathbf{0}$ leads to

$$\hat{\mathbf{w}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{t}$$

- Now, the Hessian is given by $\mathbf{H} = \nabla^2 f(\mathbf{w}) = \frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}$
  Check on your own. Or have a look at formula (98) of
  `https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf`
- For any $\boldsymbol{z} \in \mathbb{R}^{D+1}$, we have $\boldsymbol{z}^T\left(\frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\right)\mathbf{z} \geq 0$. Why?
- A little trick: We can rewrite this in the following form

$$\boldsymbol{z}^T\left(\frac{2}{N}\boldsymbol{X}^T\boldsymbol{X}\right)\mathbf{z} = \frac{2}{N}(\mathbf{X}\boldsymbol{z})^T(\boldsymbol{X}\mathbf{z}) = \frac{2}{N}\mathbf{v}^T\mathbf{v} = \frac{2}{N}\sum_{j=1}^{N} v_j^2 \geq 0$$

- Thus, the Hessian $\mathbf{H} = \nabla^2 f(\mathbf{w})$ is positive semidefinite (for all $\mathbf{w}$). This means that $\mathcal{L}$ is a convex function and that our $\hat{\mathbf{w}}$ is a global minimum.

Convex Optimization (in case you are interested)

Stephen P. Boyd

- Home
- Teaching
- Biography

Research

- Books
- Papers
- Software
- People

Classes

- EE103
- EE263
- EE363
- EE364a
- EE364b
- EE365

MOOC

- CVX101

Convex Optimization – Boyd and Vandenberghe

Convex Optimization
Stephen Boyd and Lieven Vandenberghe

Cambridge University Press

A MOOC on convex optimization, CVX101, was run from 1/21/14 to 3/14/14. If you register for it, you can access all the course materials.

More material can be found at the web sites for EE364A (Stanford) or EE236B (UCLA), and our own web pages. Source code for almost all examples and figures in part 2 of the book is available in CVX (in the examples directory), in CVXOPT (in the book examples directory), and in CVXPY. Source code for examples in Chapters 9, 10, and 11 can be found here. Instructors can obtain complete solutions to exercises by email request to us; please give us the URL of the course you are teaching.

If you find an error not listed in our errata list, please do let us know about it.

Stephen Boyd & Lieven Vandenberghe

https://web.stanford.edu/~boyd/cvxbook/