Introduction to intelligent systems

Optimization

Mikkel N. Schmidt

Technical University of Denmark,
DTU Compute, Department of Applied Mathematics and Computer Science.

Overview

1. Gradient descent
2. Linear regression (with gradient descent)
3. Neural network (with gradient descent)
4. Tasks

Feedback group

- Nicholas Borch
- Alfred Fonnesbech Aqraou
- Josefine Høgsted Voglhofer
- Rasmus Bernth Linnemann

Learning objectives

- II Gradient descent algorithm.
- I Stochastic gradient descent.
- II Gradient of cost function.
- II Neural networks: Model (layers, activation functions), parameters, cost function.

- I Understand the concepts and definitions, and know their application. Reason about the concepts in the context of an example. Use correct technical terminology.
- II As above plus: Read, manipulate, and work with technical definitions and expressions (mathematical and Python code). Carry out practical computations. Interpret and evaluate results.

Gradient descent

Gradient descent

- Iterative method for finding optimum of a function
- Start at an initial point
- Updates parameters by taking step proportional to negative of the gradient
- Repeat until convergence

Partial derivative

- Derivative of a function of *several variable* with respect to *one* of those variables, with the others *held constant*.
- Notation

$$\frac{\partial f}{\partial x_1}, \qquad \frac{\partial f(x_1, x_2)}{\partial x_1}, \qquad \frac{\partial f}{\partial x_1}(x_1, x_2)$$

- The partial derivative evaluated at a certain point

$$\left.\frac{\partial f}{\partial x_1}\right|_{x_1=5, x_2=7}$$

Partial derivative, definition

Derivative, function of single variable, $f(x)$

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Partial derivative, definition

Derivative, function of single variable, $f(x)$

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h}$$

Partial derivative, function of multiple variables, $f(x_1, x_2)$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \lim_{h\to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$

Partial derivative, definition

Derivative, function of single variable, $f(x)$

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Partial derivative, function of multiple variables, $f(x_1, x_2)$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$
$$\frac{\partial f}{\partial x_2}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h}$$

Gradient

Definition

$$\nabla f(x_1, x_2, \dots) = \begin{bmatrix} \dfrac{\partial f(x_1, x_2, \dots)}{\partial x_1} \\[2ex] \dfrac{\partial f(x_1, x_2, \dots)}{\partial x_2} \\[2ex] \vdots \end{bmatrix}$$

Exercise: Gradient calculation

Multivariate function

$$f(x, y) = x^2 \cos(y)$$

What is the gradient?

Gradient definition

$$\nabla f(x, y) = \begin{bmatrix} \dfrac{\partial f(x, y)}{\partial x} \\ \\ \dfrac{\partial f(x, y)}{\partial y} \end{bmatrix}$$

Exercise: Gradient calculation

Multivariate function

$$f(x, y) = x^2 \cos(y)$$

What is the gradient?

Partial derivatives

$$\frac{\partial f(x, y)}{\partial x} = 2x \cos(y)$$
$$\frac{\partial f(x, y)}{\partial y} = -x^2 \sin(y)$$

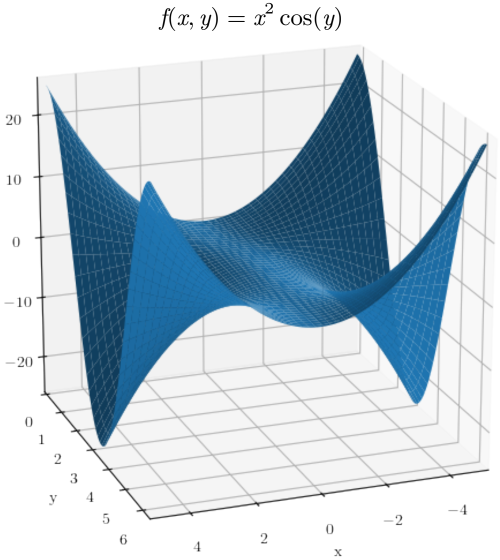
Gradient

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f(x,y)}{\partial x} \\ \frac{\partial f(x,y)}{\partial y} \end{bmatrix} = \begin{bmatrix} 2x \cos(y) \\ -x^2 \sin(y) \end{bmatrix}$$

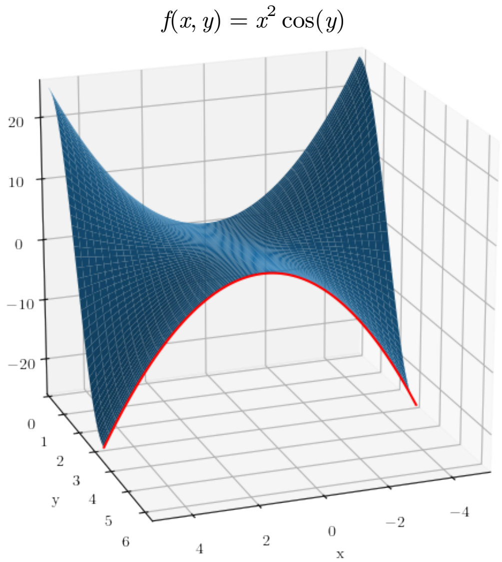
Gradient definition

$$\nabla f(x, y) = \begin{bmatrix} \dfrac{\partial f(x, y)}{\partial x} \\ \\ \dfrac{\partial f(x, y)}{\partial y} \end{bmatrix}$$

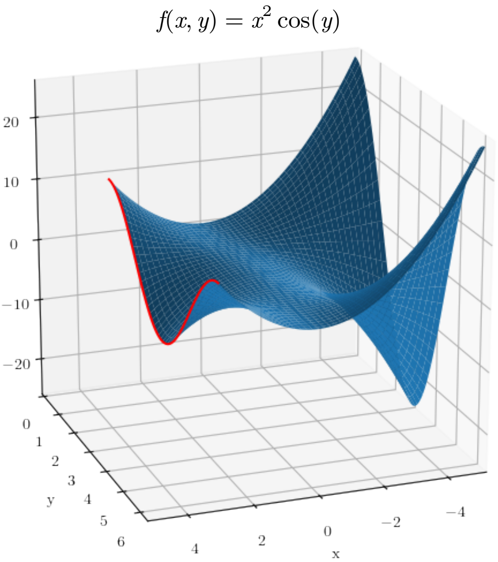
Gradient



Gradient



Gradient



Gradient descent

Initialize $x^{(0)}$
Repeat, $t = 0, 1, 2, \ldots$

$$\underbrace{x^{(t+1)}}_{\text{new parameter value}} = \underbrace{x^{(t)}}_{\text{old parameter value}} - \underbrace{\alpha}_{\text{step size}} \cdot \underbrace{\nabla f(x^{(t)})}_{\text{gradient}}$$

until convergence

Partial derivative in vector form

Partial derivative, function of multiple variables, $f(x_1, x_2)$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$
$$\frac{\partial f}{\partial x_2}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h}$$

Partial derivative in vector form

Partial derivative, function of multiple variables, $f(x_1, x_2)$

$$\frac{\partial f}{\partial x_1}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1 + h, x_2) - f(x_1, x_2)}{h}$$
$$\frac{\partial f}{\partial x_2}(x_1, x_2) = \lim_{h \to 0} \frac{f(x_1, x_2 + h) - f(x_1, x_2)}{h}$$

Partial derivative in vector form, function of a vector, $f(\bar{x})$, where $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\frac{\partial f}{\partial x_1}(\bar{x}) = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{e}_1) - f(\bar{x})}{h}$$
$$\frac{\partial f}{\partial x_2}(\bar{x}) = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{e}_2) - f(\bar{x})}{h}$$

$$\bar{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \bar{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Directional derivative

How much does the function change if we move the parameters $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in the direction $\bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

Directional derivative

How much does the function change if we move the parameters $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in the direction $\bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h}$$

Directional derivative

How much does the function change if we move the parameters $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in the direction $\bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$$\begin{aligned} \nabla_{\bar{v}} f(\bar{x}) &= \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} \\ &= \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h} \end{aligned}$$

Directional derivative

How much does the function change if we move the parameters $\bar{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ in the direction $\bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$

$$\begin{aligned}
\nabla_{\bar{v}} f(\bar{x}) &= \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} \\
&= \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h} \\
&= \nabla f(\bar{x}) \cdot \bar{v} \qquad \leftarrow \text{We will show this}
\end{aligned}$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \quad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \quad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$g(h) = f(\underbrace{x_1 + h \cdot v_1}_{z_1(h)}, \underbrace{x_2 + h \cdot v_2}_{z_2(h)}) = f(\bar{x} + h\bar{v})$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \quad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$g(h) = f(\underbrace{x_1 + h \cdot v_1}_{z_1(h)}, \underbrace{x_2 + h \cdot v_2}_{z_2(h)}) = f(\bar{x} + h\bar{v})$$

$$\left.\frac{\mathrm{d}g}{\mathrm{d}h}\right|_{h=0} = \lim_{\epsilon \to 0} \frac{g(0 + \epsilon) - g(0)}{\epsilon}$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h\to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h\to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \quad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$g(h) = f(\underbrace{x_1 + h \cdot v_1}_{z_1(h)}, \underbrace{x_2 + h \cdot v_2}_{z_2(h)}) = f(\bar{x} + h\bar{v})$$

$$\begin{aligned} \left.\frac{\mathrm{d}g}{\mathrm{d}h}\right|_{h=0} &= \lim_{\epsilon\to 0} \frac{g(0+\epsilon) - g(0)}{\epsilon} \\ &= \lim_{\epsilon\to 0} \frac{f(\bar{x} + \epsilon\bar{v}) - f(\bar{x})}{\epsilon} = \underline{\nabla_{\bar{v}} f(\bar{x})} \end{aligned}$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \qquad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$g(h) = f(\underbrace{x_1 + h \cdot v_1}_{z_1(h)}, \underbrace{x_2 + h \cdot v_2}_{z_2(h)}) = f(\bar{x} + h\bar{v})$$

$$\begin{aligned} \left.\frac{\mathrm{d}g}{\mathrm{d}h}\right|_{h=0} &= \lim_{\epsilon \to 0} \frac{g(0+\epsilon) - g(0)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{f(\bar{x} + \epsilon\bar{v}) - f(\bar{x})}{\epsilon} = \underline{\nabla_{\bar{v}} f(\bar{x})} \end{aligned}$$

$$= \frac{\partial f}{\partial z_1}\frac{\partial z_1}{\partial h} + \frac{\partial f}{\partial z_2}\frac{\partial z_2}{\partial h}$$

Directional derivative: Proof

$$\nabla_{\bar{v}} f(\bar{x}) = \lim_{h \to 0} \frac{f(x_1 + h \cdot v_1, x_2 + h \cdot v_2) - f(x_1, x_2)}{h} = \lim_{h \to 0} \frac{f(\bar{x} + h\bar{v}) - f(\bar{x})}{h}, \quad \bar{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$g(h) = f(\underbrace{x_1 + h \cdot v_1}_{z_1(h)}, \underbrace{x_2 + h \cdot v_2}_{z_2(h)}) = f(\bar{x} + h\bar{v})$$

$$\begin{aligned} \left.\frac{\mathrm{d}g}{\mathrm{d}h}\right|_{h=0} &= \lim_{\epsilon \to 0} \frac{g(0 + \epsilon) - g(0)}{\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{f(\bar{x} + \epsilon\bar{v}) - f(\bar{x})}{\epsilon} = \underline{\nabla_{\bar{v}} f(\bar{x})} \\ \\ &= \frac{\partial f}{\partial z_1}\frac{\partial z_1}{\partial h} + \frac{\partial f}{\partial z_2}\frac{\partial z_2}{\partial h} \\ &= \frac{\partial f}{\partial z_1} v_1 + \frac{\partial f}{\partial z_2} v_2 = \underline{\nabla f \cdot \bar{v}} \end{aligned}$$

Direction of steepest descent

Directional derivative

$$\nabla_{\bar{v}} f(\bar{x}) = \nabla f(\bar{x}) \cdot \bar{v}$$

- Measures how much the function changes when we move a bit in the direction $\bar{v}$

Direction of steepest descent

Directional derivative

$$\nabla_{\bar{v}} f(\bar{x}) = \nabla f(\bar{x}) \cdot \bar{v}$$

- Measures how much the function changes when we move a bit in the direction $\bar{v}$

Which direction maximizes the directional derivative?

Direction of steepest descent

Directional derivative

$$\nabla_{\bar{v}} f(\bar{x}) = \nabla f(\bar{x}) \cdot \bar{v}$$

- Measures how much the function changes when we move a bit in the direction $\bar{v}$

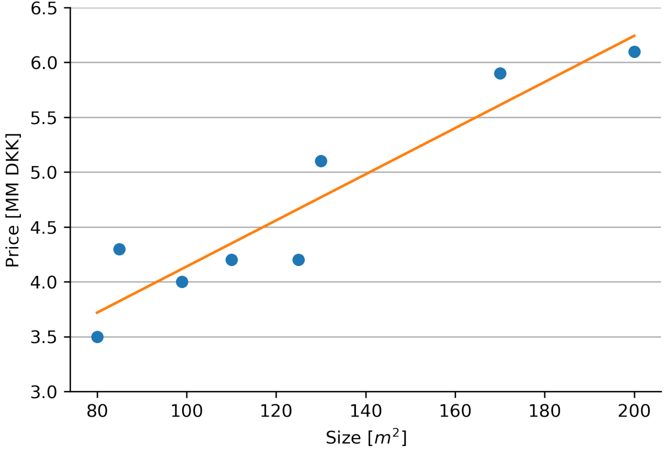
Which direction maximizes the directional derivative?
The dot product is maximal when the two vectors are parallel

$$\bar{v} = \frac{\nabla f(\bar{x})}{\|\nabla f(\bar{x})\|}$$

I.e. the gradient points in the direction of steepest ascent.

Linear regression (with gradient descent)

Remember linear regression



Gradient descent in linear regression

Linear regression

- Regression line: $f(x) = ax + b$
- Cost: Squared distance between data and regression line

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

What is the gradient?

$$\nabla E(a, b) = \begin{bmatrix} \frac{\partial E(a,b)}{\partial a} \\ \frac{\partial E(a,b)}{\partial b} \end{bmatrix}$$

Gradient descent in linear regression

Linear regression

- Regression line: $f(x) = ax + b$
- Cost: Squared distance between data and regression line

$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

What is the gradient?

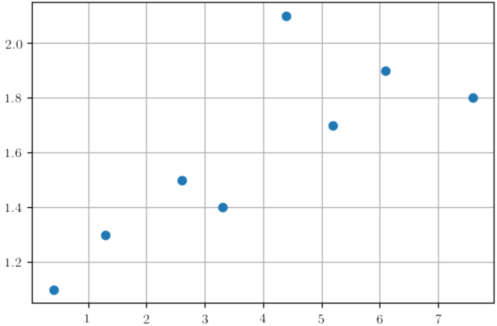
$$\nabla E(a, b) = \begin{bmatrix} \frac{\partial E(a,b)}{\partial a} \\ \frac{\partial E(a,b)}{\partial b} \end{bmatrix}$$

Solution

$$\frac{\partial E}{\partial a} = \sum_{n=1}^{N} -2(y_n - ax_n - b)x_n$$

$$\frac{\partial E}{\partial b} = \sum_{n=1}^{N} -2(y_n - ax_n - b)$$

Linear regression data

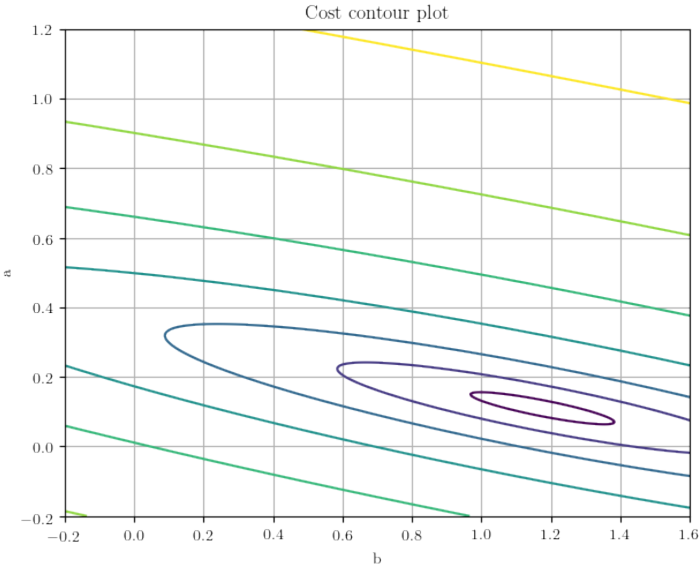


- Regression line: $f(x) = ax + b$
- Cost: Squared distance between data and regression line

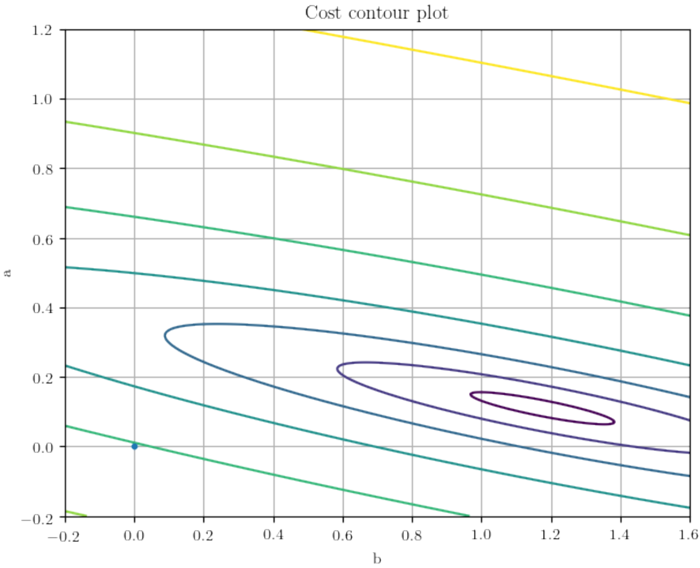
$$E = \sum_{n=1}^{N} (y_n - f(x_n))^2$$

The cost, $E(a, b)$, is a function of two variables. What does it look like?

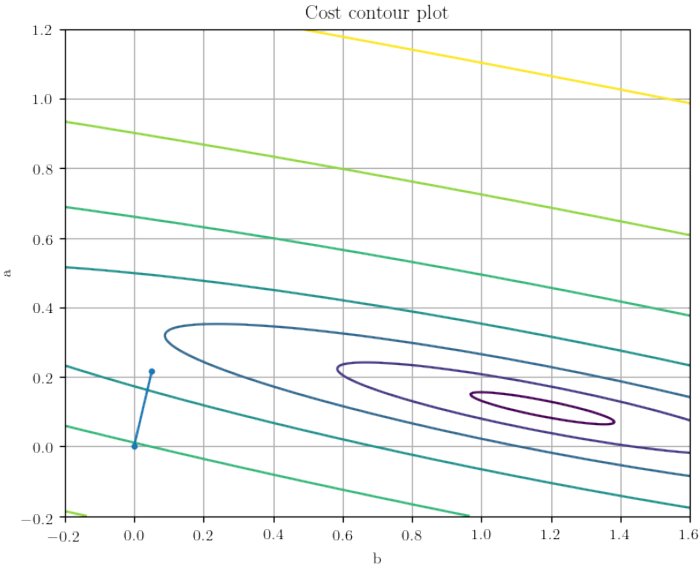
Gradient steps



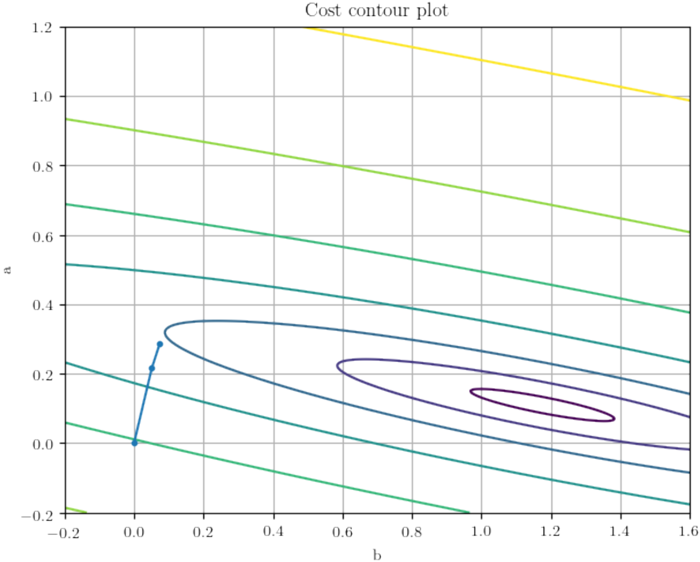
Gradient steps



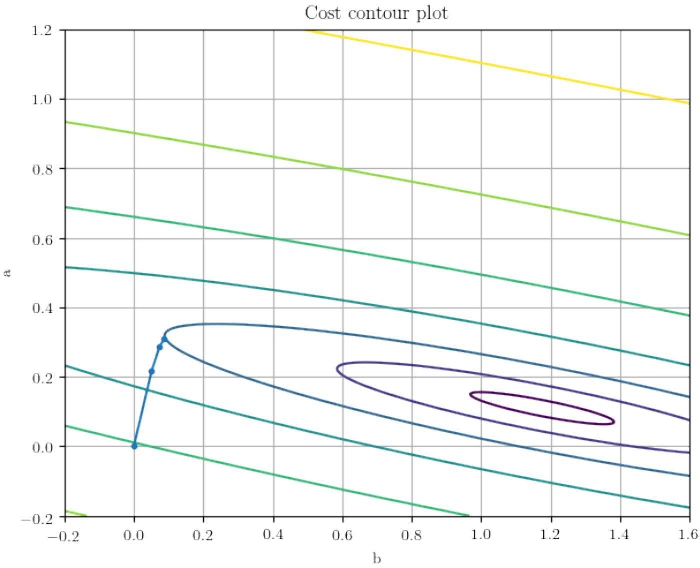
Gradient steps



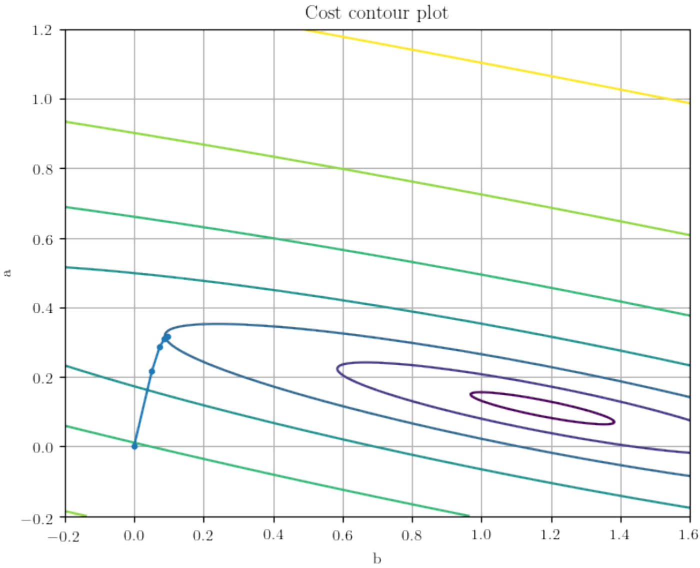
Gradient steps



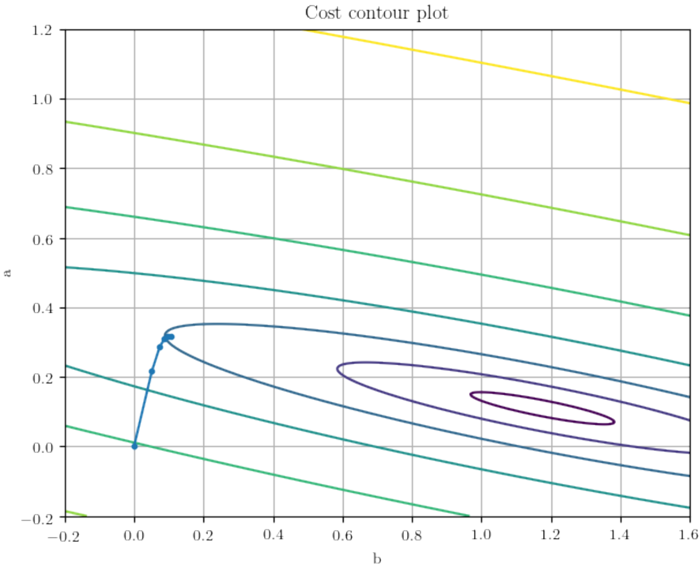
Gradient steps



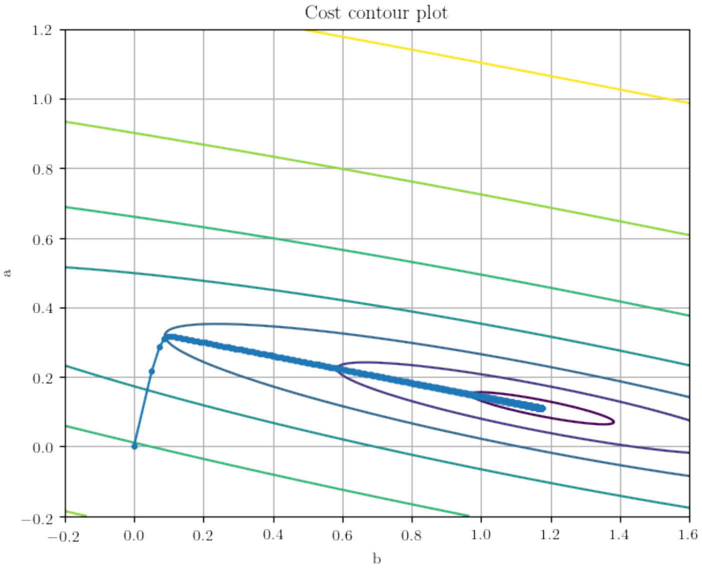
Gradient steps



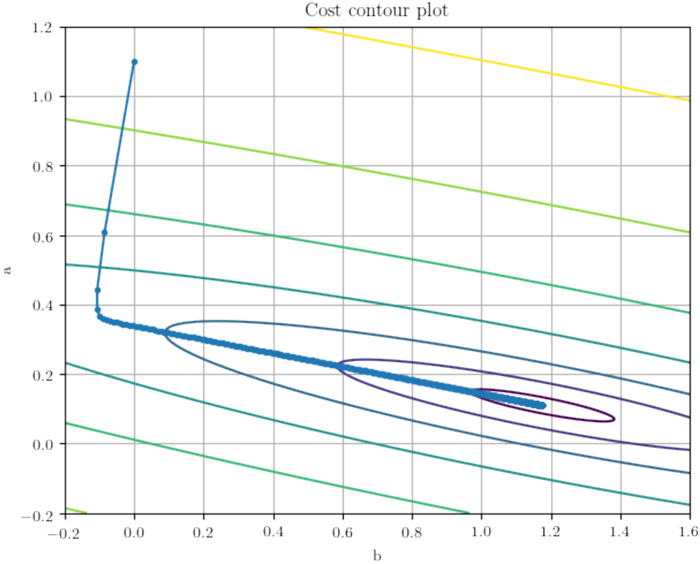
Gradient steps



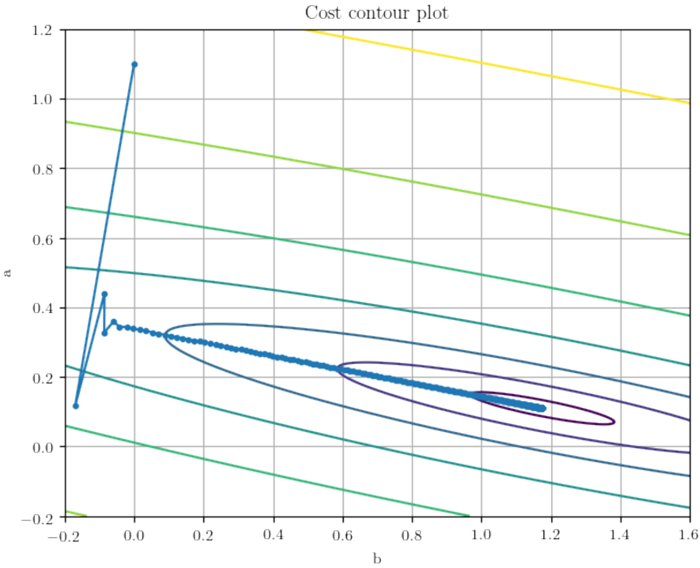
Gradient descent



Gradient descent



Gradient descent



Learning rate

- Small learning rate can lead to slow convergence
- Large learning rate may lead to divergence

Comparison with setting derivative equal to zero

Derivative equal to zero and solve

- No parameters to tune
- Closed form solution
- Slow for many features
  Need to solve N equations in N unknowns

Gradient descent

- Need to select step size
- Needs many iterations
- Fast for many features
  Need only compute the gradient

Feature scaling

- Is gradient descent sensitive to the scale of features? YES
- Features on different scale = parameters on different scale

Feature scaling

Min-max normalization

Rescale the range to $[0, 1]$

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardization

Rescale to have zero mean and unit variance

$$x' = \frac{x - \bar{x}}{\sigma_x}$$

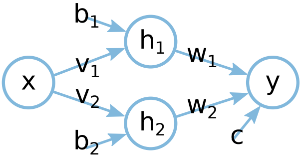
($\bar{x}$: mean, σ_x: standard deviation)

Neural network (with gradient descent)

Neural network

Cost function

$$E = \sum_{n=1}^{N} \left(y_n - \hat{y}_n\right)^2$$

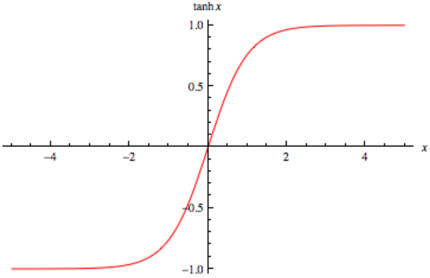


Network structure

$$\hat{y}_n = w_1 h_1(x_n) + w_2 h_2(x_n) + c$$

$$h_1(x_n) = \tanh\left(v_1 x_n + b_1\right)$$

$$h_2(x_n) = \tanh\left(v_2 x_n + b_2\right)$$



Model parameters

$$c,\ w_1,\ w_2,\ v_1,\ v_2,\ b_1,\ b_2$$

Exercise: Gradient of neural network

Compute the partial derivatives

$$\frac{\partial E}{\partial c}, \quad \frac{\partial E}{\partial w_1}, \quad \frac{\partial E}{\partial b_1}, \quad \frac{\partial E}{\partial v_1}$$

Hints

1. Use the chain rule
2. $\dfrac{\partial \tanh(x)}{\partial x} = 1 - \tanh^2(x)$
3. Don't expand terms needlessly. Express in terms of e.g. $\hat{y}_n$ and $h_1(x_n)$ where possible.

Cost function

$$E = \sum_{n=1}^{N} \left(y_n - \hat{y}_n\right)^2$$

Neural network model

$$\begin{aligned} \hat{y}_n &= w_1 h_1(x_n) + w_2 h_2(x_n) + c \\ h_1(x_n) &= \tanh\left(v_1 x_n + b_1\right) \\ h_2(x_n) &= \tanh\left(v_2 x_n + b_2\right) \end{aligned}$$

Solution: Gradient of neural network

$$\frac{\partial E}{\partial c} = -2 \sum_{n=1}^{N} \left(y_n - \hat{y}_n \right)$$

Solution: Gradient of neural network

$$\frac{\partial E}{\partial c} = -2 \sum_{n=1}^{N} \Big(y_n - \hat{y}_n \Big)$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{n=1}^{N} \Big(\big(y_n - \hat{y}_n \big) h_1(x_n) \Big)$$

Solution: Gradient of neural network

$$\frac{\partial E}{\partial c} = -2 \sum_{n=1}^{N} \left(y_n - \hat{y}_n \right)$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{n=1}^{N} \left(\left(y_n - \hat{y}_n \right) h_1 \left(x_n \right) \right)$$

$$\frac{\partial E}{\partial b_1} = -2 \sum_{n=1}^{N} \left(\left(y_n - \hat{y}_n \right) w_1 \left(1 - h_1^2(x_n) \right) \right)$$

Solution: Gradient of neural network

$$\frac{\partial E}{\partial c} = -2 \sum_{n=1}^{N} \left(y_n - \hat{y}_n \right)$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{n=1}^{N} \left(\left(y_n - \hat{y}_n \right) h_1 \left(x_n \right) \right)$$

$$\frac{\partial E}{\partial b_1} = -2 \sum_{n=1}^{N} \left(\left(y_n - \hat{y}_n \right) w_1 \left(1 - h_1^2 (x_n) \right) \right)$$

$$\frac{\partial E}{\partial v_1} = -2 \sum_{n=1}^{N} \left(\left(y_n - \hat{y}_n \right) w_1 \left(1 - h_1^2 (x_n) \right) \right) x_n$$

Analysis of gradient

- Gradient scales with the error $y_n - \hat{y}_n$
- If $h_1(x_n)$ saturates at -1 or +1, the term $1 - h_1^2(n)$ is zero

Partial derivatives

$$\frac{\partial E}{\partial c} = -2 \sum_{n=1}^{N} \Big(y_n - \hat{y}_n \Big)$$

$$\frac{\partial E}{\partial w_1} = -2 \sum_{n=1}^{N} \Big((y_n - \hat{y}_n) h_1(x_n) \Big)$$

$$\frac{\partial E}{\partial b_1} = -2 \sum_{n=1}^{N} \Big((y_n - \hat{y}_n) w_1 (1 - h_1^2(n)) \Big)$$

$$\frac{\partial E}{\partial v_1} = -2 \sum_{n=1}^{N} \Big((y_n - \hat{y}_n) w_1 (1 - h_1^2(n)) \Big) x_n$$

Tasks

Tasks

Tasks today

1. Work through the notebook `08-GradientDescentLinearRegression.ipynb`
2. Work through the notebook `08-GradientDescentNeuralNet.ipynb`
3. Today's feedback group
   - Nicholas Borch
   - Alfred Fonnesbech Aqraou
   - Josefine Høgsted Voglhofer
   - Rasmus Bernth Linnemann

Lab report hand in

- Lab 3: Image segmentation (Deadline: Thursday 26 October 20:00)