Maths fundamentals for AI and deep learning

Linear Algebra

Scope: Focused on the subset of linear algebra most relevant to deep learning.

Scalars

- A scalar is a single number
- Integers, real numbers, rational numbers, etc.
- We denote it with italic font:

$$a, n, x$$

Vectors

- A vector is a 1-D array of numbers:

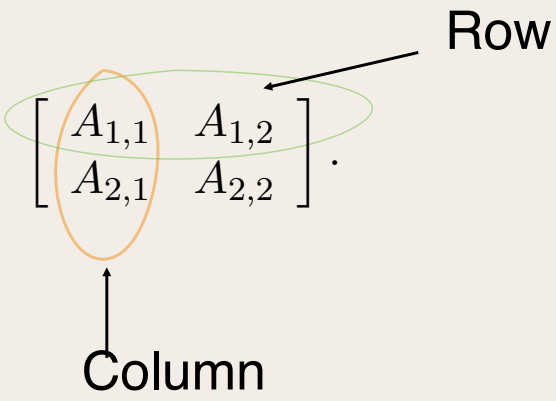
$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}. \tag{2.1}$$

- Can be real, binary, integer, etc.

- Example notation for type and size: $\mathbb{R}^n$

Matrices

- A matrix is a 2-D array of numbers:



$$\begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}. \tag{2.2}$$

- Example notation for type and shape: $\boldsymbol{A} \in \mathbb{R}^{m \times n}$

Tensors

- A tensor is an array of numbers, that may have
  - zero dimensions, and be a scalar
  - one dimension, and be a vector
  - two dimensions, and be a matrix
  - or more dimensions.

Matrix Transpose

$$(\boldsymbol{A}^\top)_{i,j} = A_{j,i}. \tag{2.3}$$

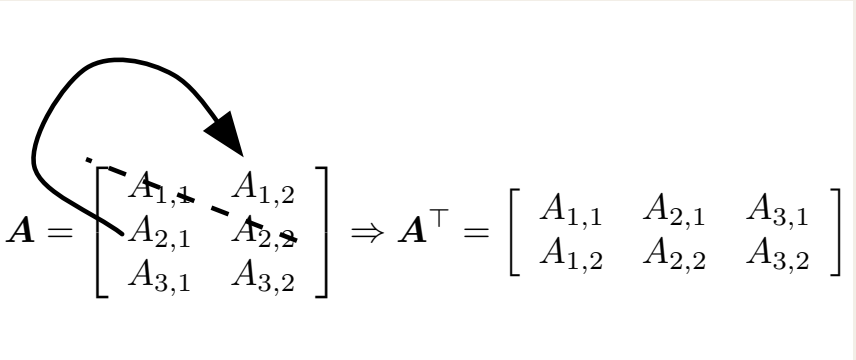


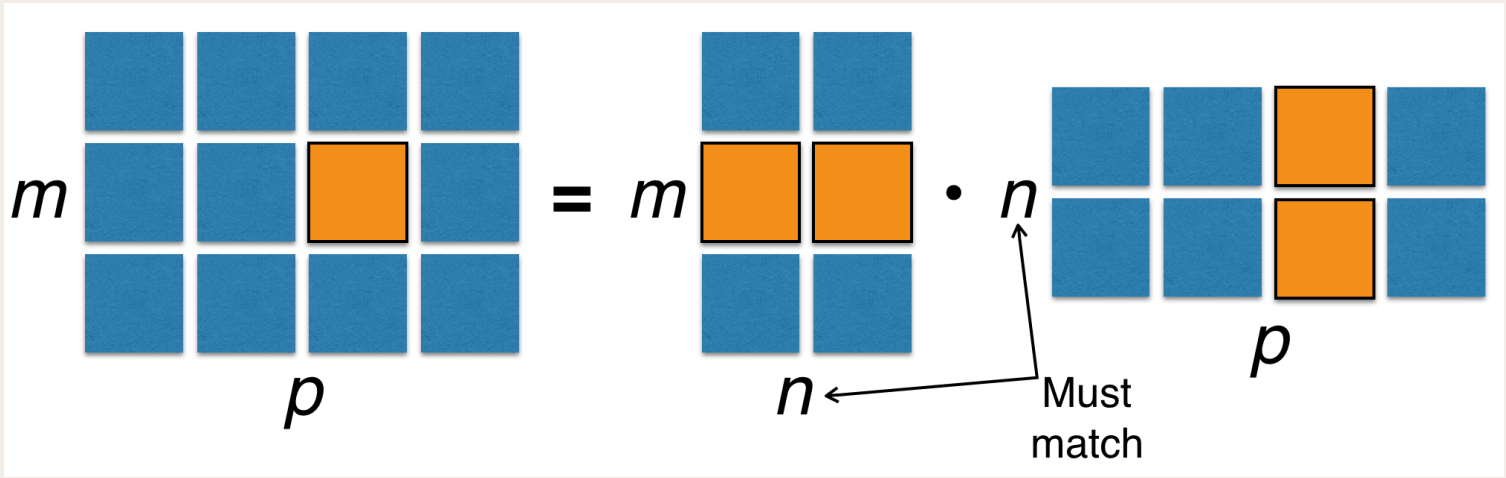
Figure 2.1: The transpose of the matrix can be thought of as a mirror image across the main diagonal.

$$(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top. \tag{2.9}$$

Matrix (Dot) Product

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}. \tag{2.4}$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}. \tag{2.5}$$



Identity Matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Figure 2.2: *Example identity matrix*: This is $\boldsymbol{I}_3$.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{I}_n \boldsymbol{x} = \boldsymbol{x}. \tag{2.20}$$

Systems of Equations

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.11}$$

expands to

$$\boldsymbol{A}_{1,:}\boldsymbol{x} = b_1 \tag{2.12}$$

$$\boldsymbol{A}_{2,:}\boldsymbol{x} = b_2 \tag{2.13}$$

$$\dots \tag{2.14}$$

$$\boldsymbol{A}_{m,:}\boldsymbol{x} = b_m \tag{2.15}$$

Solving Systems of Equations

- A linear system of equations can have:
  - No solution
  - Many solutions
  - Exactly one solution: this means multiplication by the matrix is an invertible function

Matrix Inversion

- Matrix inverse:

$$\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}_n. \tag{2.21}$$

- Solving a system using an inverse:

$$\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \tag{2.22}$$

$$\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.23}$$

$$\boldsymbol{I}_n\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b} \tag{2.24}$$

- Numerically unstable, but useful for abstract analysis

Invertibility

- Matrix can’t be inverted if…
  - More rows than columns
  - More columns than rows
  - Redundant rows/columns (“linearly dependent”, “low rank”)

Norms

- Functions that measure how “large” a vector is
- Similar to a distance between zero and the point represented by the vector
  - $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
  - $f(\boldsymbol{x} + \boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ (the *triangle inequality*)
  - $\forall \alpha \in \mathbb{R}, f(\alpha \boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$

Norms

- L^p norm

$$||\boldsymbol{x}||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

- Most popular norm: L2 norm, p=2
- L1 norm, p=1: $||\boldsymbol{x}||_1 = \sum_i |x_i|.$ (2.31)
- Max norm, infinite p:

$$||\boldsymbol{x}||_\infty = \max_i |x_i|. \quad (2.32)$$

Special Matrices and Vectors

- Unit vector:

$$||\boldsymbol{x}||_2 = 1. \tag{2.36}$$

- Symmetric Matrix:

$$\boldsymbol{A} = \boldsymbol{A}^\top. \tag{2.35}$$

- Orthogonal matrix: $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{I}.$ (2.37)

$$\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$$

Learning linear algebra

- Do a lot of practice problems
- Start out with lots of summation signs and indexing into individual entries
- Eventually you will be able to mostly use matrix and vector product notation quickly and easily

Probability and Information Theory

Probability Mass Function

- The domain of P must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, 0 \leq P(x) \leq 1$. An impossible event has probability 0 and no state can be less probable than that. Likewise, an event that is guaranteed to happen has probability 1, and no state can have a greater chance of occurring.
- $\sum_{x \in \mathrm{x}} P(x) = 1$. We refer to this property as being **normalized**. Without this property, we could obtain probabilities greater than one by computing the probability of one of many events occurring.

Example: uniform distribution: $P(\mathrm{x} = x_i) = \frac{1}{k}$

Probability Density Function

- The domain of p must be the set of all possible states of x.
- $\forall x \in \mathrm{x}, p(x) \geq 0$. Note that we do not require $p(x) \leq 1$.
- $\int p(x)dx = 1$.

Example: uniform distribution: $u(x; a, b) = \frac{1}{b-a}$.

Computing Marginal Probability with the Sum Rule

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y). \tag{3.3}$$

$$p(x) = \int p(x, y) dy. \tag{3.4}$$

Conditional Probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}. \tag{3.5}$$

Chain Rule of Probability

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}). \tag{3.6}$$

Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x} = x, \mathrm{y} = y) = p(\mathrm{x} = x)p(\mathrm{y} = y). \tag{3.7}$$

Conditional Independence

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x} = x, \mathrm{y} = y \mid \mathrm{z} = z) = p(\mathrm{x} = x \mid \mathrm{z} = z)p(\mathrm{y} = y \mid \mathrm{z} = z). \tag{3.8}$$

Expectation

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x) f(x), \tag{3.9}$$

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x) f(x) dx. \tag{3.10}$$

linearity of expectations:

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{x}}[f(x)] + \beta \mathbb{E}_{\mathrm{x}}[g(x)], \tag{3.11}$$

Variance and Covariance

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right]. \tag{3.12}$$

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}\left[f(x)\right])\left(g(y) - \mathbb{E}\left[g(y)\right]\right)\right]. \tag{3.13}$$

Covariance matrix:

$$\mathrm{Cov}(\mathbf{x})_{i,j} = \mathrm{Cov}(\mathrm{x}_i, \mathrm{x}_j). \tag{3.14}$$

Bernoulli Distribution

$$P(\mathrm{x} = 1) = \phi \tag{3.16}$$

$$P(\mathrm{x} = 0) = 1 - \phi \tag{3.17}$$

$$P(\mathrm{x} = x) = \phi^{x}(1 - \phi)^{1-x} \tag{3.18}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi \tag{3.19}$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi) \tag{3.20}$$

Gaussian Distribution

Parametrized by variance:

$$\mathcal{N}(x; \mu, \sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{3.21}$$

Parametrized by precision:

$$\mathcal{N}(x; \mu, \beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x-\mu)^2\right). \tag{3.22}$$

Gaussian Distribution

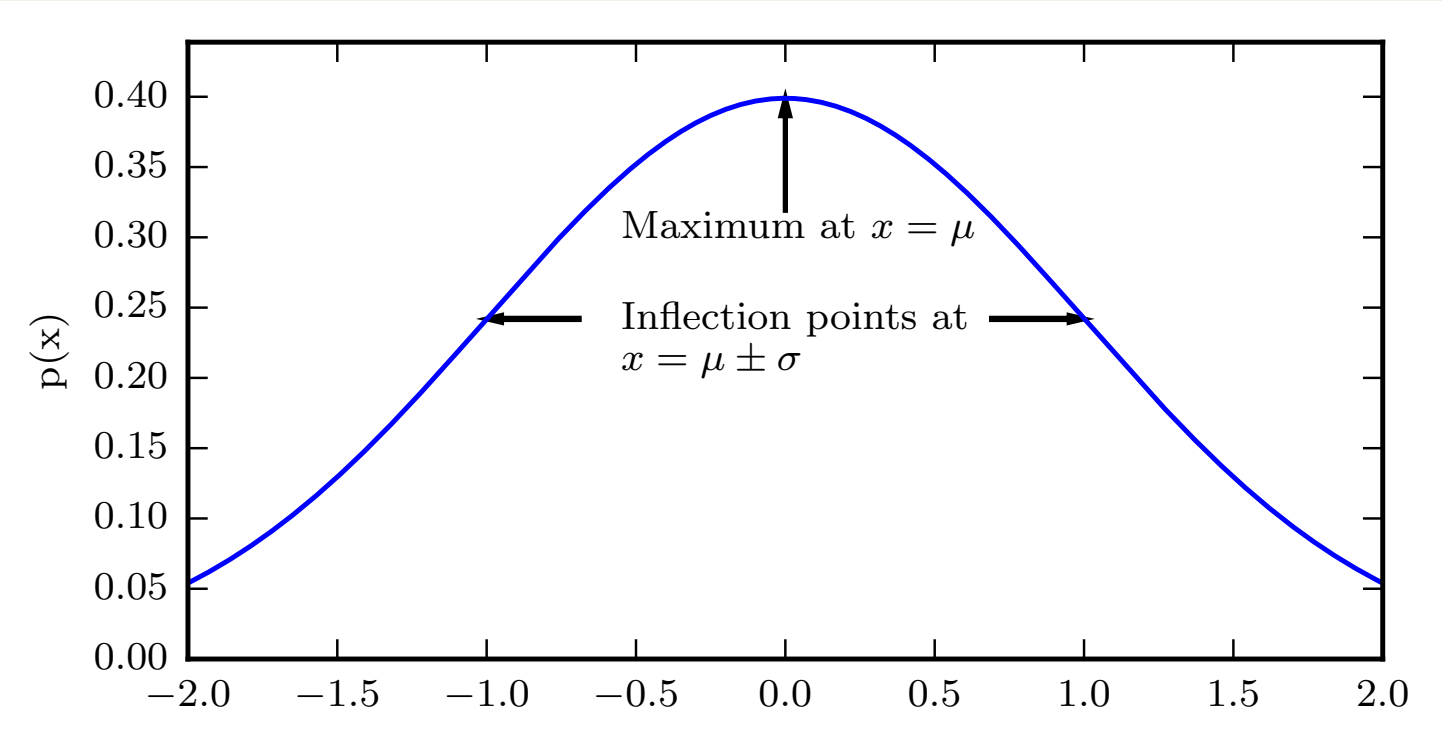


Figure 3.1

Multivariate Gaussian

Parametrized by covariance matrix:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^n \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3.23}$$

Parametrized by precision matrix:

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\beta}^{-1}) = \sqrt{\frac{\det(\boldsymbol{\beta})}{(2\pi)^n}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\beta}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{3.24}$$

More Distributions

Exponential:

$$p(x; \lambda) = \lambda \mathbf{1}_{x \geq 0} \exp(-\lambda x). \tag{3.25}$$

Laplace:

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x - \mu|}{\gamma}\right). \tag{3.26}$$

Dirac:

$$p(x) = \delta(x - \mu). \tag{3.27}$$

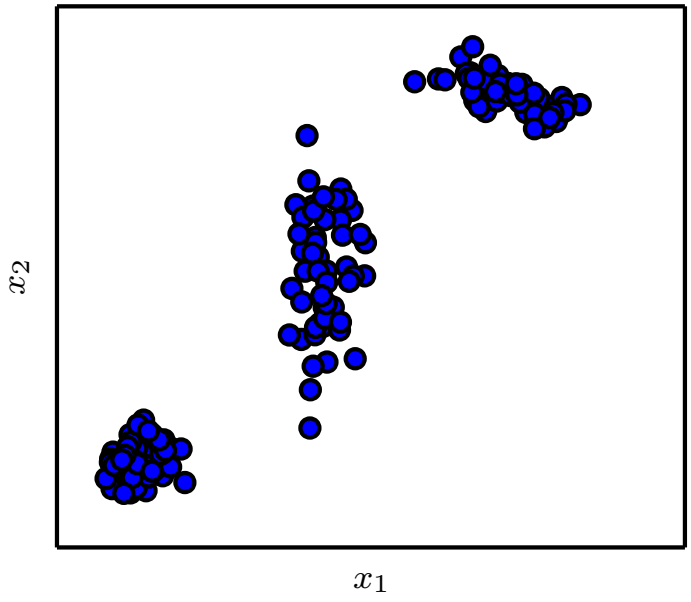
Empirical Distribution

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \delta(\boldsymbol{x} - \boldsymbol{x}^{(i)}) \tag{3.28}$$

Mixture Distributions

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i) \qquad (3.29)$$

Gaussian mixture with three components



Logistic Sigmoid

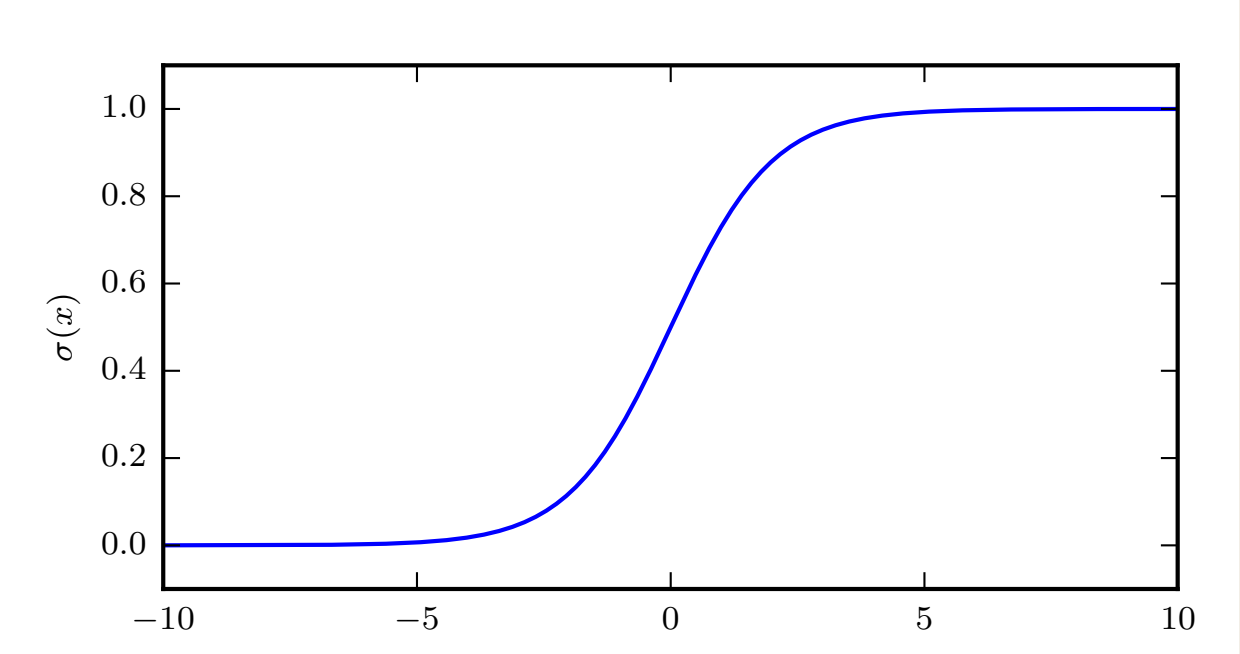


Figure 3.3: The logistic sigmoid function.

Commonly used to parametrize Bernoulli distributions

Softplus Function

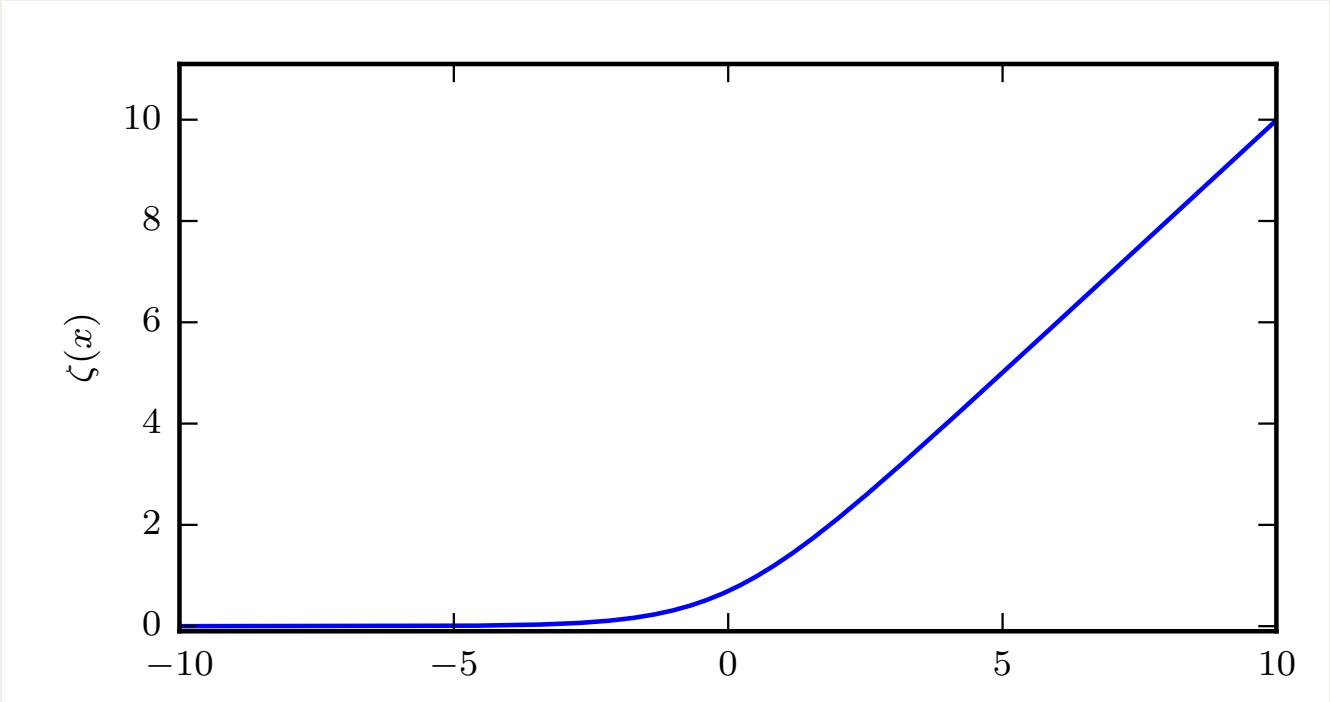


Figure 3.4: The softplus function.

Bayes' Rule

$$P(\mathrm{x} \mid \mathrm{y}) = \frac{P(\mathrm{x})P(\mathrm{y} \mid \mathrm{x})}{P(\mathrm{y})}. \tag{3.42}$$

Information Theory

Information:

$$I(x) = -\log P(x). \tag{3.48}$$

Entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]. \tag{3.49}$$

KL divergence:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]. \tag{3.50}$$

The KL Divergence is Asymmetric

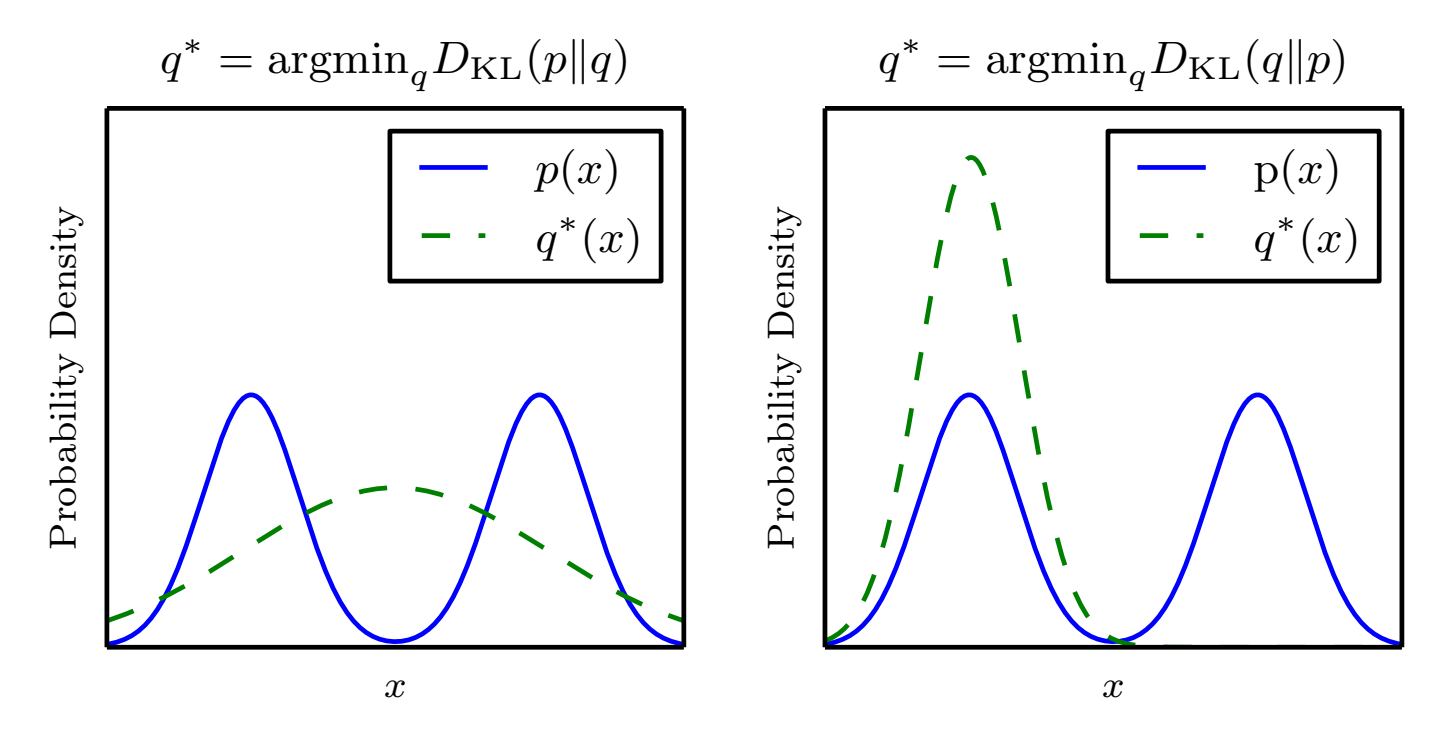


Figure 3.6

Numerical Computation

Overflow and Underflow

$$\mathrm{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}. \tag{4.1}$$

The exponentiation can underflow when the argument is very negative, or overflow when it is very positive.

Gradient Descent

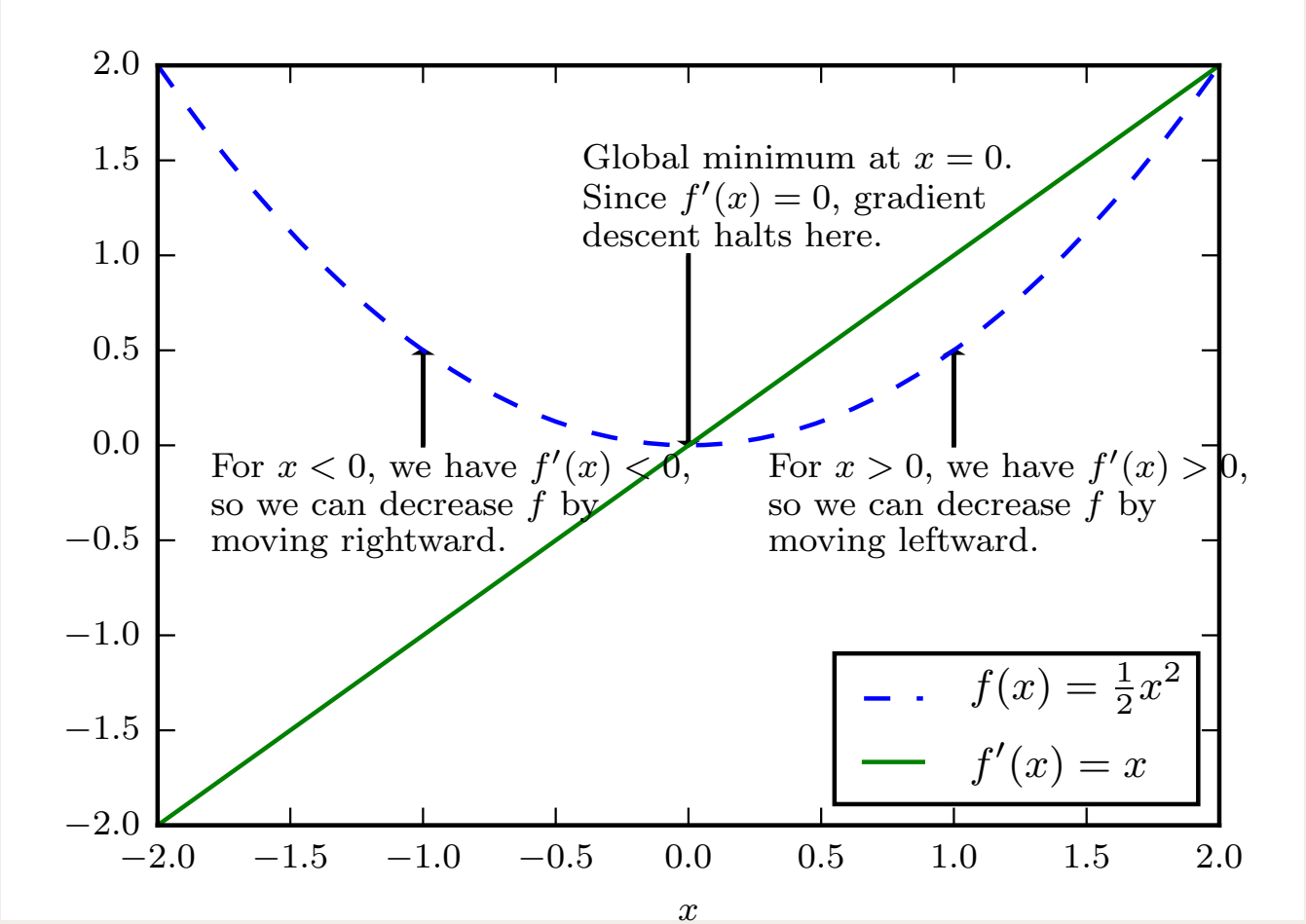


Figure 4.1

(Goodfellow 2016)

Approximate Optimization

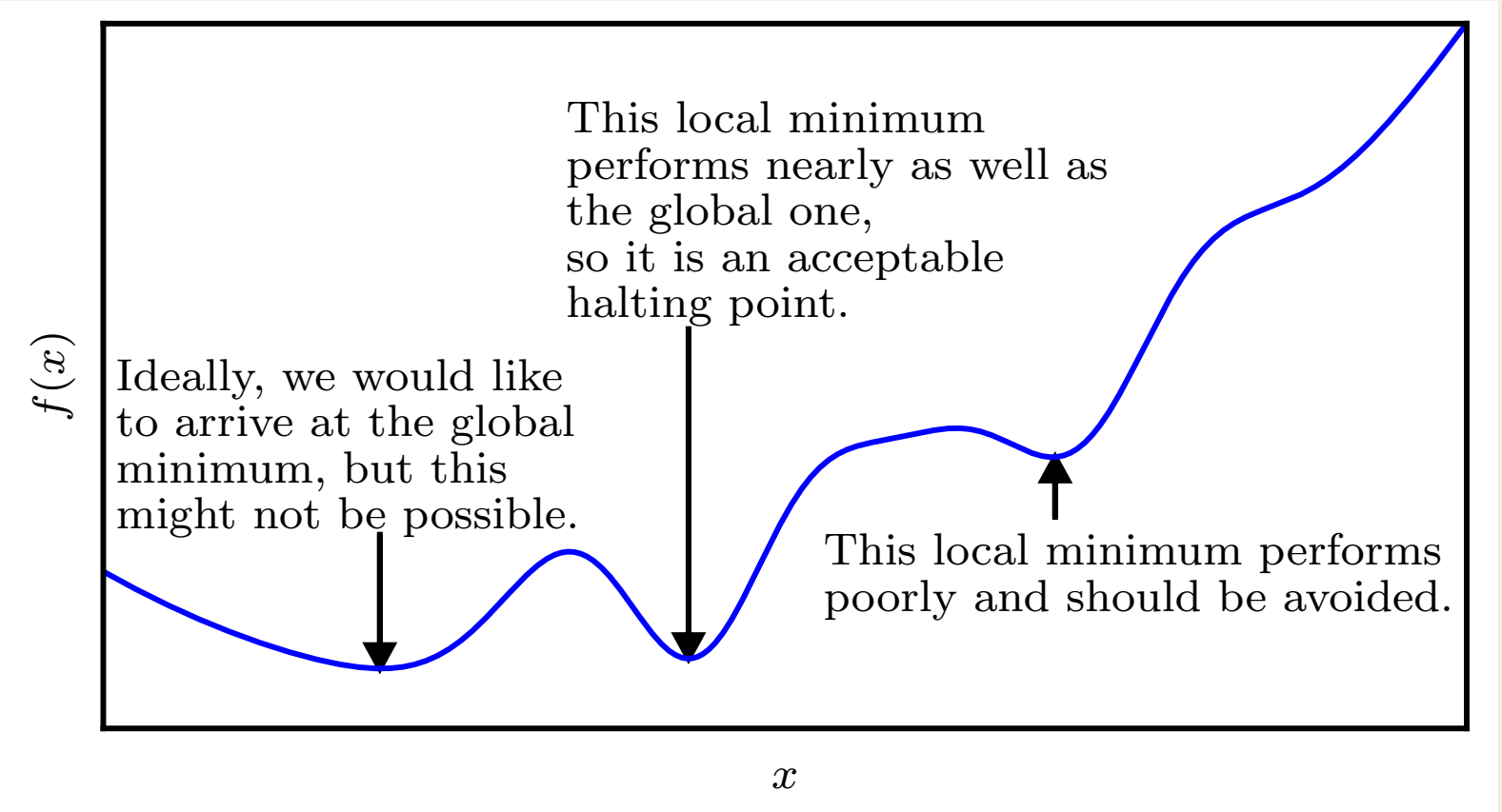


Figure 4.3

Saddle Points

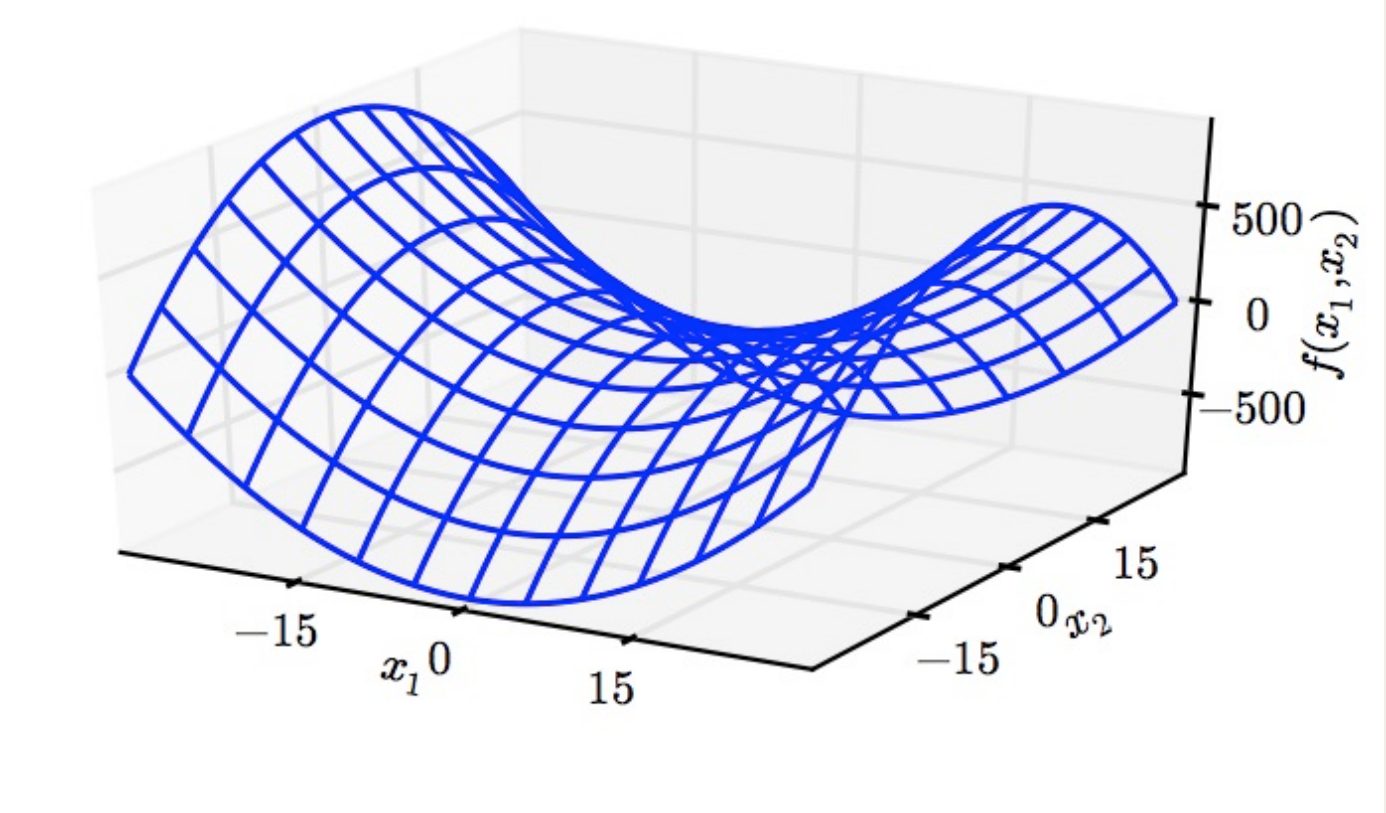


Figure 4.5 (Goodfellow 2016)

Gradient Descent and Poor Conditioning

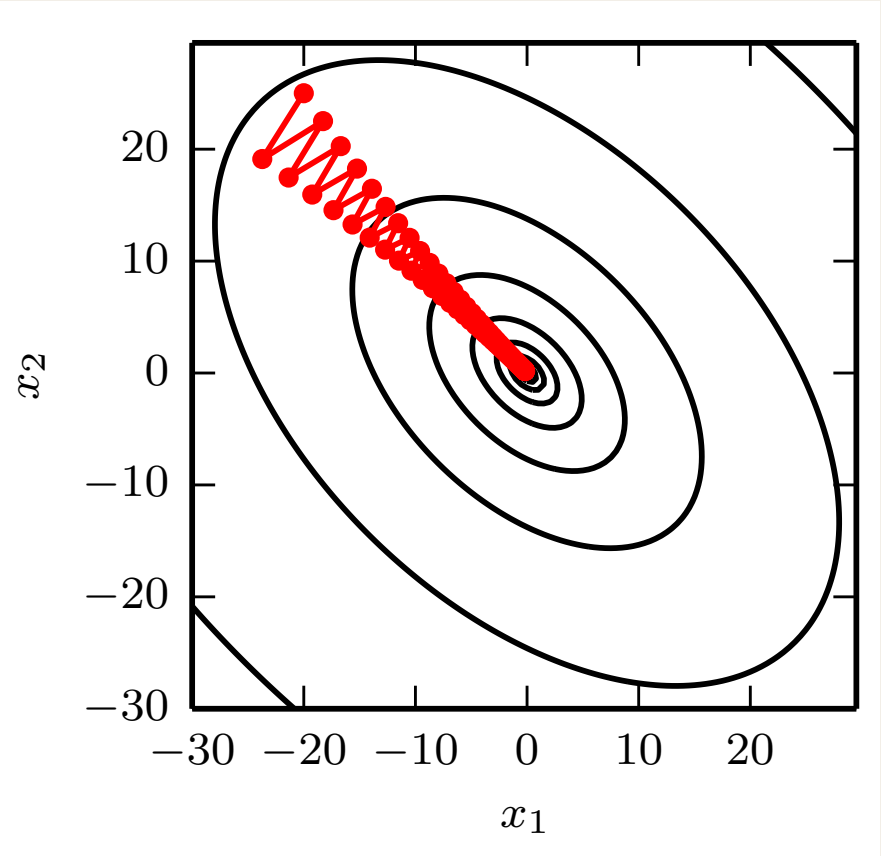


Figure 4.6

(Goodfellow 2016)

Backpropagation: a simple example

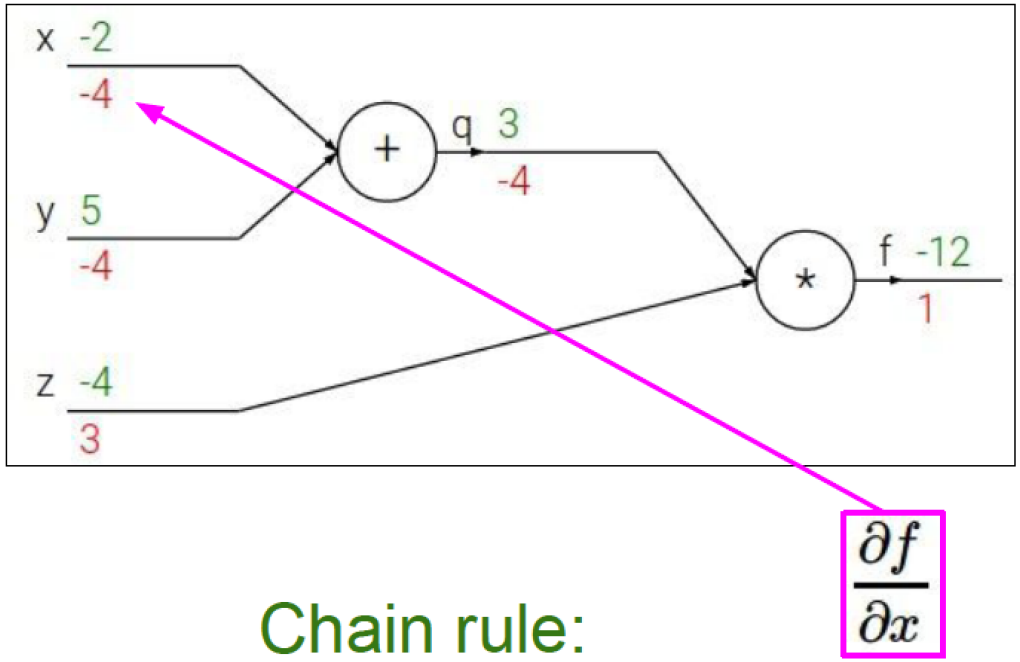
$$f(x, y, z) = (x + y)z$$

e.g. x = -2, y = 5, z = -4

$$q = x + y \qquad \frac{\partial q}{\partial x} = 1, \frac{\partial q}{\partial y} = 1$$

$$f = qz \qquad \frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$$

Want: $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$



Chain rule:

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x}$$

Upstream gradient

Local gradient

References

- *Deep Learning, www.deeplearningbook.org, Ian Goodfellow et. al.*
- *Lecture slides for Chapter 2, 3， 4 of Deep Learning, Ian Goodfellow, 2016-06-24*
- *Linear Algebra, Georgi Shilov, Publisher : Dover Publications Inc.; New edition (1 Mar. 1978), ISBN-10 : 048663518X, ISBN-13 : 978-0486635187*
- *Stanford CS class CS231n: Convolutional Neural Networks for Visual Recognition*