# 15) Tensor Operations and Stochastic Gradient Descent (SGD)

Vitor Kamada

August 2019

# Reference

Goodfellow et al. (2016): Ch 2, 4, and 8.

https://www.deeplearningbook.org/

Chollet (2018): Ch 2.

https://www.manning.com/books/deep-learning-with-python

# Tensor Rank (R)

## Scalar (0R)

$$s$$

## Vector (1R)

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_r \end{bmatrix}$$

## Matrix (2R)

$$M = \begin{bmatrix} m_{11} & m_{12} & \cdots & m_{1c} \\ m_{12} & m_{22} & \cdots & m_{2c} \\ \vdots & \vdots & \ddots & \vdots \\ m_{r1} & m_{r2} & \cdots & m_{rc} \end{bmatrix}$$

# Data Tensors

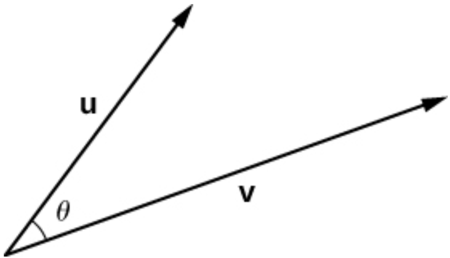
| | Cross-Sectional (2R) | Time Series (3R) |
|---|---|---|
| **Sample** | 100 Firms | 100 Firms |
| **Time** | | 10 years |
| **Features** | Sales, R&D, Size | Sales, R&D, Size |

| | Image (4R) | Video (5R) |
|---|---|---|
| **Sample** | 128 | 4 |
| **Frames** | | 240 |
| **Height** | 256 pixels | 256 pixels |
| **Width** | 256 pixels | 256 pixels |
| **Colors** | 3 | 3 |

# Dot Product

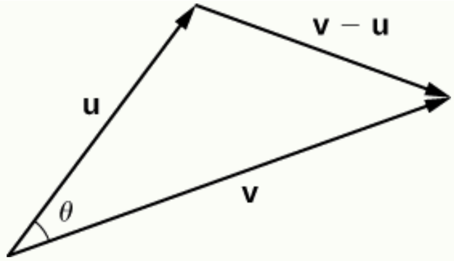
$$\vec{u} \cdot \vec{v} = u_1 v_1 + u_2 v_2$$

$$\vec{u} \cdot \vec{u} = ||\vec{u}||^2$$




$$\vec{u} \cdot \vec{v} = ||\vec{u}||||\vec{v}|| cos\theta$$

# Law of Cosines

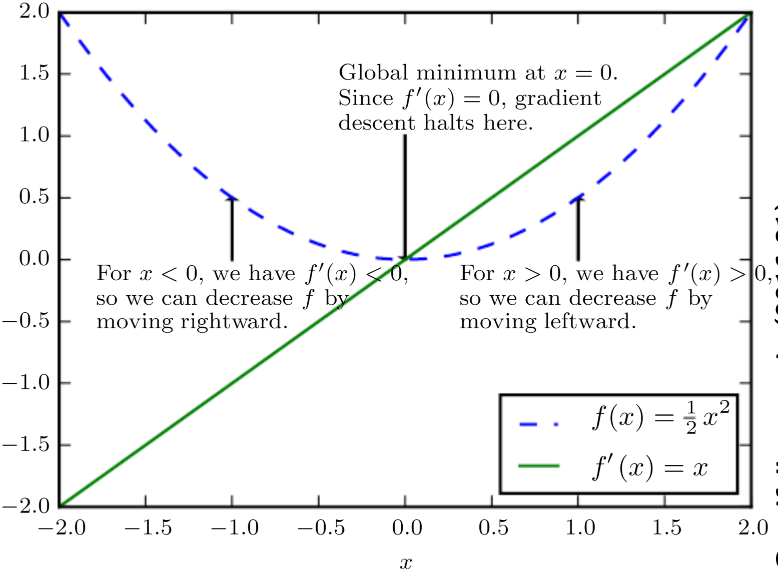



$$||\vec{u} - \vec{v}||^2 = ||\vec{u}||^2 + ||\vec{v}||^2 - 2||\vec{u}||||\vec{v}||cos\theta$$

$$||\vec{u} - \vec{v}||^2 = ||\vec{u}||^2 + ||\vec{v}||^2 - 2\vec{u} \cdot \vec{v}$$

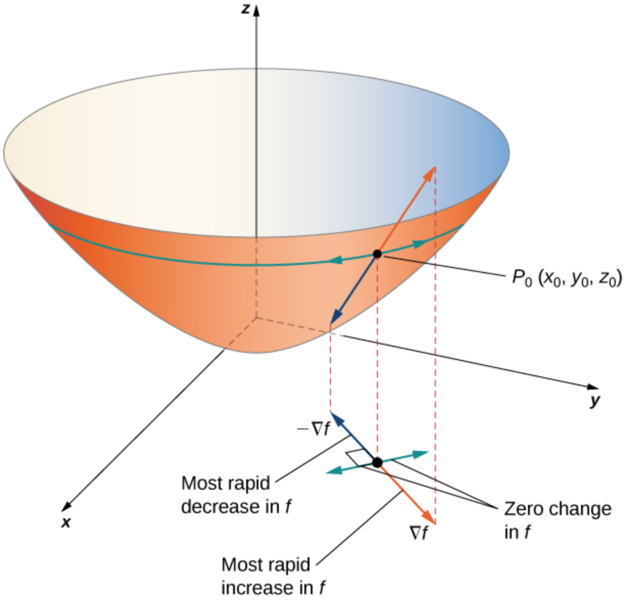
$$\vec{u} \cdot \vec{v} = ||\vec{u}||||\vec{v}||cos\theta$$

# Minimizing $f(x) = \frac{1}{2}x^2$




**Goodfellow et al. (2016:81)**

# Directional Derivative and Gradient




**Source:** LibreTexts (2019).
Calculus - Early Transcendentals (Stewart), Ch 4.1.

# Directional Derivatives

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

$$\hat{\mathbf{u}} = (cos\theta)\hat{\mathbf{i}} + (sin\theta)\hat{\mathbf{j}}$$

$$||\hat{\mathbf{u}}|| = [cos\theta]^2 + [sin\theta]^2 = 1$$

$$D_{\vec{u}} f(x, y) = \frac{\partial z}{\partial x} cos\theta + \frac{\partial z}{\partial y} sin\theta$$

# Gradient

$$\vec{\nabla} f(x, y) = \frac{\partial z}{\partial x}\widehat{\mathbf{i}} + \frac{\partial z}{\partial y}\widehat{\mathbf{j}}$$

$$D_{\vec{u}} f(x_0, y_0) = \vec{\nabla} f(x_0, y_0) \cdot \widehat{\mathbf{u}}$$

$$= || \vec{\nabla} f(x_0, y_0) |||| \widehat{\mathbf{u}} || cos\varphi$$

$$= || \vec{\nabla} f(x_0, y_0) || cos\varphi$$

# Gradient Descent

$$D_{\vec{u}} f(x_0, y_0) = ||\vec{\nabla} f(x_0, y_0)|| cos\varphi$$

If $\varphi = \pi$ then $cos\varphi = -1$

$\therefore \vec{\nabla} f(x_0, y_0)$ and $\widehat{\mathbf{u}}$ point in opposite directions

Min Value of $D_{\vec{u}} f(x_0, y_0)$ is $-||\vec{\nabla} f(x_0, y_0)||$

# Deterministic vs SGD

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, \theta)$$

$$\nabla_\theta J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L(x_i, y_i, \theta)$$

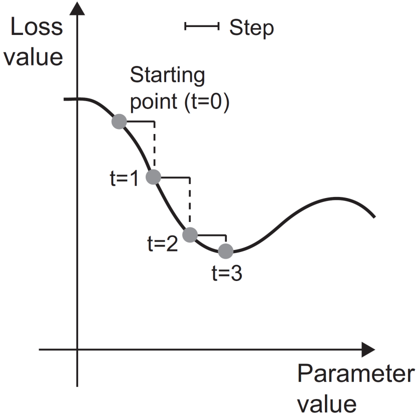
---

$m$: minibatch (power of 2)

$$g = \frac{1}{m} \sum_{i=1}^{m} \nabla_\theta L(x_i, y_i, \theta)$$

**Epoch**: each iteration over all the training data

# Stochastic Gradient Descent

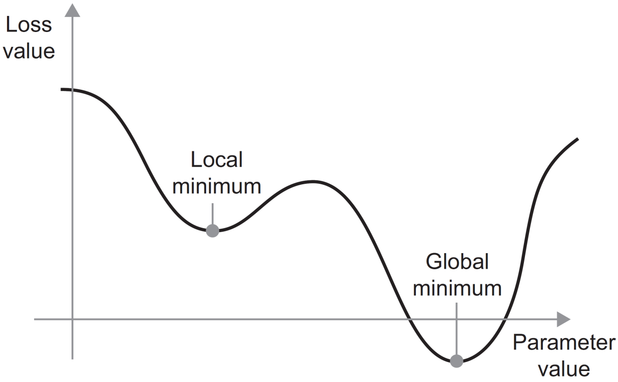



Chollet (2018:49)

$$\theta \leftarrow \theta - \epsilon g$$

$\epsilon$ : **learning rate or step**

# Momentum or Velocity

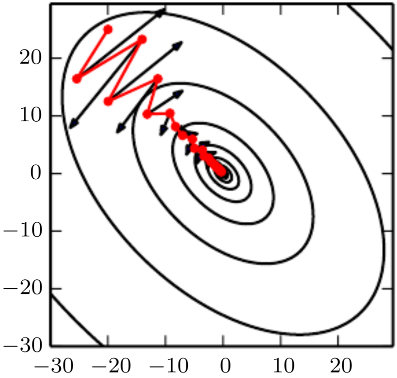



Chollet (2018:51)

$$v \leftarrow \alpha v - \epsilon g, \qquad \alpha \in [0, 1)$$

**$v$: Exponentially Decaying Average of the Gradient**

# Red Path Gradient with Momentum




Goodfellow et al. (2016:293)

**Black Path Gradient wastes time moving back and forth**

# Nesterov Momentum

$$g \leftarrow \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} L[(x_i; \theta), y_i]$$

---

$$g \leftarrow \nabla_\theta \frac{1}{m} \sum_{i=1}^{m} L[(x_i; \theta + \alpha v), y_i]$$

$$v \leftarrow \alpha v - \epsilon g$$

$$\theta \leftarrow \theta + v$$

# AdaGrad (Adaptive Gradient)

$$r \leftarrow r + g \odot g$$

$$\Delta\theta \leftarrow -\frac{\epsilon}{\delta+\sqrt{r}} \odot g$$

**Adapts learning rates ($\epsilon$), scaling them inversely proportional to past squared values of the gradient ($r$)**

$\delta = 10^{-7}$ for numerical stability

# RMSProp

$$r \leftarrow \rho r + (1-\rho) g \odot g$$

$$\Delta\theta = -\frac{\epsilon}{\sqrt{\delta + r}} \odot g$$

**Discard extreme past, using exponentially decaying average ($\rho$)**

$$\theta \leftarrow \theta + \Delta\theta$$

# RMSProp + Nesterov Momentum

$$\tilde{\theta} \leftarrow \theta + \alpha v$$

$$g \leftarrow \nabla_{\tilde{\theta}} \frac{1}{m} \sum_{i=1}^{m} L[f(x_i; \tilde{\theta}), y_i]$$

$$r \leftarrow \rho r + (1 - \rho) g \odot g$$

$$v \leftarrow \alpha v - \frac{\epsilon}{\sqrt{r}} \odot g$$

$$\theta \leftarrow \theta + v$$