# Practical Deep Neural Networks

## GPU computing perspective

Softmax Regression

Yuhuang Hu  Chu Kiong Loo

Advanced Robotic Lab
Department of Artificial Intelligence
Faculty of Computer Science & IT
University of Malaya

# Outline


1. Introduction
2. Logistic Regression
3. Softmax Regression
4. Stochastic Gradient Descent

# Outline

1 Introduction

2 Logistic Regression

3 Softmax Regression

4 Stochastic Gradient Descent

# Assumed Prerequisites

- ☆ Basic Linear Algebra (DL book chapter 2)
- ☆ Basic Probability and Information Theory (DL book chapter 3)
- ☆ Basic Numerical Computation (DL book chapter 4)
- ☆ Machine Learning Basics (DL book chapter 5)

# Suggested Readings

- UFLDL Tutorial: Logistic Regression, Softmax Regression and Stochastic Gradient Descent.
- CS231n: Linear classification: Support Vector Machine, Softmax and Optimization: Stochastic Gradient Descent.
- DL Book Chapter 4 Numerical Computation 4.3 DL Book Chapter 8 Numerical Optimization 8.3

# Outline

1 Introduction

2 Logistic Regression

3 Softmax Regression

4 Stochastic Gradient Descent

# Logistic Function

$$f(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{W}^\top \mathbf{x})}$$

$$p(y = 1|\mathbf{x}) = f(\mathbf{x})$$

$$p(y = 0|\mathbf{x}) = 1 - f(\mathbf{x})$$

# Logistic Regression — Binary Classifier

$$L(X, \mathbf{y}|\mathbf{W}) = -\frac{1}{N}\sum_i \left(y^i \log(P(y=1|\mathbf{x}^i)) + (1-y^i)\log(p(y=0|\mathbf{x}^i))\right)$$

$$W^\star = \arg\min_W L(X, \mathbf{y})$$

# Outline

1 Introduction

2 Logistic Regression

3 Softmax Regression

4 Stochastic Gradient Descent

# Extending Logistic Function — Softmax Function

$$P(y = k|\mathbf{x}; \mathbf{W}) = \frac{\exp(\mathbf{W}^{(k)\top}\mathbf{x})}{\sum_{j=1}^{K} \exp(\mathbf{W}^{(j)\top}\mathbf{x})}$$

# Softmax Regression — Multi-classes Classifier

$$L(X, \mathbf{y}|\mathbf{W}) = -\frac{1}{N}\sum_{i}\sum_{k=1}^{K}\mathbf{1}\{y^i = k\}\log P(y^i = k|\mathbf{x}^i, \mathbf{W})$$

$$\mathbf{W}^\star = \arg\min_{\mathbf{W}} L(X, \mathbf{y})$$

# Outline

1 Introduction

2 Logistic Regression

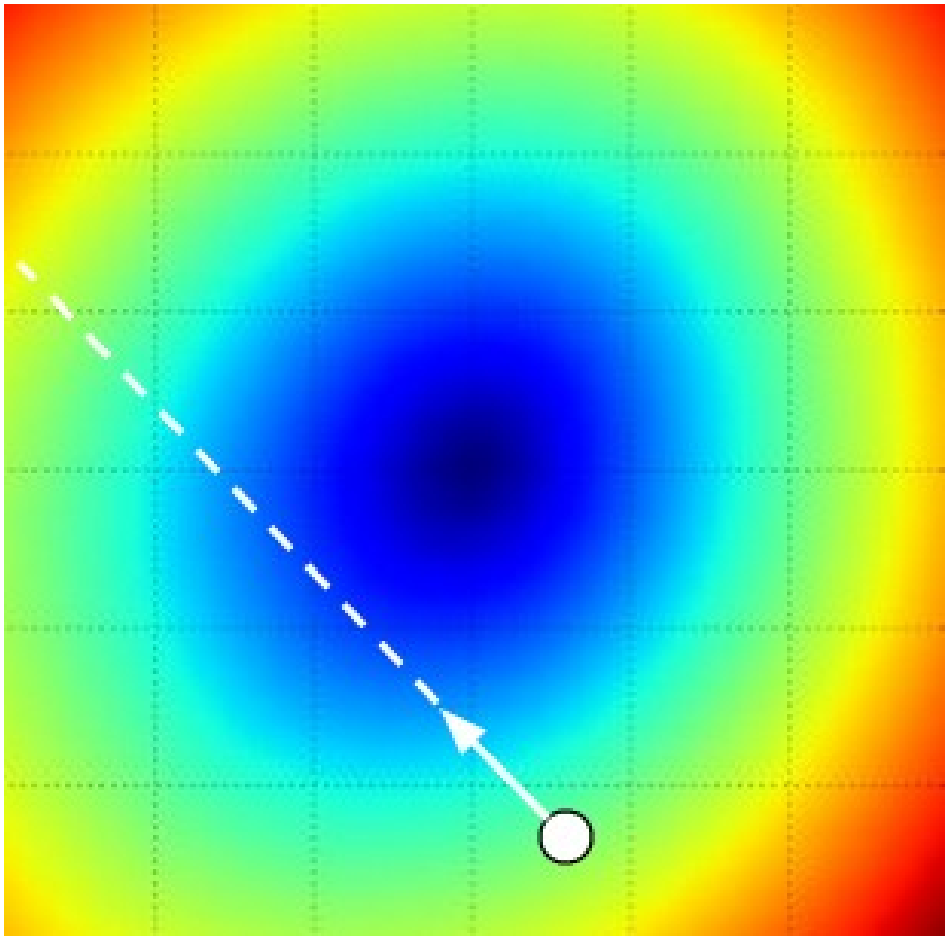
3 Softmax Regression

4 Stochastic Gradient Descent

# Cost

- Target: to learn some optimal parameters $\theta$
- Strategy: to minimize some cost $L$ respect to $\theta$
- Cost choice: cross-entropy cost, mean-squared error cost
- Solution: Gradient Descent!

# Gradient optimization



Demo 1; Demo 2

# Stochastic Gradient Descent (SGD)

$$\theta^* = \theta - \alpha \frac{\partial}{\partial \theta} L(\theta)$$

# Variants: momentum SGD

$$V^* = \mu V - \alpha \nabla L(\theta)$$
$$\theta^* = \theta + V^*$$

# Variants: Nesterov's Accelerated Gradient (NAG)

$$V^* = \mu V - \alpha \nabla L(\theta + \mu V)$$
$$\theta^* = \theta + V^*$$

# Q&A