Python for Big Data

Il-Youp Kwak, PhD

Problems with Big Data

- Data Processing
- Memory problems
  - Too big to load data
  - Modeling difficulties

> Python can handle

- **ids.txt 에 있는 사람들의 데이터만 new_data.txt에 저장하라**

data1.txt

ID	field1	field2	field3
Tom	1	3	1
John	1	3	3
Tom	1	1	1
Karl	1	1	2
Karl	1	3	2
Karl	2	2	2
Andy	1	2	3

data2.txt

ID	field1	field2	field3
Tom	2	3	1
Jane	1	1	3
Jane	2	2	1
Max	3	1	2
Karl	1	4	2
Max	3	2	2
Max	2	2	1

ids.txt

ID
John
Karl
Andy
Tim

(가정) data1.txt, data2.txt 가 너무 커서, 메모리에 읽을 수 없음

Solution

- **Process line by line (using python, perl, awk, etc)**

Open a file to write
Read line by line from data files
If current id is in the list, write a line in a file

We also have problems in modeling side

- How to handle big data in modeling?
- Where is bottle neck?
- Mostly, memory problems, why?

Ex) Binary Classification Model

- We are Classifying Cat or Dog $f : \mathbf{x} \xrightarrow{f_\theta} \mathbb{R}_{[0,1]}$
- Define a Loss function $L(\theta; X, \mathbf{y})$
- Minimize Loss w.r.t θ given data $argmin_\theta L(\theta; X, \mathbf{y})$
- Iterate gradient updates $\hat{\theta}^* = \hat{\theta_0} - \alpha * \frac{dL(\theta; X, \mathbf{y})}{d\theta}$
- $f_{\hat{\theta}}(\mathbf{x})$ is your classification model

Ex) Binary Classification Model

- Minimize Loss w.r.t θ given data $argmin_\theta L(\theta; X, \mathbf{y})$

  This part may suffer (memory problem)

- We can consider using mini-batch Gradient Descent, Stochastic Gradient Descent

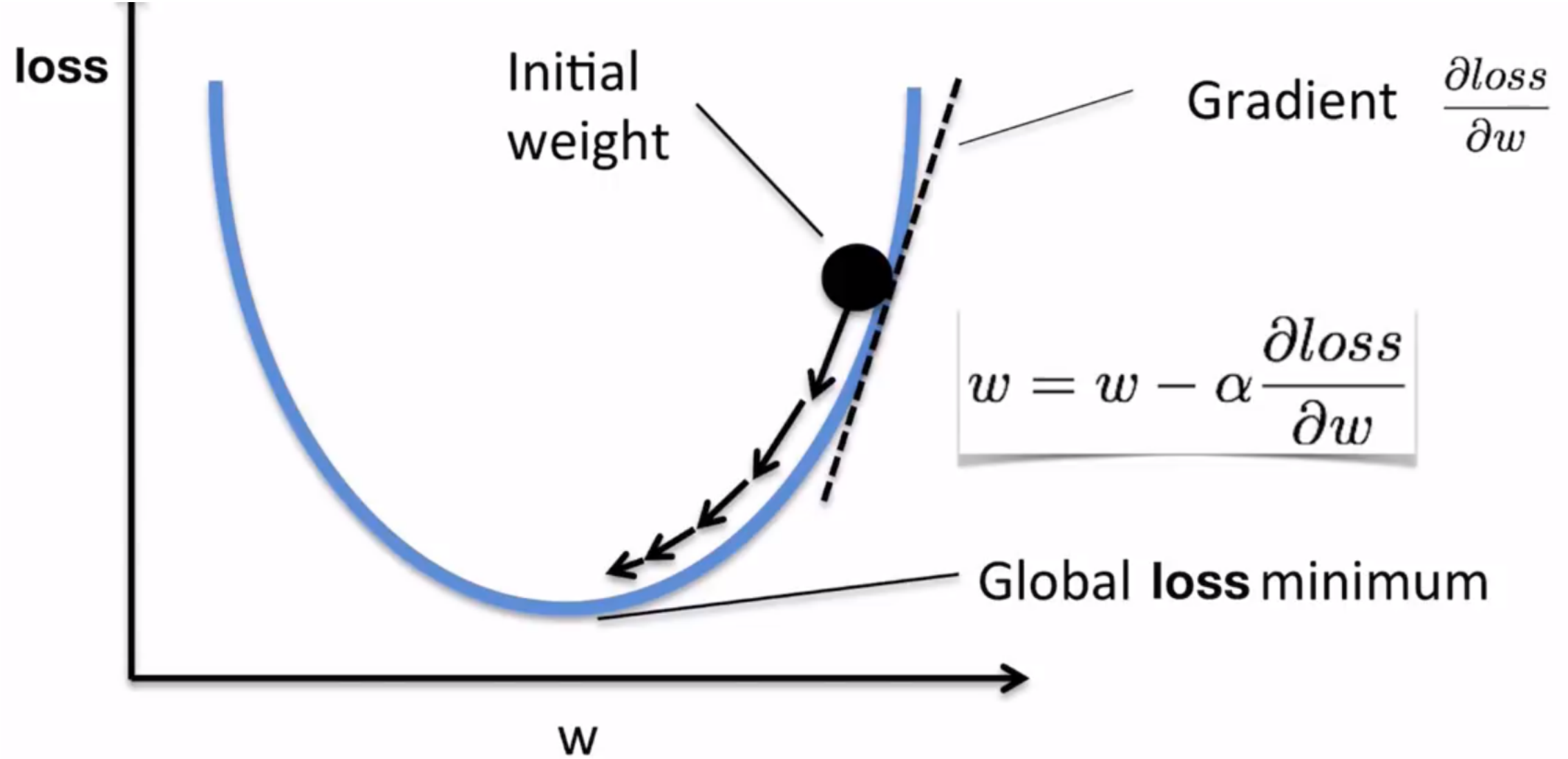
- We may need a data generator

Why Gradient Descent?

- We can handle memory side better
  (Control Batch Size)

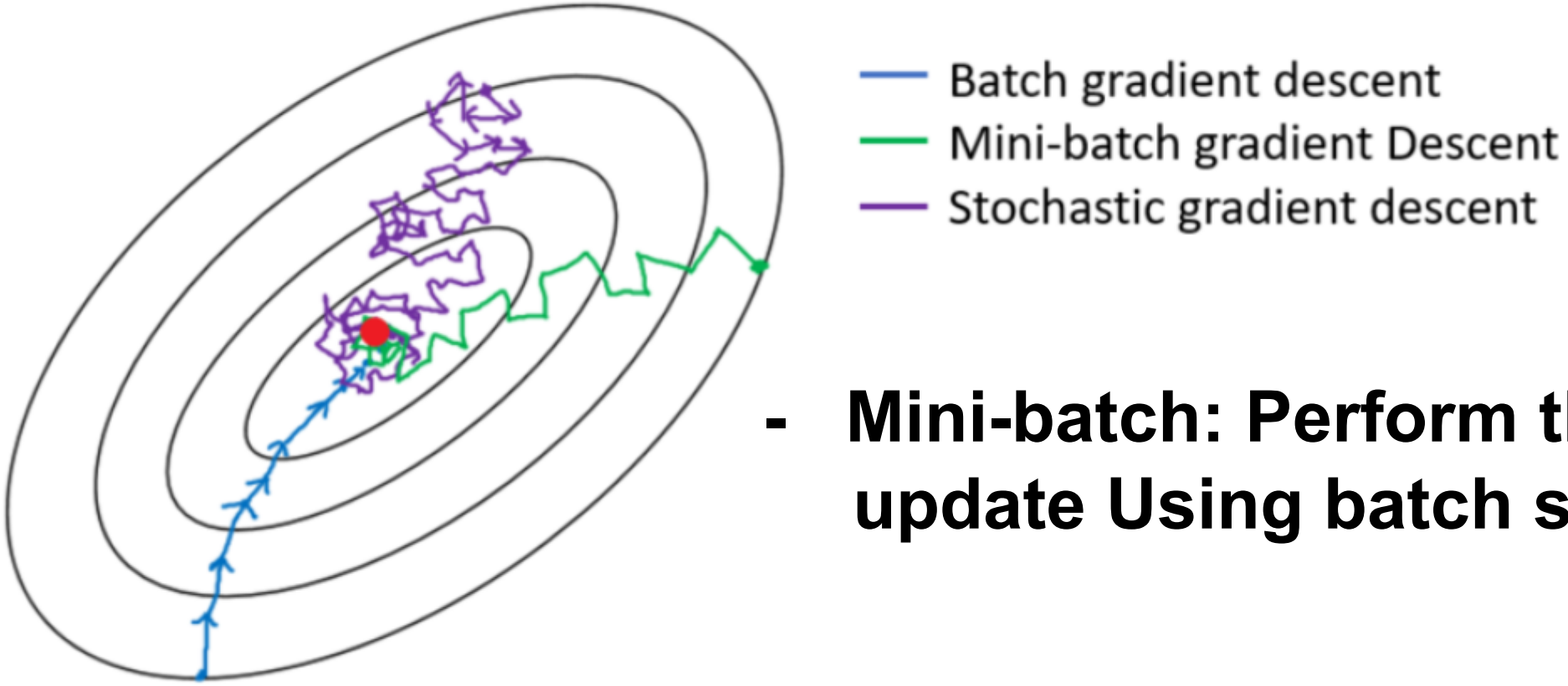
- We can cover online-learning side for streaming data

What is Gradient Descent?



What is Stochastic Gradient Descent?

- Perform the gradient update using each sample



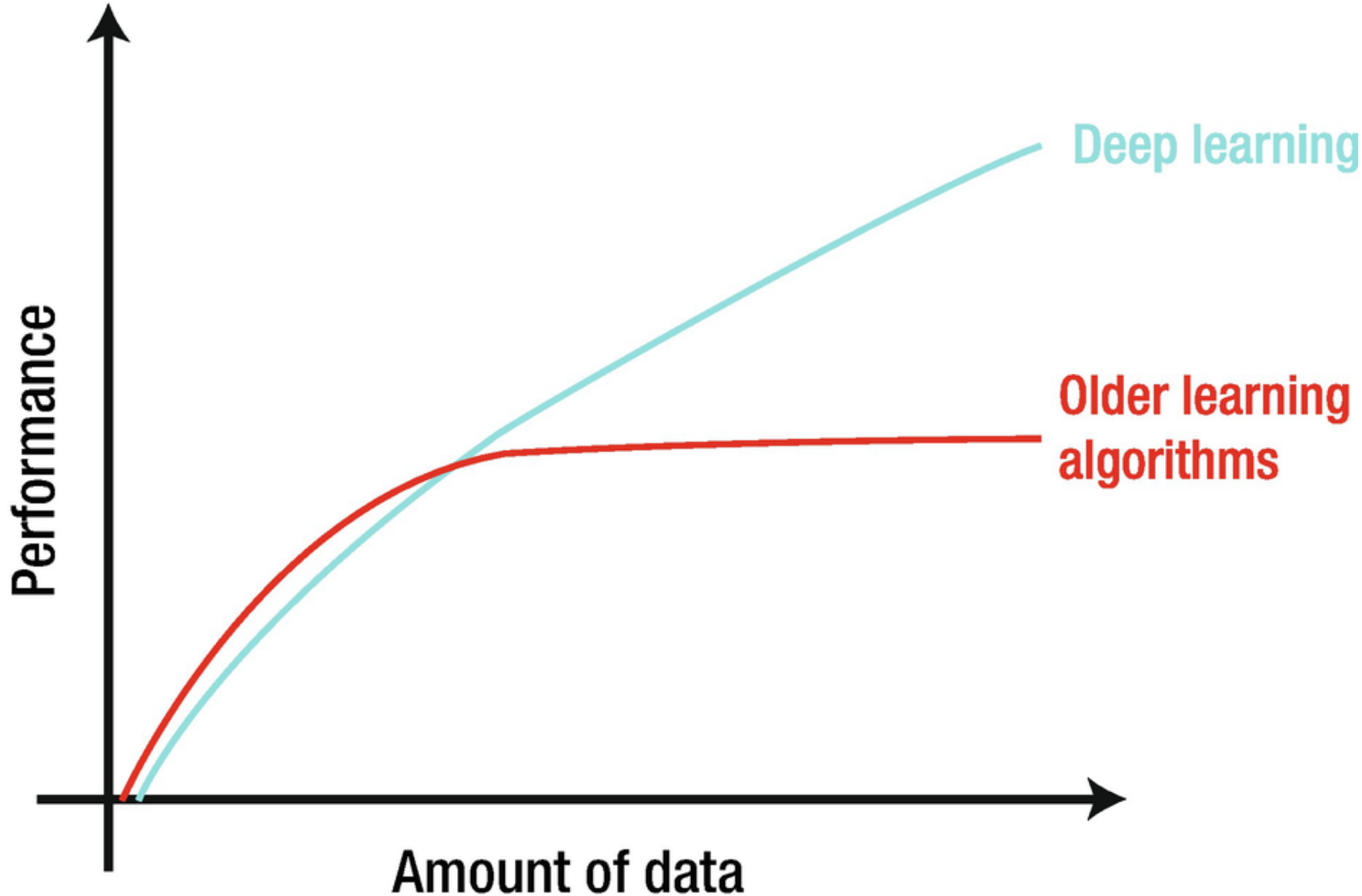
- Mini-batch: Perform the update Using batch samples

Why Generator?

- Do not load data in your memory

- In modeling, you can feed data from file using generator

Why deep learning?



Deep Learning Concepts

- Use all data to maximize performance
- Higher model complexity

  Ex) $y = f_1(f_2(f_3(f_4(f_5(x))))) \,, \quad f_i(x) = sigmoid(Wx + b)$
- Make a deep non-linear model that explain the structure of data the best
- Define Loss function and get optimized solution

Python Practice

https://wikidocs.net/book/1

HW 2 (1)

- R for Data Science 의 flight 데이터를 dplyr 과 ggplot 을 활용하여 시각화 시키고, 해당 그래프를 요약하는 한 문장의 설명을 달 것.
- 목표: 그림은 이쁘게, 설명은 그림을 명확하게 설명할 수 있도록

HW 2 (2)

- list compression 을 사용하여 1~100 사이의 수중 3의 배수를 나열하라

HW 2 (3) 아래 문제를 해결할 것

- ids.txt 에 있는 사람들의 데이터만 new_data.txt에 저장하라

data1.txt

ID	field1	field2	field3
Tom	1	3	1
John	1	3	3
Tom	1	1	1
Karl	1	1	2
Karl	1	3	2
Karl	2	2	2
Andy	1	2	3

data2.txt

ID	field1	field2	field3
Tom	2	3	1
Jane	1	1	3
Jane	2	2	1
Max	3	1	2
Karl	1	4	2
Max	3	2	2
Max	2	2	1

ids.txt

ID
John
Karl
Andy
Tim

(가정) data1.txt, data2.txt 가 너무 커서, 메모리에 읽을 수 없음

HW 2 (4)

- yield 를 사용한 Generator 를 만드는 연습을 해보기

Thank you!
Q & A