Chapter 4. Numerical Computation

4.1 Overflow and Underflow

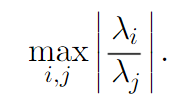
- Caused by Rounding Error

- /0, -, 🡪 Nan value ( Not a number value )

- near zero 🡪 zero

- Libraries are already solved these issues w/ several creativity ways

- Theano : SW packages which automatically detects and stabilizes these issues

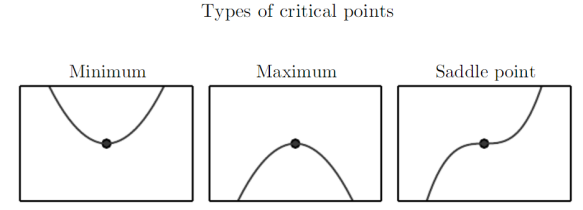
4.2 Poor Conditioning

- Condition Number :

- Big Condition number 🡪 big rounding error @ output

4.3 Gradient-Based Optimization

- objective function = criterion = cost function = loss function = error function

 - notation for value which minimize criterion : x∗= arg min f(x).

- critical point = stationary point

- local minimum / global minimum / saddle point

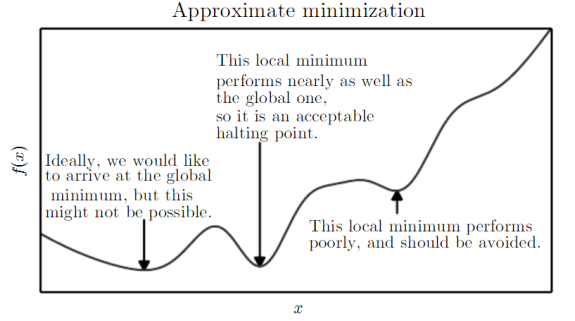
- Gradient Descent on 2D

f(x −e sign(f(x)))

- Gradient Descent on 3D or more(steepest descent)

x’= x −e∇xf(x)

- e : learning rate

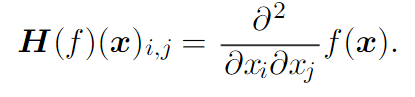
 - line search = strategy which swapping several e

4.3.1 Beyond the Gradient : Jacobian and Hessian Matricies

- Jacobian matrix : ex) partial derivatives of electric field



- Hessian matrix

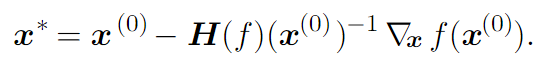


- if all second partial derivatives are cont, Hessian matrix is symmetric

- 🡪 can be decomposed into an orthogonal basis of eigenvectors

- Newton’s method (related to Limited-memory BFGS)

: using Taylor series until 2nd order



4.4 Constrained Optimization

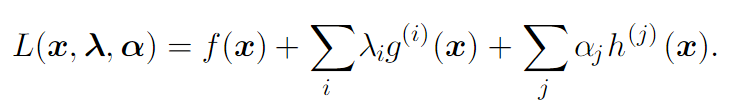
: Optimization in some set S

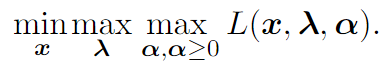
1. simple gradient descent, if the step goes out of S, project again into S

2. design unconstrained optimization problem whose solution can solve original.

ex) changing a system of coordinates

3. KKT ( Karush-Kuhn-Tucker) approach





4.5 Example : Linear Least Squares

Chapter 5. Machine Learning Basics

5.1 Learning Algorithms

5.1.1 The Task, T

- Classification / Classification w/ missing inputs / Regression / Transcription

/ Machine translation / Structured output / Anomaly detection

/ Synthesis and sampling / Imputation of missing values / Denoising

/ Density estimation or probability mass function estimation

5.1.2 The Performance Measure, P

: Cost function

5.1.3 The Experience, E

- supervised / unsupervised learning

5.1.4 Example : Linear Regression

5.2 Capacity, Overfitting and Underfitting

- training error / generalization error ( = test error )

- data generating process : probability distribution which generates train and test data

- i.i.d : Independent from each other, and that the train and test set are Identically

Distributed

- Capacity : ability to fit a wide variety of function. (ex. # of parameters )

low capacity 🡪 struggle to fit training set

high capacity 🡪 overfitting

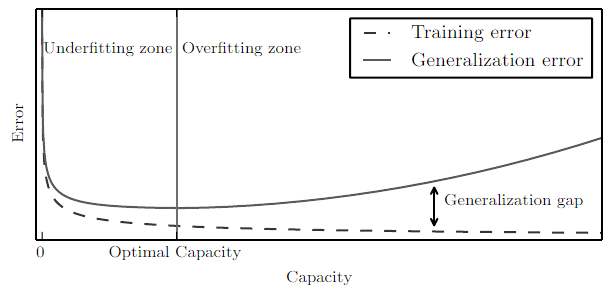
Occam’s razor : If observations are equally well, then choose “simplest” one

VC dimension ( Vapnik-Chervonenkis dimension ) : Quantitative expression of

capacity, but not practical

- Bayes error : minimum error that happens even using “oracle” that know exactly

probability distribution for generating data



5.2.1 The No Free Lunch Theorem

- “averaged over all possible data generating distributions, no algorithm is better

than others when classifying previously unobserved points.”

🡪 we should focus on the kinds of data generating distributions we care about

5.2.2 Regularization

: ways of expressing preferences on learning algorithm

: any modification we make to a learning algorithm that is intended to reduce its generalization error

ex) weight decay to make small lambda ( prevent overfitting )



5.3 Hyperparameters and Validation Sets

- Validation sets

: subset of training set to determine hyperparameter

- hyperparameter

: which control the behavior of the learning algorithm

: too difficult to train or usually control capacity

- if training hyperparameter as normal parameter, hyperparameter goes

maximum, and makes overfitting

5.3.1 Cross-Validation

: when dataset is too small

: way to approximating generalization error at the price of

increased computational cost

- k-fold cross-validation / double k-fold cross-validation algorithm

**5.4 Estimators, Bias and Variance**

5.4.1 Point Estimation

- point estimator = statistic

: predicted value as best ( but maybe not true )

- point estimator : Best function between input and output

5.4.2 Bias

5.4.3 Variance and Standard Error

5.4.4 Trading off Bias and Variance to Minimize Mean Squared Error

5.4.5 Consistency

5.5 Maximum Likeliyhood Estimation ( **MLE** )