# MM Optimization Algorithms

**Chathuranga Weeraddana**

April 2022

# Lecture 1: Introduction

# Course Information

- Examiner: Carlo Fischione (carlofi@kth.se)
- Instructor: Chathuranga Weeraddanana (chatw@kth.se)
- Lectures: Wednesday [1] 13:00-15:00 CET, $7$ weeks

[1] There is one exception. See the course webpage.

# Course Information

- Course Website:
  - `https://chathurangaw.staff.uom.lk/files/KTH/courseinfo.html`
- Textbooks:
  - Kenneth Lange, *MM Optimization Algorithms*
- Evaluation:
  - based on homeworks + take home exam + mini project
  - Grade: binary

# Course Information

- Any other related information:
  - contact Carlo or myself

# My Sincere Gratitude

- to Prof. Kenneth Lange (Computational Genetics at UCLA)
  - for sharing some recently updated materials
  - they were very useful when preparing the slides

# History

- roots trace back to
  - A.G. McKendrick (1926, epidemiology)
  - F. Yates (1934, multiple classification)
  - E. Weiszfeld (1937, facilities location)
  - C.A.B. Smith (1957, gene counting)
  - H.O. Hartley (1958, EM algorithms)
- J.M. Ortega & W.C. Rheinboldt (1970, enunciation)
- J.D Leeuw (1977, multidimensional scaling)
- A.P. Dempster et al. (1977, EM algorithms)
- H. Voss and U. Eckhardt (1980, a firm theoretical foundation)

# MM Optimization Algorithms
# Application Domains

# Applicaton Domains

- logistic regression
- quantile regression
- discriminant analysis
- factor analysis
- matrix completion
- image restoration
- DC programming
- signomial programming
- many others

# Problem

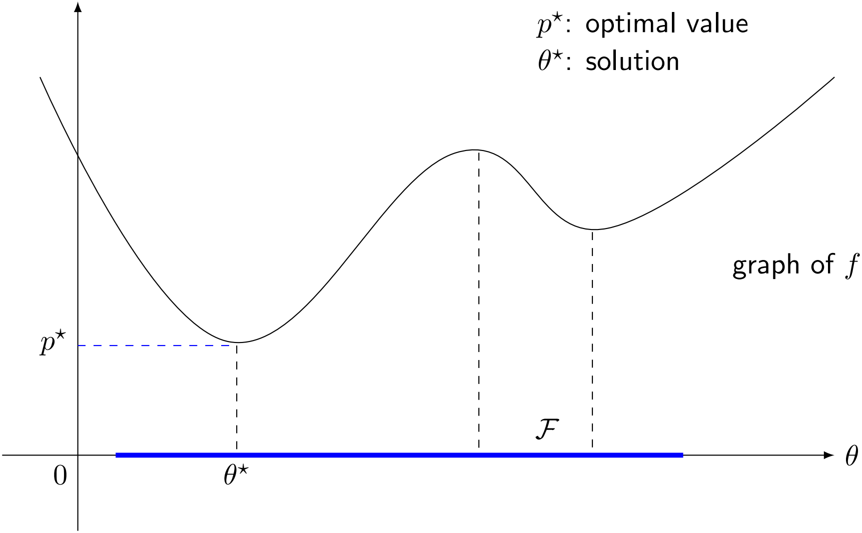
- a general formulation of an optimization problem [2]

$$\begin{array}{ll} \underset{\theta}{\text{minimize}} & f(\theta) \\ \text{subject to} & \theta \in \mathcal{F} \end{array}$$

- the decision variable is $\theta$
- $f$ and $\mathcal{F}$ depend on the application
- $f$ encodes what we want to optimize
- $\mathcal{F}$ encodes the underlying constraints

[2] see under the *Additional Reading*: *A Brief on Optimization*.

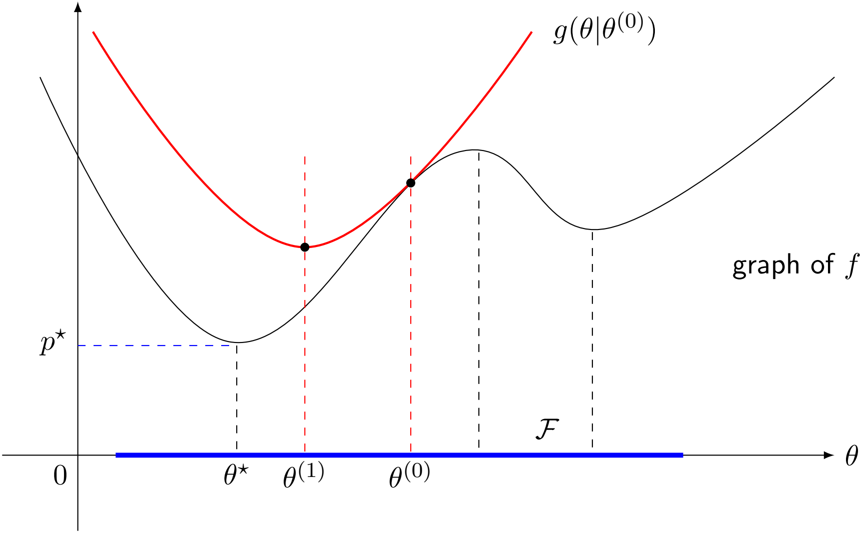
# Geometric Interpretation



# What is MM?

- MM stands for
  - majorize and minimize in a minimization problem
  - minorize and maximize in a maximization problem

# Majorize and Minimize

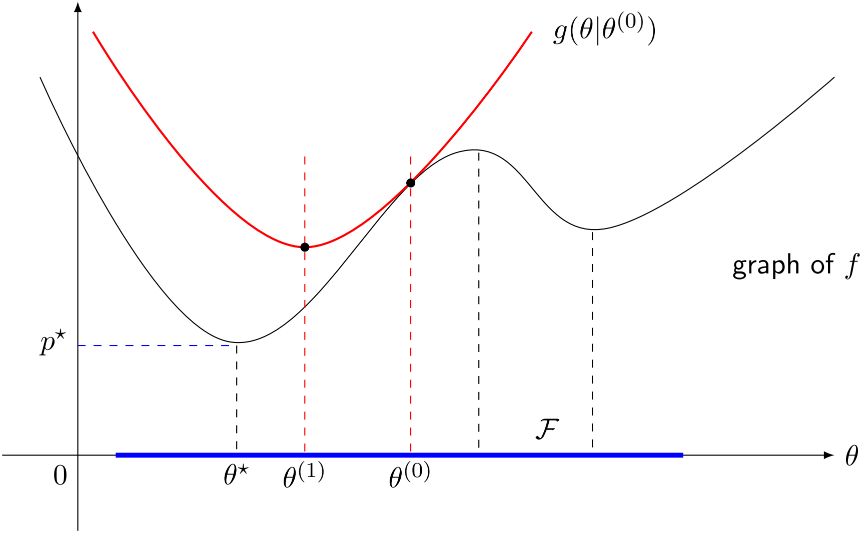


# The MM Principle

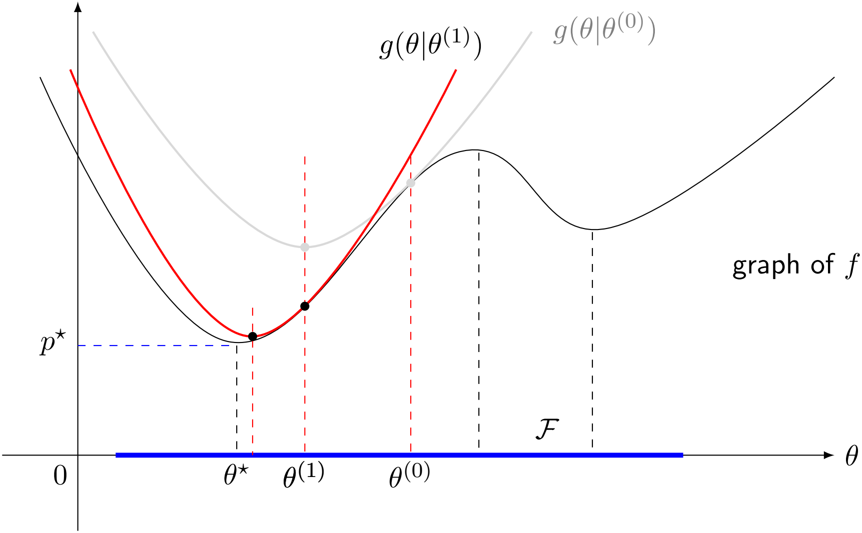
- is not an algorithm
- a useful principle for constructing optimization algorithms
- the resulting algorithms are called MM algorithms
  - majorize and minimize in an iterative mannar

# The MM Algorithm
# A Geometric Interpretation

# Geometric Interpretation



# Geometric Interpretation



# The MM Algorithm: Key Idea

- majorize and minimize in an iterative mannar

# Minorize and Maximize

- applied for maximization problems in a similar mannar

# Why MM Algorithms?

- MM principle simplifies optimization by
  - separating the variables of a problem
  - avoiding large matrix inversions
  - restoring the symmetry
  - turning a non-smooth problem into a smooth problem

# Some Notation and Definitions

# Majorization Function

- $g\big(\theta|\theta^{(n)}\big)$ is said to majorize $f(\theta)$ at $\theta^{(n)}$ provided

$$f\big(\theta^{(n)}\big) = g\big(\theta^{(n)}|\theta^{(n)}\big) : \qquad \text{tangency at } \theta^{(n)}$$

$$f(\theta) \leq g\big(\theta|\theta^{(n)}\big) \ \text{ for all } \theta : \quad \text{domination}$$

- $g\big(\ \cdot\ |\theta^{(n)}\big)$ is a majorization function of $f(\cdot)$ at $\theta^{(n)}$

# Majorization Function

- majorization relation between functions is closed under
  - sums
  - nonnegative products
  - limits
  - composition with an increasing function

# Minorization Function

- $g\big(\ \cdot\ |\theta^{(n)}\big)$ is a minorization function of $f(\cdot)$ at $\theta^{(n)}$ when
  - $-g\big(\theta|\theta^{(n)}\big)$ majorizes $-f(\theta)$ at $\theta^{(n)}$

# The MM Algorithm

# MM Algorithm

**Algorithm 1** MM Algorithm

**Input:** $\theta^{(0)} \in \mathcal{F}$, $n = 0$

1: Compute $g\big( \cdot \, | \theta^{(n)} \big)$

2: $\theta^{(n+1)} = \underset{\theta \in \mathcal{F}}{\arg\min} \quad g\big( \theta | \theta^{(n)} \big)$

3: $n := n + 1$ and go to step 1

# Descent Property

- MM (minimize/majorize) algorithm is a descent algorithm
- i.e., $f\left(\theta^{(n+1)}\right) \leq f\left(\theta^{(n)}\right)$ for all $n \in \mathbb{Z}$
- simple to verify the descent property

$$
\begin{aligned}
f\left(\theta^{(n+1)}\right) &\leq \inf_{\theta \in \mathcal{F}} \; g\left(\theta|\theta^{(n)}\right) && (1)\\
&\leq g\left(\theta^{(n)}|\theta^{(n)}\right) && (2)\\
&= f\left(\theta^{(n)}\right) && (3)
\end{aligned}
$$

# Some Common Tricks with Convexity and Lipschitz Continuity

# Affine Lower Bound

- suppose $f$ is convex and differentiable
- then we have

$$\begin{aligned} f(\theta) &\geq f\big(\theta^{(n)}\big) + \nabla f\big(\theta^{(n)}\big)^{\mathsf{T}}\big(\theta - \theta^{(n)}\big) \\ &= g\big(\theta|\theta^{(n)}\big) \end{aligned}$$

- $g\big(\theta|\theta^{(n)}\big)$ minorizes $f(\theta)$ at $\theta^{(n)}$
- e.g., $f(\theta) = -\log\ \theta \geq -\log\theta^{(n)} - \big(1/\theta^{(n)}\big)\big(\theta - \theta^{(n)}\big)$

# Jensen's Inequality

- suppose $f$ is convex
- then we have

$$f\big(\alpha x + (1-\alpha)y\big) \leq \alpha f(x) + (1-\alpha) f(y), \quad \alpha \in [0,1]$$

# Jensen's Inequality

- let $u, v > 0$ and let

$$\alpha = \frac{u^{(n)}}{u^{(n)} + v^{(n)}}, \quad x = \frac{u^{(n)} + v^{(n)}}{u^{(n)}}\ u, \quad y = \frac{u^{(n)} + v^{(n)}}{v^{(n)}}\ v$$

- thus, from the Jennsen's inequality, we get

$$f\left(u+v\right) \leq \frac{u^{(n)}}{u^{(n)}+v^{(n)}} f\left(\frac{u^{(n)}+v^{(n)}}{u^{(n)}}\ u\right) + \frac{v^{(n)}}{u^{(n)}+v^{(n)}} f\left(\frac{u^{(n)}+v^{(n)}}{v^{(n)}}\ v\right)$$

- $u$ and $v$ can be positive functions of $\theta$, e.g., $u(\theta)$ and $v(\theta)$

# Jensen's Inequality

- i.e.,

$$f\big(u(\theta)+v(\theta)\big) \le \frac{u(\theta^{(n)})}{u(\theta^{(n)})+v(\theta^{(n)})} f\left(\frac{u(\theta^{(n)})+v(\theta^{(n)})}{u(\theta^{(n)})}\ u(\theta)\right) + \frac{v(\theta^{(n)})}{u(\theta^{(n)})+v(\theta^{(n)})} f\left(\frac{u(\theta^{(n)})+v(\theta^{(n)})}{v(\theta^{(n)})}\ v(\theta)\right) = g(\theta|\theta^{(n)})$$

- $g\big(\theta|\theta^{(n)}\big)$ majorizes $f\big(u(\theta)+v(\theta)\big)$ at $\theta^{(n)}$
- e.g., $f(\theta) = -\log\ \theta =$ ?

# Quadratic Upper Bound

- suppose $f$ is twice differentiable and gradient Lipschitz continuous [3], i.e.,

$$\|\nabla f(\theta) - \nabla f(\beta)\|_2 \leq L\|\theta - \beta\|_2 \quad \text{for all } \theta, \beta$$

- then we have

$$\begin{aligned} f(\theta) &\leq f(\theta^{(n)}) + \nabla f(\theta^{(n)})^{\mathsf{T}}(\theta - \theta^{(n)}) + \frac{L}{2}\|\theta - \theta^{(n)}\|_2^2 \\ &= g\big(\theta|\theta^{(n)}\big) \end{aligned}$$

- $g\big(\theta|\theta^{(n)}\big)$ majorizes $f(\theta)$ at $\theta^{(n)}$

- e.g.,. $\cos\ \theta \leq \cos\theta^{(n)} - (\sin\theta^{(n)})(\theta - \theta^{(n)}) + (1/2)(\theta - \theta^{(n)})^2$

[3] The following condition is equivalent to a bound on the Hessian $\nabla^2 f(\theta)$ of $f$. For example, $LI - \nabla^2 f(\theta) \succeq 0$ is positive semidefinite ($LI - \nabla^2 f(\theta) \succeq 0$).

# Some Related MM Examples

# Minimize $\cos\theta$

- $\cos(\cdot)$ is twice differentiable and gradient Lipschitz continuous with constant $1$
- i.e.,

$$\begin{aligned} f(\theta) &= \cos\ \theta \\ &\le \cos\theta^{(n)} - (\sin\theta^{(n)})(\theta - \theta^{(n)}) + (1/2)(\theta - \theta^{(n)})^2 \\ &= g(\theta|\theta^{(n)}) \end{aligned}$$

- minimize the majorization function $g(\ \cdot\ |\theta^{(n)})$
- thus we have

$$\theta^{(n+1)} = \theta^{(n)} + \sin\theta^{(n)}$$

# Bradley–Terry Model

- prob. model: predicts the outcome of a paired comparison
- let us consider a sports league with $m$ teams
- $i$th team's skill level is parameterized by $\theta_i$, $i = 1, \ldots, m$
- probability that $i$ beats $j$ is given by

$$p_{ij}(\theta) = \frac{\theta_i}{\theta_i + \theta_j}$$

# Bradley–Terry Model

- let $b_{ij}$ be the number of times $i$ has beaten $j$ (data)
- ML estimate [4] of the model parameters $\theta \in \mathbb{R}^m_{++}$?
- the likelihood function of data has the form

$$p_\theta(b) = \prod_{i,j} (p_{ij}(\theta))^{b_{ij}}$$

- the log-likelihood function $f(\theta) = \log\ p_\theta(b)$
- the log-likelihood function $f$ should be maximized over $\theta$

[4] For a concise description of ML estimation, see § 7.1.1 *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004.

# Bradley–Terry Model

- let us find a minorization function:

$$
\begin{aligned}
f(\theta) = \log\ p_\theta(b) &= \log\ \prod_{i,j}\ (p_{ij}(\theta))^{b_{ij}} \\
&= \sum_{i,j}\ b_{ij} \log\ \left(\frac{\theta_i}{\theta_i + \theta_j}\right) \\
&= \sum_{i,j}\ b_{ij} \left[\log\ \theta_i - \log\ (\theta_i + \theta_j)\right] \\
&\geq \sum_{i,j}\ b_{ij} \left[\log\ \theta_i + g_{ij}\left(\theta|\theta^{(n)}\right)\right],
\end{aligned}
$$

where

$$
g_{ij}\left(\theta|\theta^{(n)}\right) = -\log\left(\theta_i^{(n)} + \theta_j^{(n)}\right) - \frac{1}{\theta_i^{(n)} + \theta_j^{(n)}}\left(\theta_i + \theta_j - \theta_i^{(n)} - \theta_j^{(n)}\right)
$$

# Bradley–Terry Model

- as a result

$$f(\theta) \geq \sum_{i,j} b_{ij} \left[ \log \ \theta_i - \log\left(\theta_i^{(n)} + \theta_j^{(n)}\right) - \frac{\theta_i + \theta_j}{\theta_i^{(n)} + \theta_j^{(n)}} + 1 \right]$$
$$= g\big(\theta | \theta^{(n)}\big)$$

- maximize the minorization function $g\big( \ \cdot \ | \theta^{(n)}\big)$

- thus we have

$$\theta_i^{(n+1)} = \frac{\sum_{j \neq i} b_{ij}}{\sum_{j \neq i} (b_{ij} + b_{ji}) / (\theta_i^{(n)} + \theta_j^{(n)})}$$

# An example Based on Jensen's

- you will be solving a problem in your homework
- based on the inequalities discussed in page 26

# Key Themes

- helpful majorizations and minorizations techniques?
- next 2-3 lectures