MM Optimization Algorithms with Machine Learning Applications

Chathuranga Weeraddana

09 May 2022

Based on a Course at Digital Future

- Course Website:
  - digital future events
  - MM Opt. Alg. with Machine Learning Applications
- Textbooks:
  - Kenneth Lange, *MM Optimization Algorithms*, 2016, `SIAM`

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

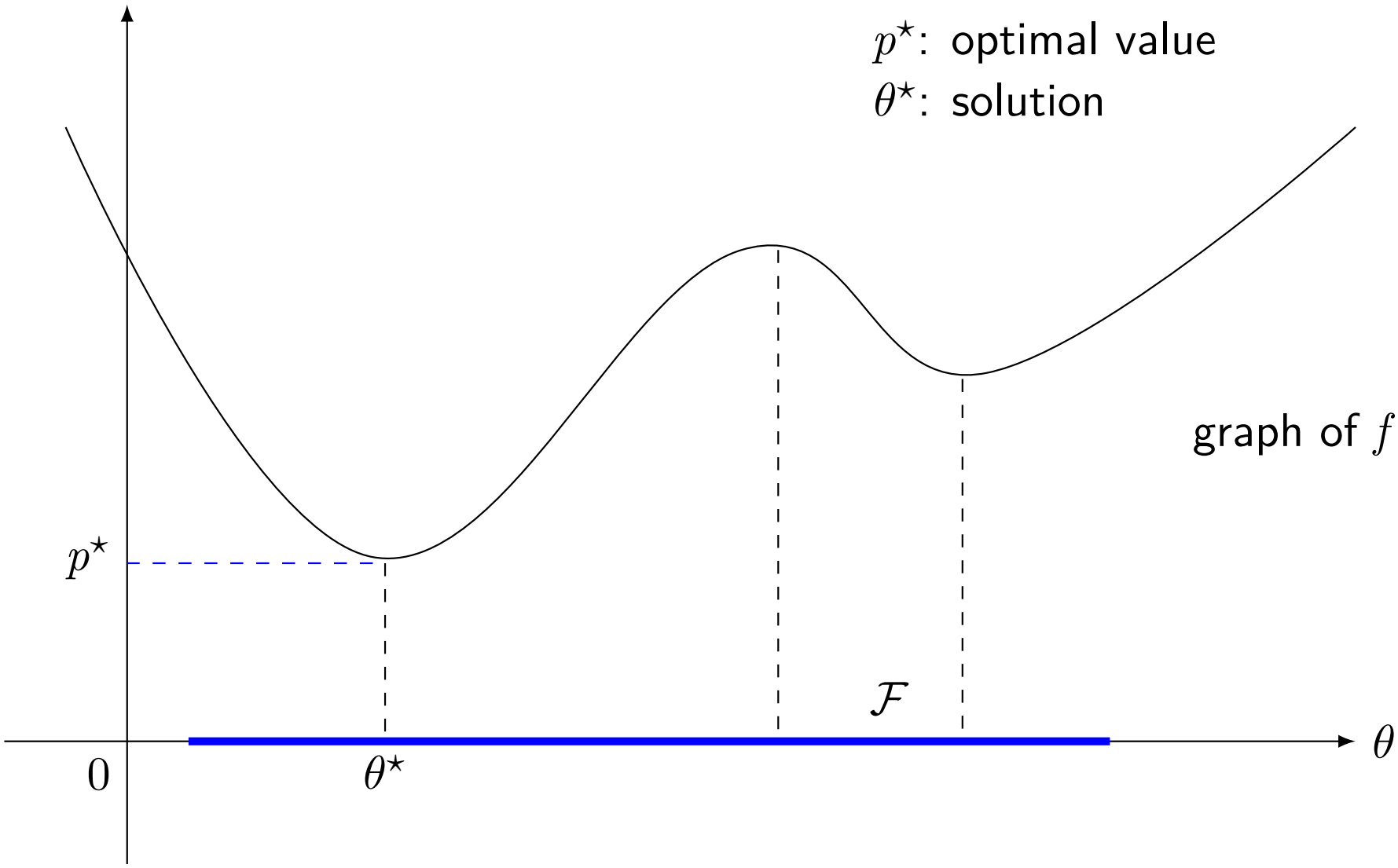
In General an ML Problem?

- a general formulation of an optimization problem

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

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

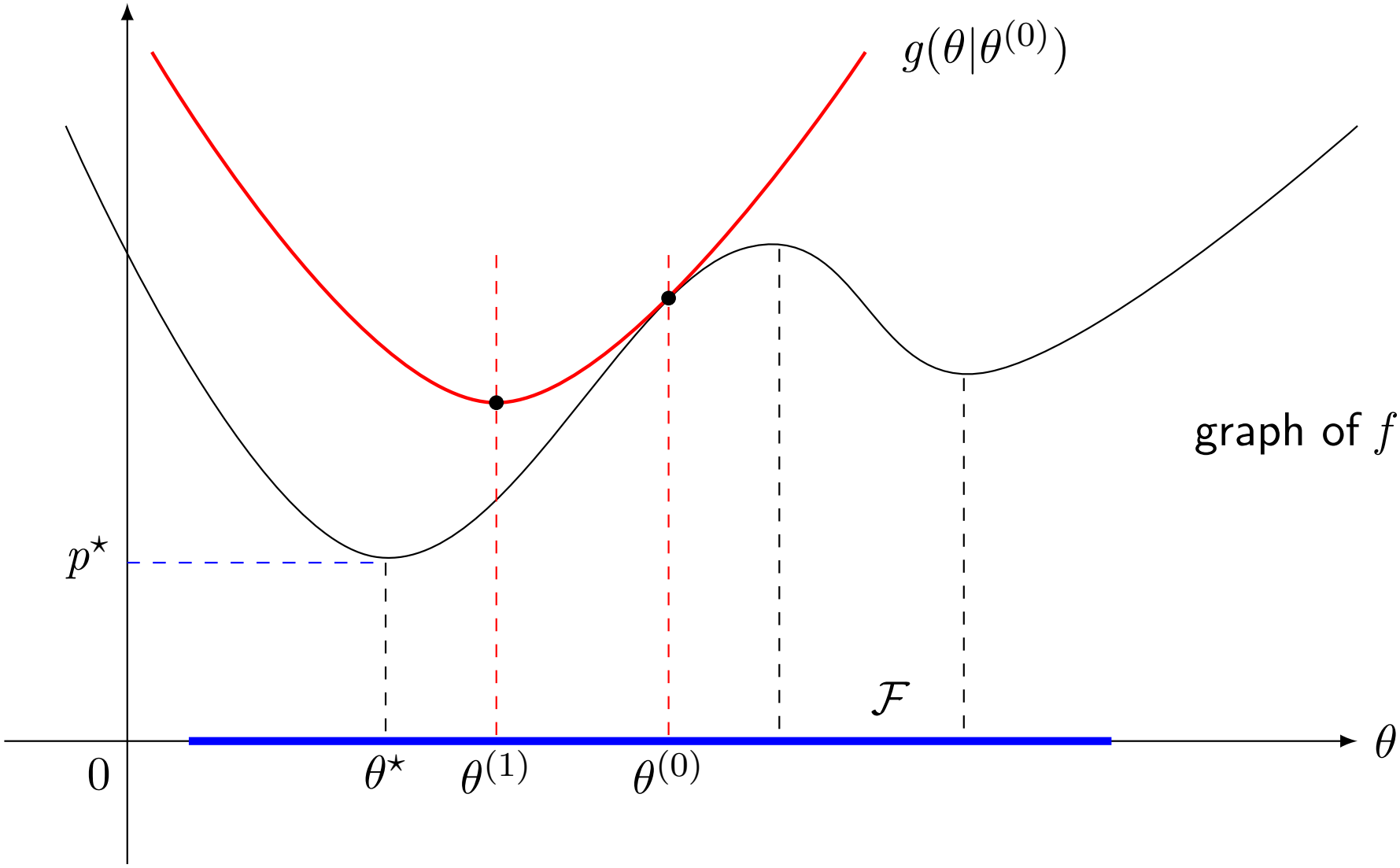
Geometric Interpretation



What is MM?

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

Majorize and Minimize

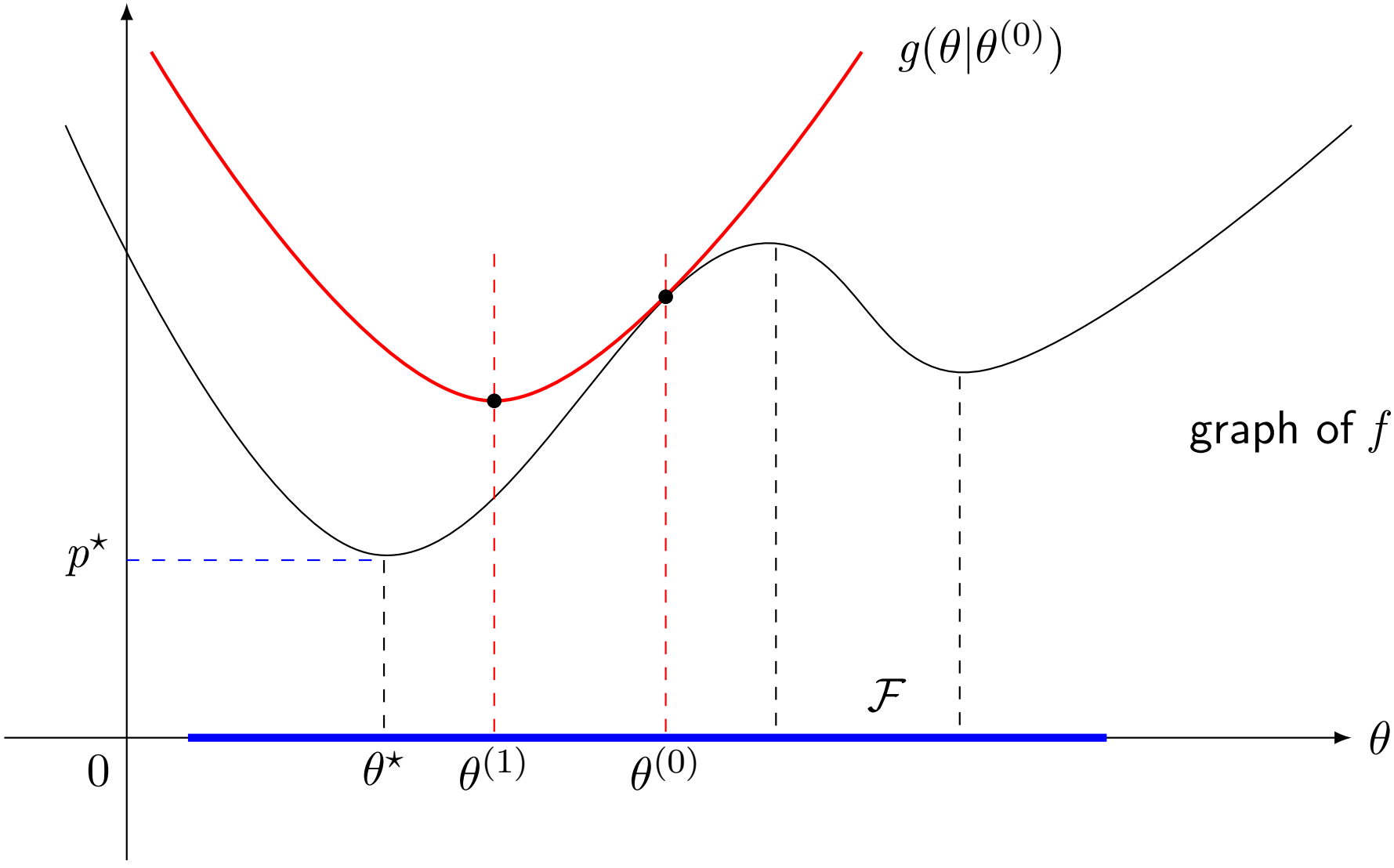


The MM Principle

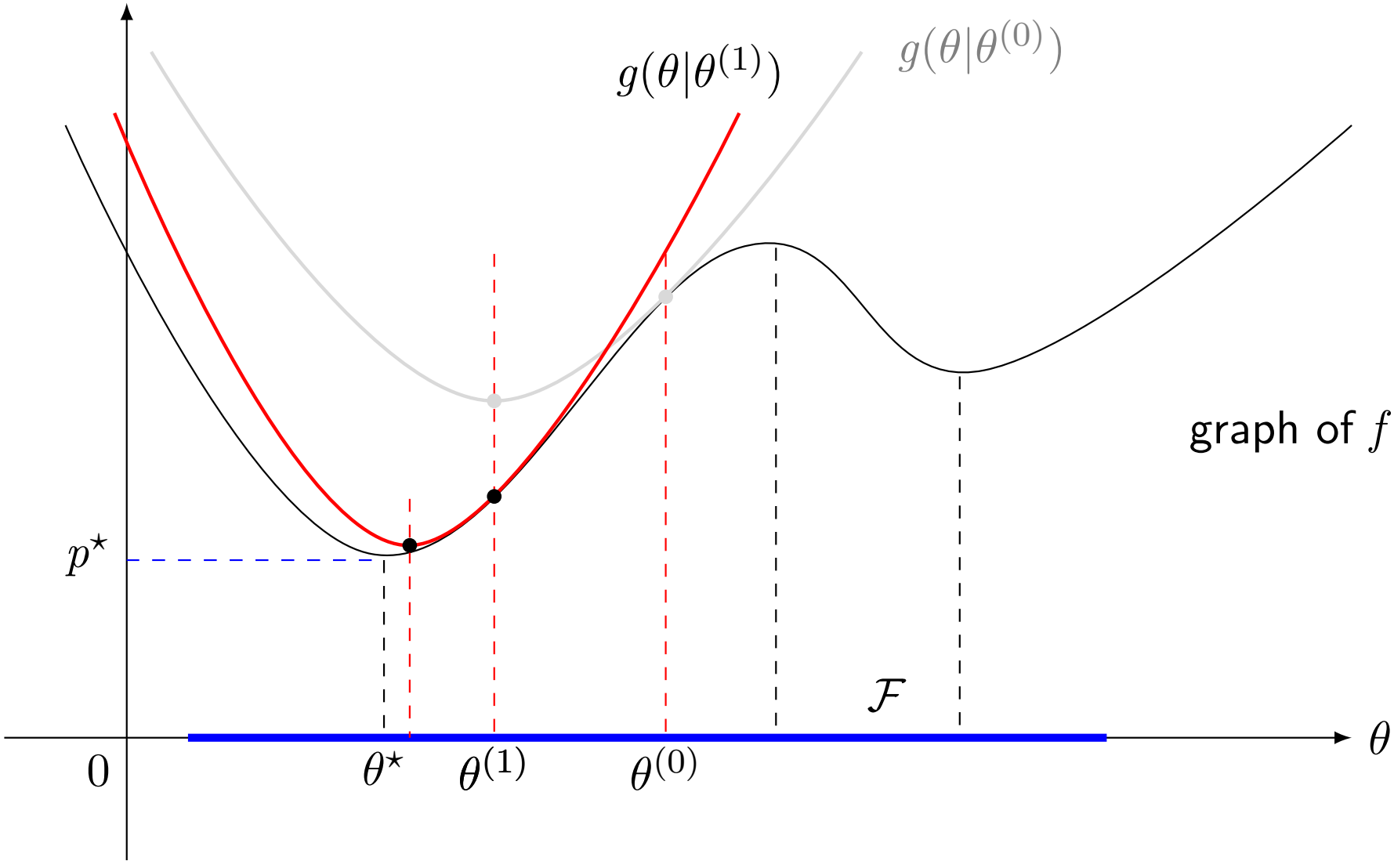
- is not an algorithm
- a useful principle for constructing optimization algorithms
- the resulting algorithms are called MM algorithms
  - majorize and minimize in an iterative mannar

The MM Algorithm
A Geometric Interpretation

Geometric Interpretation



Geometric Interpretation



The MM Algorithm: Key Idea

- majorize and minimize in an iterative mannar

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

Why MM Algorithms?

- MM principle simplifies optimization by
  - separating the variables of a problem
  - avoiding large matrix inversions
  - restoring the symmetry
  - nonconvex settings: designing smart heuristics
  - turning a non-smooth problem into a smooth problem

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

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

Majorizations and Minorizations

- it involves ingenuity and skill
- a few basic themes
  - Jensen's inequality
  - Cauchy-Schwarz inequality
  - supporting Hyperplane inequality
  - quadratic upper bound principle
  - weighted arithmetic-geometric mean and its derivatives
  - inequalities from partial minimization
- list is still growing

A Toy Examples

Minimize $\cos\theta$

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

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

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

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

A Few More Examples

K-Mean Clustering

- m subjects
- each subject i is associated with a vector $y_i \in \mathbb{R}^d$
- subjects must be assigned to one of K clusters
- $\mu_k \in \mathbb{R}^d$, the center of cluster k
- subjects are assigned to clusters based on proximity [1]
- the set of subjects assigned to cluster k is $\mathcal{C}(\mu_k)$

[1] Applications: Topic discovery, Patient clustering, Customer market segmentation, Survey response clustering, and many others.

- Loyd algorithm (1957): key idea
  - choose cluster centers μ_ks arbitrarily
  - centers to clusters: compute $\mathcal{C}(\mu_k)$ [2]
  - clusters to centers: $\mu_k =$ centroid of points in $\mathcal{C}(\mu_k)$
  - iterate above two steps
- we can formulate the problem of K-mean clustering as

$$\text{minimize} \quad f(\mu) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \|y_i - \mu_k\|^2$$

- Loyd alg. $\rightarrow$ not necessarily yield the optimal $\mu_k, \mathcal{C}(\mu_k)$

[2] Ties are broken such that $\cap_k \mathcal{C}(\mu_k) = \emptyset$.

What if y_is are Incomplete?

- e.g., suppose $y_i \in \mathbb{R}^3$
  - $y_1 = (1, 0.5, __)$, $y_2 = (__, __, 3)$, $y_3 = (0.2, 0.4, 2)$, $\ldots$
  - 1st and 2nd indexes of y_1 is observed
  - 3rd index of y_2 is observed
- let $\mathcal{O}_i$ denote the set of indexes observed in subject i
  - $\mathcal{O}_1 = \{1, 2\}$, $\mathcal{O}_2 = \{3\}$, and $\mathcal{O}_3 = \emptyset$ in the example above
- incomplete data destroy the simple two steps of Loyd alg.

Apply MM Principle

- with incomplete y_is, the objective function is given by

$$f(\mu) = \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \left[\sum_{j \in \mathcal{O}_i} (y_{ij} - \mu_{kj})^2 \right]$$

- we can simply majorize f as

$$\begin{aligned} f(\mu) &\leq \sum_{k=1}^{K} \sum_{i \in \mathcal{C}(\mu_k)} \left[\sum_{j \in \mathcal{O}_i} (y_{ij} - \mu_{kj})^2 + \sum_{j \notin \mathcal{O}_i} (\mu_{kj}^{(n)} - \mu_{kj})^2 \right] \\ &= g(\mu | \mu^{(n)}) \end{aligned}$$

since $\sum_{j \notin \mathcal{O}_i} (\mu_{kj}^{(n)} - \mu_{kj})^2 \geq 0$

- symmetry of the data is restored
- thus the Loyd algorithm can be applied as it is

Bradley–Terry Model

- prob. model: predicts the outcome of a paired comparison
- let us consider a sports league with m teams
- ith team's skill level is parameterized by θ_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 [3] of the model parameters $\theta \in \mathbb{R}^m_{++}$?
- the likelihood function of data has the form

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

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

[3] 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:

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

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

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

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

- thus we have

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

Thank You!