MM Optimization Algorithms

Chathuranga Weeraddana

April 2022

Lecture 2: Key Inequalities for MM (Part I)

Majorizations and Minorizations

- it involves ingenuity and skill
- a list helpful majorizations and minorizations
- next 2-3 lectures we review a few basic themes
- list is still growing

Jensen's Inequality

Jensen's Inequality

- recall: when f is convex, then we have

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

- more generally

$$f\big(\textstyle\sum_i \alpha_i t_i\big) \leq \textstyle\sum_i \alpha_i f(t_i), \tag{1}$$

where $\sum_i \alpha_i = 1$ and $\alpha_i \geq 0$ for all i

A Different Useful Form

- suppose $a \in \mathbb{R}^N$ and $\theta \in \mathbb{R}^N$ and all are possitive
- in (1), let

$$\alpha_i = \frac{a_i \theta_i^{(n)}}{a^\mathsf{T} \theta^{(n)}} \quad \text{and} \quad t_i = \frac{a^\mathsf{T} \theta^{(n)}}{\theta_i^{(n)}} \; \theta_i$$

- then from (1), we get

$$\begin{aligned} f\left(a^\mathsf{T} \theta\right) &\leq \sum_{i=1}^{N} \frac{a_i \theta_i^{(n)}}{a^\mathsf{T} \theta^{(n)}} \; f\left(\frac{a^\mathsf{T} \theta^{(n)}}{\theta_i^{(n)}} \; \theta_i\right) \\ &= g\left(\theta | \theta^{(n)}\right) \end{aligned} \tag{2}$$

Counting with Poisson

- probability model: Poisson
- it predicts number of events over some period of time
- probability that there is y events is given by

$$p_\mu(Y = y) = \frac{\mu^y e^{-\mu}}{y!}$$

- let μ modeled as an affine function of $u \in \mathbb{R}^N$, i.e., $\mu = \theta^\mathsf{T} u$
- u : the explanatory variable, θ : the model parameter

Counting with Poisson

- $\big(u(j), y(j)\big)$, $j = 1, \dots, m$: a number of observations (data)
- ML estimate of the model parameters $\theta \in \mathbb{R}^N_{++}$?
- the likelihood function of data has the form

$$p_\theta\big(\,(u(j), y(j))_j\,\big) = \prod_{j=1}^{m} \frac{\big(\theta^\mathsf{T} u(j)\big)^{y(j)} e^{-\theta^\mathsf{T} u(j)}}{y(j)!}$$

- the log-likelihood function $f(\theta) = \log\ p_\theta\big(\,(u(j), y(j))_j\,\big)$
- the log-likelihood function f should be maximized over θ

Counting with Poisson

- let us compute a minorization function:

$$
\begin{aligned}
f(\theta) &= \log \; p_\theta\big(\, (u(j), y(j))_j \, \big) \\
&= \sum_j \; y(j) \log\left(u(j)^\mathsf{T}\theta\right) - u(j)^\mathsf{T}\theta - \log(y(j)!) \\
&\overset{(2)}{\geq} \sum_{j=1}^{m} \left[y(j) \sum_{i=1}^{N} w_{jin} \log\left(s_{jin}\theta_i\right) - u(j)^\mathsf{T}\theta \right] + s \\
&= g(\theta|\theta^{(n)}),
\end{aligned}
$$

where

$$
w_{jin} = \frac{u_i(j)\theta_i^{(n)}}{u(j)^\mathsf{T}\theta^{(n)}} \text{ and } s_{jin} = \frac{u(j)^\mathsf{T}\theta^{(n)}}{\theta_i^{(n)}}
$$

Counting with Poisson

- as a result of maximizing $g(\theta|\theta^{(n)})$, we have

$$\theta_i^{(n+1)} = \left(\sum_{j=1}^m y(j) w_{ijn} \right) / 1^\mathsf{T} u(j)$$

- for an arbitrary explanatory $u \in \mathbb{R}^N$, the Poisson model is

$$p_{\theta^\star}(Y = y) = \frac{\left(\theta^{\star\mathsf{T}} u\right)^y \exp\left(-\theta^{\star\mathsf{T}} u\right)}{y!},$$

where $\theta^\star$ is given by the MM algorithm after the convergence

Finite Mixture Model

- used for [1]
  - categorizing age groups of animals
  - medical diagnosis and prognosis
  - latent structure analysis
- probability distribution is modeled as

$$p_{\phi,\pi}(y) = \sum_{k=1}^{c} \pi_k \; p_{k\phi}(y) \tag{3}$$

- $\theta = (\phi, \pi) = (\phi, \pi_1, \ldots, \pi_c)$: the model parameter

[1] For more examples, see § 2 of *Statistical Analysis of Finite Mixture Distributions* by D. M. Titterington, A.F.M. Smith and U.E. Makov, 1985.

Finite Mixture Model

- e.g., Gaussian mixture model
  - $\phi = (\mu_1, \ldots, \mu_c, \Sigma_1, \ldots, \Sigma_c)$
  - $p_{k\phi}(\cdot)$ is a Gaussian density, more specifically

$$p_{k\phi}(y) = \frac{1}{\sqrt{(2\pi)^l |\Sigma_k|}} \exp\left(-\frac{(y-\mu_k)^\mathsf{T} \Sigma_k^{-1} (y-\mu_k)}{2}\right) \quad (4)$$

  - $\theta = (\mu_1, \ldots, \mu_c, \Sigma_1, \ldots, \Sigma_c, \pi_1, \ldots, \pi_c)$

Finite Mixture Model

- $\big(y(j)\big)$, $j = 1, \ldots, m$: a number of observations (data)
- ML estimate of the model parameters θ?
- the likelihood function of data has the form

$$
\begin{aligned}
p_\theta\big(\,(y(j))_j\,\big) &= \prod_{j=1}^{m} p_{\phi,\pi}(y(j)) \\
&= \prod_{j=1}^{m} \sum_{k=1}^{c} \pi_k \; p_{k\phi}(y(j))
\end{aligned}
$$

- the log-likelihood function $f(\theta) = \log \; p_\theta\big(\,(y(j))_j\,\big)$
- the log-likelihood function f should be maximized over θ

Finite Mixture Model

- let us compute a minorization function:

$$
\begin{aligned}
f(\theta) &= \log\ p_\theta\big(\,(y(j))_j\,\big) \\
&= \sum_j \ \log\left(\sum_{k=1}^{c} \pi_k\ p_{k\phi}(y(j))\right) \\
&\overset{(2)}{\geq} \sum_{j=1}^{m}\left[\sum_{k=1}^{c} w_{jkn} \log\left(s_{jkn}\pi_k\ p_{k\phi}(y(j))\right)\right] \\
&= g(\theta|\theta^{(n)}),
\end{aligned}
$$

where

$$
w_{jkn} = \frac{\pi_k^{(n)}\ p_{k,\phi^{(n)}}(y(j))}{\sum_{i=1}^{c} \pi_i^{(n)}\ p_{i,\phi^{(n)}}(y(j))} \text{ and } s_{jkn} = w_{jkn}^{-1}
$$

Finite Mixture Model

- let us minimize $g(\theta|\theta^{(n)})$ which is given by [2]

$$
\begin{aligned}
g(\theta|\theta^{(n)}) &= \sum_{k=1}^{c}\sum_{j=1}^{m} w_{jkn}\log\pi_k + \sum_{k=1}^{c}\sum_{j=1}^{m} w_{jkn}\log p_{k\phi}(y(j)) \\
&= \sum_{k=1}^{c} \alpha_{kn}\log\pi_k + \sum_{k=1}^{c}\sum_{j=1}^{m} w_{jkn}\log p_{k\phi}(y(j))
\end{aligned}
$$

  where $\alpha_{kn} = \sum_{j=1}^{m} w_{jkn}$

- ϕ and $\pi = (\pi_1, \ldots, \pi_c)$ are separate $\rightarrow$ minimize separately

[2] Irrelevant constants are dropped.

Finite Mixture Model

- maximization with respect to π

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{c} \alpha_{kn} \log \pi_k \\ \text{subject to} & \sum_{k=1}^{c} \pi_k = 1 \\ & \pi_k \geq 0,\ k = 1, \dots, c \end{array} \tag{5}$$

- closed form solution of the problem above is

$$\begin{aligned} \pi_k^{(n+1)} &= \alpha_{kn} / (\textstyle\sum_{\bar{k}=1}^{c} \alpha_{\bar{k}n}) \\ &= \big(\textstyle\sum_{j=1}^{m} w_{jkn} \big) / m \end{aligned}$$

Finite Mixture Model

- suppose $p_{k\phi}$ is given by (4)
- maximization with respect to $\phi = (\mu_1, \ldots, \mu_c, \Sigma_1, \ldots, \Sigma_c)$

$$\begin{array}{ll} \text{minimize} & \sum_{k=1}^{c} \sum_{j=1}^{m} w_{jkn} \log p_{k\phi}(y(j)) \\ \text{subject to} & \Sigma_k \succeq 0,\ k = 1, \ldots, c \end{array} \qquad (6)$$

- alternating optimization to solve (6) in closed form

$$\mu_k^{(n+1)} = (1/m) \sum_{j=1}^{m} y(j)$$

$$\Sigma_k^{(n+1)} = \frac{1}{\sum_{j=1}^{m} w_{jkn}} \sum_{\bar{j}=1}^{m} w_{\bar{j}kn} \left(y(\bar{j}) - \mu_k^{(n+1)}\right) \left(y(\bar{j}) - \mu_k^{(n+1)}\right)^{\mathsf{T}}$$

Finite Mixture Model

- as a result of maximizing $g(\theta|\theta^{(n)})$, we have

$$\theta_i^{(n+1)} = \left(\underbrace{\mu_1^{(n+1)}, \ldots, \Sigma_1^{(n+1)}, \ldots}_{\phi^{(n+1)}}, \underbrace{\pi_1^{(n+1)}, \ldots, \pi_c^{(n+1)}}_{\pi^{(n+1)}} \right)$$

- thus, the pdf model $p_{\phi^\star,\pi^\star} : \mathbb{R}^l \to \mathbb{R}$ is [compare with (3)]

$$p_{\phi^\star,\pi^\star}(y) = \sum_{k=1}^{c} \pi_k^\star \, p_{k\phi^\star}(y)$$

where $\theta^\star = (\phi^\star, \pi^\star)$ is given by the MM algorithm

Cauchy-Schwarz Inequality

Cauchy-Schwarz Inequality

- suppose $x, y \in \mathbb{R}^N$
- Cauchy-Schwarz inequality is given by

$$|y^\mathsf{T} x| \leq ||y||\ ||x||$$

- i.e., $-||y||\ ||x|| \leq y^\mathsf{T} x \leq ||y||\ ||x||$

MDS

- MDS stands for multi dimensional scaling
- there are n objects
- we are also given their pairwise dissimilarity $d_{ij} \geq 0$
- need to represent n objects by using points in $\mathbb{R}^p$
- those points are given by $x_k \in \mathbb{R}^p, \ k = 1, \ldots, n$

MDS

- we want to compute $X \in \mathbb{R}^{p \times n}$, where

$$X = [x_1 \cdots x_n]$$

- the variable X is computed by minimizing f where

$$\begin{aligned} f(X) &= \sum_i \sum_{j \neq i} \; (d_{ij} - ||x_i - x_j||)^2 \\ &= \sum_i \sum_{j \neq i} d_{ij}^2 + \sum_i \sum_{j \neq i} ||x_i - x_j||_{ij}^2 \\ &\qquad - 2 \sum_i \sum_{j \neq i} d_{ij} ||x_i - x_j|| \end{aligned}$$

- function f should be minimized over X

MDS

- let us compute a majorization function to the last term
- we have from the Cauchy-Schwarz inequality

$$
\begin{aligned}
-d_{ij}||x_i - x_j|| &\leq d_{ij} \frac{\left(x_i^{(n)} - x_j^{(n)}\right)^{\mathsf{T}} (x_i - x_j)}{\left\| x_i^{(n)} - x_j^{(n)} \right\|} \\
&= g_{ij}(X|X^{(n)})
\end{aligned}
$$

- thus a majorization function for f is given by

$$
\begin{aligned}
f(X) &\leq \sum_i \sum_{j \neq i} ||x_i - x_j||_{ij}^2 + 2 \sum_i \sum_{j \neq i} g_{ij}(X|X^{(n)}) + d \\
&= g(X|X^{(n)})
\end{aligned}
$$

MDS

- f is not differentiable
- $g(\cdot | X^{(n)})$ is not only differentiable, but also quadratic
- further processing: $||x_i - x_j||^2$ can also be majorized
  - why?

MDS

- f is not differentiable
- $g(\,\cdot\,|X^{(n)})$ is not only differentiable, but also quadratic
- further processing: $||x_i - x_j||^2$ can also be majorized
  - why? to enable separability
- a small trick based on the convexity of $||\cdot||^2$, i.e.,

MDS

- how?

$$
\begin{aligned}
||x_i - x_j||^2 &= \left\| x_i - x_j + (1/2)\big(x_i^{(n)} - x_i^{(n)} + x_j^{(n)} - x_j^{(n)}\big) \right\|^2 \\
&= \left\| \left(x_i - (1/2)\big(x_i^{(n)} + x_j^{(n)}\big) \right) - \left(x_j - (1/2)\big(x_i^{(n)} + x_j^{(n)}\big) \right) \right\|^2 \\
&= \left\| \frac{1}{2}\left(2x_i - \big(x_i^{(n)} + x_j^{(n)}\big) \right) - \frac{1}{2}\left(2x_j - \big(x_i^{(n)} + x_j^{(n)}\big) \right) \right\|^2 \\
&\leq 2 \left\| x_i - \frac{1}{2}\big(x_i^{(n)} + x_j^{(n)}\big) \right\|^2 + 2 \left\| x_j - \frac{1}{2}\big(x_i^{(n)} + x_j^{(n)}\big) \right\|^2 \\
&= \tilde{g}_{ij}(X|X^{(n)})
\end{aligned}
$$

MDS

- thus the new majorization function for f is given by

$$f(X) \leq \sum_i \sum_{j \neq i} \tilde{g}_{ij}(X|X^{(n)}) + 2 \sum_i \sum_{j \neq i} g_{ij}(X|X^{(n)}) + d$$
$$= h(X|X^{(n)})$$

- $h(\,\cdot\,|X^{(n)})$ is quadratic and separable

- minimize $h(\,\cdot\,|X^{(n)})$
  - closed form: up to each element x_{im} of x_i, i.e.,

$$x_{im}^{(n+1)} = r_i(x_{im}^{(n)})$$

  - you may compute r_i

Supporting Hyperplane Inequality

Supporting Hyperplane Inequality

- for a convex function it produces an affine minorization
- for a concave function it produces an affine majorization
- suppose f is convex, then

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

where $v^{(n)} \in \partial f\big(x^{(n)}\big)$

Maximizing a Convex over Compact Set

- maximizing a convex f over compact $\mathcal{C} \subset \mathbb{R}^n$
- not a convex problem
- however, the maximizing $g(\ \cdot \ | x^{(n)})$ turns out to be promising
- related to the well-known support function $\sigma_{\mathcal{C}}$ of $\mathcal{C}$ given by

$$\sigma_{\mathcal{C}}(y) = \sup_{x \in \mathcal{C}} y^{\mathsf{T}} x$$

Maximizing a Convex over Compact Set

- e.g.,

$$\begin{array}{ll} \text{maximize} & (1/2)(x-a)^{\mathsf{T}} P (x-a) \\ \text{subject to} & ||x|| = 1 \end{array}$$

- P is positive semidefinite and $a \in \mathbb{R}^{(n)}$
- the solution of the problem above is

$$x_k^{(n+1)} = \frac{1}{\left\| P\left(x^{(n)} - a\right) \right\|} \; P\left(x^{(n)} - a\right)$$

Concave-Convex Principle

- minimizing a difference of convex functions f and h
- i.e., $f - h$ is to be minimized
- not a convex problem
- consider the following majorization for $-h$

$$-h(x) \leq -h\big(x^{(n)}\big) - v^{(n)\mathsf{T}}\big(x - x^{(n)}\big)$$

where $v^{(n)} \in \partial h\big(x^{(n)}\big)$

Concave-Convex Principle

- thus a majorization function for $f - h$ is given by

$$\begin{aligned} f(x) - h(x) &\leq f(x) - h\big(x^{(n)}\big) - v^{(n)\mathsf{T}}\big(x - x^{(n)}\big) \\ &= g(x|x^{(n)}) \end{aligned}$$

- note that $g(\,\cdot\,|x^{(n)})$ is convex and we have

$$x^{(n+1)} = \arg\min_{x}\; g\big(x|x^{(n)}\big)$$

Concave-Convex Principle

- e.g., minimizing a quadratic over a compact and convex set
- let P be symmetric and indefinite, $\mathcal{C}$ compact and convex
- consider the problem

$$\begin{array}{ll} \text{minimize} & x^{\mathsf{T}} P x \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

- not a convex problem
- we can express $x^{\mathsf{T}} P x$ in the form $f(x) - h(x)$, f, h convex

Concave-Convex Principle

- the spectral decomposition of P

$$P = V\Lambda V^{\mathsf{T}} = \underbrace{\sum_{\{i|\lambda_i>0\}} \lambda_i v_i v_i^{\mathsf{T}}}_{Q} - \underbrace{\sum_{\{j|\lambda_j<0\}} |\lambda_j| v_j v_j^{\mathsf{T}}}_{R}$$
$$= Q - R$$

where $Q, R \succeq 0$

- as a result, we have

$$\begin{aligned} x^{\mathsf{T}} P x &= x^{\mathsf{T}} Q x - x^{\mathsf{T}} R x \\ &\leq x^{\mathsf{T}} Q x - 2x^{(n)\mathsf{T}} R x + c \\ &= g\big(x|x^{(n)}\big) \end{aligned}$$

Concave-Convex Principle

- thus the following problem is to be solved

$$\begin{array}{ll} \text{maximize} & g\big(x|x^{(n)}\big) = x^\mathsf{T} Q x - 2x^{(n)\mathsf{T}} R x + c \\ \text{subject to} & x \in \mathcal{C} \end{array}$$

- this is a constrained (convex) quadratic problem where

$$x^{(n+1)} = \underset{x \in \mathcal{C}}{\arg\min}\ g\big(x|x^{(n)}\big)$$

Concave-Convex Principle

- another example: weighted sum-rate maximization

$$\begin{array}{ll} \text{maximize} & \sum_{i=1}^{N} \log\left[1 + \mathtt{SINR}_i(p)\right] \\ \text{subject to} & 1^{\mathsf{T}} p \leq p_{\mathtt{tot}} \\ & p \succeq 0 \end{array}$$

  where $p = [p_1 \ldots p_N]^{\mathsf{T}}$ and

$$\mathtt{SINR}_i(p) = \frac{\alpha_i p_i}{\sigma^2 + \sum_{j \neq i} \alpha_i p_j}$$

- you may try this