MM Optimization Algorithms

Chathuranga Weeraddana

March 2022

Lecture 4: Key Inequalities for MM (Part III)

Majorization and Partial Optimization

Partial Minimization

- variety of functions can be represented as partial minima [1]

$$f(x) = \min_{y \in \mathcal{Y}} g(x, y) \tag{1}$$

- such a function f can readily be majorized at $x^{(n)} \in \mathbb{R}^p$, i.e.,

$$\begin{aligned} f(x) &= \min_{y \in \mathcal{Y}} g(x, y) \\ &\leq g\big(x, y^{(n)}\big) = h(x \mid x^{(n)}) \end{aligned} \tag{2}$$

  where $y^{(n)} = \arg\min_{y \in \mathcal{Y}} g\big(x^{(n)}, y\big)$

- $h(x \mid x^{(n)})$ is the restriction of g to set $\{(x, y^{(n)}) \mid x \in \mathbb{R}^p\}$

[1] It is assumed that the minimum over $y \in \mathcal{Y}$ is attained for each x.

Examples

Block Descent

- suspect you are given the problem

$$\begin{array}{ll} \text{minimize} & g(x, y) \\ \text{subject to} & x \in \mathcal{X} \\ & y \in \mathcal{Y} \end{array}$$

- the problem is equivalent to

$$\begin{array}{ll} \text{minimize} & f(x) = \min_{y \in \mathcal{Y}} g(x, y) \\ \text{subject to} & x \in \mathcal{X} \end{array}$$

- MM principle: $f(x) \leq h\big(x|x^{(n)}\big) = g\big(x, y^{(n)}\big)$

- can the constraint be of the form $(x, y) \in \mathcal{Z}$?

A Simple Problem

- consider the following problem

$$\begin{array}{ll} \text{minimize} & \|Ax - y\| \\ \text{subject to} & y \in \mathcal{Y} \end{array}$$

  where the decision variables are x, y

- the iterative algorithm reduces to:

$$x^{(n+1)} = (A^\mathsf{T} A)^{-1} A^\mathsf{T} P_\mathcal{Y}(Ax^{(n)})$$

  where $P_\mathcal{Y}(\,\cdot\,)$ is the projection onto $\mathcal{Y}$

Distance Between Two Sets

- $\mathcal{X}$, $\mathcal{Y}$ $\rightarrow$ two disjoint closed sets
- compute $\texttt{dist}(\mathcal{X}, \mathcal{Y})$ the optimal value of

$$\begin{array}{ll} \text{minimize} & \|x - y\| \\ \text{subject to} & x, \in \mathcal{X}, y \in \mathcal{Y} \end{array}$$

  where the decision variables are x, y

- the iterative algorithm reduces to:

$$x^{(n+1)} = P_{\mathcal{X}}\big(P_{\mathcal{Y}}(x^{(n)})\big)$$

- optimality if $\mathcal{X}$ (or both sets) is nonconvex?

Proximal Minimization Algorithm

- suppose you are given the problem

$$\underset{x}{\text{minimize}} \quad f(x)$$

- trivial to see: $f(x) = \min_{y} \left[f(x) + (1/2\mu)\|x - y\|^2 \right]$, $\mu > 0$

- thus a majorization function of f is given by $h\big(x|x^{(n)}\big)$ where

$$h\big(x|x^{(n)}\big) = f(x) + (1/2\mu)\|x - x^{(n)}\|^2$$

- output of $h\big(\ \cdot \ |x^{(n)}\big)$ minimization compromises between
  - minimizing f and being near to $x^{(n)}$ (controlled by μ)

- the algorithm if MM principle is applied

$$x^{(n+1)} = \texttt{prox}_{\mu f}\big(x^{(n)}\big)$$

where $\texttt{prox}_{\mu f}$ is called the proximal operator of μf,

$$\texttt{prox}_{\mu f}(v) = \arg\min_x \; f(x) + (1/2\mu)\|x - v\|^2$$

- the resulting algorithm is a proximal minimization algorithm
    - also called: proximal iteration or the proximal point algorithm [2]

[2] See § 4.1 of *Proximal Algorithms* by N. Parikh and S. Boyd, now Foundations and Trends in Optimization 2013.

- why compute a sequence of proximal operators?
    - subproblems usually admits easy closed-form solutions
    - can be solved sufficiently quickly
    - minimizing of $(f+$ quadratic$)$ is *easier* than minimizing f
        - handle ill-conditioned situations $\rightarrow$ higher reliability
        - fewer iterations or faster convergence
    - amenable to distributed optimization
- an application: iterative refinement $\rightarrow$ a homework exercise

Schur Compliment Majorization

- more specifically
  - we are given a vector x
  - we have a parameter matrix D given by

$$D = \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}$$

- now it is useful to bound

$$1) \quad x^{\mathsf{T}} A^{-1} x \qquad\qquad 2) \quad \ln |A|$$

Bounding $x^\intercal A^{-1} x$

- we have [3]

$$
\begin{aligned}
f(x) &= \inf_z \; g(x,z) \\
&= \inf_x \; \begin{bmatrix} x \\ z \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}^{-1} \begin{bmatrix} x \\ z \end{bmatrix} \\
&= x^\mathsf{T} A^{-1} x \\
&= \begin{bmatrix} x \\ B^\mathsf{T} A^{-1} x \end{bmatrix}^\mathsf{T} \begin{bmatrix} A & B \\ B^\mathsf{T} & C \end{bmatrix}^{-1} \begin{bmatrix} x \\ B^\mathsf{T} A^{-1} x \end{bmatrix}
\end{aligned}
$$

[3] See § A.5.5, *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004.

- from the last two equations we get for all A, B and C

$$x^{\mathsf{T}} A^{-1} x = \begin{bmatrix} x \\ B^{\mathsf{T}} A^{-1} x \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}^{-1} \begin{bmatrix} x \\ B^{\mathsf{T}} A^{-1} x \end{bmatrix}$$

$$\leq \begin{bmatrix} x \\ z^{(n)} \end{bmatrix}^{\mathsf{T}} \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix}^{-1} \begin{bmatrix} x \\ z^{(n)} \end{bmatrix} \tag{3}$$

with $z^{(n)} = B^{(n)\mathsf{T}} \left(A^{(n)}\right)^{-1} x$

- note that the inequality holds with equality when

$$A = A^{(n)},\ B = B^{(n)},\ C = C^{(n)}$$

Bounding $\ln |A|$

- the Schur complement of A in the matrix D is given by

$$C - B^{\mathsf{T}} A^{-1} B$$

- for clarity let us define G as

$$G = (C - B^{\mathsf{T}} A^{-1} B)^{-1}$$

- we have the following determinant identity

$$|D| = |A| \times |C - B^{\mathsf{T}} A^{-1} B| = |A|/|G|$$

- now we can bound $\ln|A|$ as follows:

$$\begin{aligned}\ln|A| &= \ln|D| + \ln|G| \\ &\le \ln|D| + \ln|G^{(n)}| + \mathtt{Tr}\big[(G^{(n)})^{-1}(G - G^{(n)})\big] \\ &= \ln|D| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[(G^{(n)})^{-1}G\big] \\ &= \ln|D| + \ln|G^{(n)}| - \mathtt{Tr}(I) + \mathtt{Tr}\big[D^{-1}F^{(n)}\big] \qquad (4)\end{aligned}$$

- the last equality follows from that $F^{(n)} \triangleq \begin{bmatrix} 0 & 0 \\ 0 & (G^{(n)})^{-1} \end{bmatrix}$ and

$$D^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BGB^{\mathsf{T}}A^{-1} & -A^{-1}BG \\ -GB^{\mathsf{T}}A^{-1} & G \end{bmatrix}$$

- note that the inequality holds with equality when

$$A = A^{(n)},\ B = B^{(n)},\ C = C^{(n)}$$

Fenchel Conjugate

- Fenchel conjugate [4] of a function f

$$f^*(x) = \sup_y \{x^\mathsf{T} y - f(y)\} \tag{5}$$

- in general we have for

$$\begin{aligned} f^*(x) &= \sup_y \{x^\mathsf{T} y - f(y)\} && (6) \\ &\geq \bar{y}^\mathsf{T} x - f(\bar{y}) && (7) \\ &= g\left(x | x^{(n)}\right) && (8) \end{aligned}$$

where $\bar{y} \in \partial f^*\left(x^{(n)}\right) = \arg\max_y \{x^{(n)\mathsf{T}} y - f(y)\}$

[4] For more details see pages 15-17.

Appendices

Legendre-Fenchel Transform

- for any function $f : \mathbb{R}^N \to \bar{\mathbb{R}}$ define [5]

$$f^*(x) = \sup_y \, \{x^\mathsf{T} y - f(y)\} \tag{9}$$

- f^* is called the conjugate to f
- biconjugate to f is given by $f^{**} = \left(f^*\right)^*$, where

$$f^{**}(y) = \sup_x \, \{y^\mathsf{T} x - f^*(x)\} \tag{10}$$

---

[5] $\bar{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$.

- the mapping $f \to f^*$ from $\texttt{fcns}(R^N)$ [6] into $\texttt{fcns}(\mathbb{R}^N)$
  - is called the Legendre-Fenchel Transform [7]
- if f is proper, lsc, and convex, so is f^* and $f^{**} = f$

[6] $\texttt{fcns}(\mathbb{R}^N)$: the collection of all extended-real-valued functions on $\mathbb{R}^N$

[7] See pp. 473-476 *Variational Analysis* by R. T. Rockafellar and R. J-B Wets, 3rd printing 2009.

- for any proper, lsc, convex function f

$$\bar{x} \in \partial f(\bar{y}) \iff \bar{y} \in \partial f^*\bar{(}x) \iff f(\bar{y}) + f^*(\bar{x}) = \bar{x}^\mathsf{T}\bar{y}$$

  where

$$\partial f(\bar{y}) = \arg\max_x \{\bar{y}^\mathsf{T}x - f^*(x)\} \quad \partial f^*(\bar{x}) = \arg\max_y \{\bar{x}^\mathsf{T}y - f(y)\}$$

- in general,

$$f(y) + f^*(x) \geq x^\mathsf{T}y \qquad \text{for all} \quad x, y$$

- see Proposition 11.3, R. T. Rockafellar