MM Optimization Algorithms

Chathuranga Weeraddana

April 2022

Lecture 3: Key Inequalities for MM (Part II)

Quadratic Upper Bound Principle

Quadratic Upper Bound Principle

- key mechanisms
  - majorization via gradient Lipschitz continuity
  - majorization via bounded Hessian
  - minorization via strong convexity

Gradient Lipschitz Continuity

- suppose
  - f is differentiable
  - gradient Lipschitz continuous with constant L, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L||x - y|| \text{ for all } x, y \tag{1}$$

- from (44) [1]

$$
\begin{aligned}
f(x) &= f(y) + \int_0^1 \nabla f\big(y + t(x-y)\big)^{\mathsf{T}} (x-y) dt && (2)\\
&= f(y) + \nabla f(y)^{\mathsf{T}} (x-y) + \\
&\qquad \int_0^1 \left[\nabla f\big(y + t(x-y)\big) - \nabla f(y)\right]^{\mathsf{T}} (x-y) dt && (3)\\
&\leq f(y) + \nabla f(y)^{\mathsf{T}} (x-y) + \\
&\qquad \int_0^1 \|\nabla f\big(y + t(x-y)\big) - \nabla f(y)\| \|(x-y)\| dt && (4)\\
&\leq f(y) + \nabla f(y)^{\mathsf{T}} (x-y) + L\|x-y\|^2 \int_0^1 t \ dt && (5)\\
&= f(y) + \nabla f(y)^{\mathsf{T}} (x-y) + (L/2)\|x-y\|^2 = g(x|y) && (6)
\end{aligned}
$$

[1] See page 34 and substitute $a = y$ and $b = x - y$.

Bounded Hessian

- suppose
  - f is twice differentiable
  - f has bounded Hessians, i.e.,

$$\exists\ B \succ 0 \text{ s.t } B - \nabla^2 f(x) \succeq 0 \text{ for all } x \tag{7}$$

- from (46) [2]

$$
\begin{aligned}
f(x) = f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + & \qquad (8) \\
\int_0^1 \int_0^t (x-y)^{\mathsf{T}} \nabla^2 f\big(y + \tau(x-y)\big)(x-y)\ d\tau dt & \qquad (9) \\
\leq f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + & \\
\int_0^1 \int_0^t (x-y)^{\mathsf{T}} B(x-y)\ d\tau dt & \qquad (10) \\
= f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + (1/2)(x-y)^{\mathsf{T}} B(x-y) & \qquad (11) \\
= g(x|y) & \qquad (12)
\end{aligned}
$$

[2] See page 35 and substitute $a = y$ and $b = x - y$.

Strong Convexity

- suppose
  - f is twice differentiable and strongly convex
  - thus,
$$\exists\ m > 0 \text{ s.t } \nabla^2 f(x) - mI \succeq 0 \text{ for all } x \tag{13}$$

- from (46)

$$
\begin{aligned}
f(x) &= f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + && (14)\\
&\qquad \int_0^1 \int_0^t (x-y)^{\mathsf{T}} \nabla^2 f\big(y + \tau(x-y)\big)(x-y)\; d\tau dt && (15)\\
&\geq f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + \\
&\qquad m \int_0^1 \int_0^t (x-y)^{\mathsf{T}}(x-y)\; d\tau dt && (16)\\
&= f(y) + \nabla f(y)^{\mathsf{T}}(x-y) + (m/2)\|x-y\|^2 && (17)\\
&= g(x|y) && (18)
\end{aligned}
$$

Examples

Landweber's Method

- consider a positive definite matrix $A \in \mathbb{S}^n$
- we have to find the solution $x^\star$ for $Ax = -b$
- the task requires order of n^3 flops
  - A^{-1} is to be computed
  - if n is very large the task may be computationally challending
- but we may rely on MM principle $\rightarrow$ avoid matrix inversion

- note that

$$x^\star = \arg\min_x \; f(x) = (1/2)\; x^\mathsf{T} A x + b^\mathsf{T} x \tag{19}$$

- moreover, we have [3]

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\| &\leq \lambda_{\max}(A)\; \|x - y\| & (20)\\ &\leq \|A\|_* \|x - y\| & (21) \end{aligned}$$

where $\|\cdot\|_*$ is any matrix norm

[3] See p. 497, *Matrix Analysis and Applied Linear Algebra* by C. D. Meyer, 2000.

- thus, with any $\rho > \|A\|_*$, we have

$$f(x) \leq f\big(x^{(n)}\big) + \nabla f\big(x^{(n)}\big)^{\mathsf{T}}\big(x - x^{(n)}\big) + \frac{\rho}{2}\|x - x^{(n)}\|^2 \quad (22)$$
$$= g\big(x|x^{(n)}\big) \quad (23)$$

- x can be updated as

$$x^{(n+1)} = x^{(n)} - (1/\rho)\nabla f\big(x^{(n)}\big) \quad (24)$$
$$= x^{(n)} - (1/\rho)\big(Ax^{(n)} + b\big) \quad (25)$$

- what if A is positive semidefinite? indefinite? not symmetric?

Jacobi Iterations

- symmetric matrix D and N with positive definite $D + N$
- we have to find the solution $x^\star$ for $(D+N)x = -b$
- note that

$$x^\star = \arg\min_x \; f(x) = (1/2)\; x^\mathsf{T}(D+N)x + b^\mathsf{T}x \tag{26}$$

- e.g., computing the Newton step for

$$f(x) = \sum_{i=1}^{N} f_i(x_i) + r(Ax - b) \tag{27}$$

$x \in \mathbb{R}^M$, $A \in \mathbb{R}^{p \times M}$, $b \in \mathbb{R}^p$, r is a regularization function

  - Newton's step $\Delta x_{\mathtt{nt}}$ for f at x is given by

$$\nabla^2 f(x) \Delta x_{\mathtt{nt}} = -\nabla f(x) \tag{28}$$

  - the Hessian $\nabla^2 f(x)$ is of the form

$$\nabla^2 f(x) = D + N \tag{29}$$

  - identify D and N $\rightarrow$ structured matrix $+N$

- e.g., continues ..
  - if N is low rank $\rightarrow$ things can be handled efficiently
  - if N is not 'sufficiently' low rank $\rightarrow$ apply MM principle
- find $L > \|N\|_*$ and majorize $(1/2)\ x^\mathsf{T} N x$
- thus f is majorized

- more specifically, we have

$$
\begin{aligned}
f(x) &= (1/2)\ x^{\mathsf{T}}(D+N)x + b^{\mathsf{T}}x && (30)\\
&= (1/2)\ x^{\mathsf{T}}Dx + b^{\mathsf{T}}x + (1/2)\ x^{\mathsf{T}}Nx && (31)\\
&\leq (1/2)\ x^{\mathsf{T}}Dx + b^{\mathsf{T}}x + g(x|x^{(n)}) && (32)
\end{aligned}
$$

since $(1/2)\ x^{\mathsf{T}}Nx \leq g(x|x^{(n)})$, where

$$g(x|x^{(n)}) = (\rho/2)\|x - x^{(n)}\|^2 + x^{(n)\mathsf{T}}Nx - (1/2)x^{(n)\mathsf{T}}Nx^{\mathsf{T}}$$

- x can be updated as

$$x^{(n+1)} = (D + \rho I)^{-1} \left(\rho x^{(n)} - N x^{(n)} - b \right) \tag{33}$$

- $(D + \rho I)$ is efficiently invertible $\rightarrow$ has a rich structure
  - e.g., block diagonal

Nonnegative Quadratic Programming

The Related Problem

- consider the following problem

$$\begin{array}{ll} \text{minimize} & (1/2)x^{\mathsf{T}}Rx + s^{\mathsf{T}}x \\ \text{subject to} & Cx \preceq d \end{array}$$

- $x \in \mathbb{R}^N$, positive definite R, suppose it has some structure

- $C \in \mathbb{R}^{p \times N}$, $d \in \mathbb{R}^p$

- Lagrangian L is given by

$$L(x,\mu) = (1/2)x^{\mathsf{T}}Rx + s^{\mathsf{T}}x + \mu^{\mathsf{T}}(Cx - d) \tag{34}$$

$\mu \in \mathbb{R}^p$

- minimizer $x(\mu)$ of the Lagrangian is given by

$$x(\mu) = -R^{-1}(s + C^{\mathsf{T}}\mu) \tag{35}$$

- dual function h is given by

$$h(\mu) = -(1/2)\mu^{\mathsf{T}} P\mu + q^{\mathsf{T}}\mu + r \tag{36}$$

where

$$P = CR^{-1}C^{\mathsf{T}}, \quad q = -(d + CR^{-1}s), \quad r = -\frac{1}{2}s^{\mathsf{T}}R^{-1}s$$

- P lacks any structure even if R does

Dual Problem

- dual problem is given by

$$\begin{array}{ll} \text{maximize} & h(\mu) = -(1/2)\mu^\mathsf{T} P\mu + q^\mathsf{T}\mu + r \\ \text{subject to} & \mu \succeq 0 \end{array}$$

- how to solve the dual problem?
  - interior-point methods [4] $\rightarrow$ recall P lacks any structure
    - not easily implemented for large P
  - coordinate descent
  - MM principle

[4] See § 11.3.1 of *Convex Optimization* by S. Boyd and L. Vandenberghe, 2004.

Coordinate Descent to Solve the Dual

- restrict h to a line $\mathcal{L}_i = \left\{\mu^{(n)} + te_i \;\middle|\; t \in \mathbb{R}\right\}$
- minimize h over the restriction $\mathcal{L}_i$ [5]

$$\underset{t \in \mathbb{R}}{\text{maximize}} \quad h_i(t) = h(\mu^{(n)} + te_i)$$

- we can unfold $h(\mu^{(n)} + te_i)$ in a straightforward manner, i.e., [6]

$$\begin{aligned} h_i(t) &= -(1/2)(\mu^{(n)} + te_i)^{\mathsf{T}} P(\mu^{(n)} + te_i) + q^{\mathsf{T}}(\mu^{(n)} + te_i) \\ &= -(P_{ii}/2)\; t^2 + \left(q_i - \textstyle\sum_{k=1}^{p} P_{ik}\mu_i^{(n)}\right) t + \text{irrelevant const.} \end{aligned}$$

[5] Let us first ignore the constrain $\mu \succeq 0$ and assimilate it later.

[6] The constant r is dropped since it is irrelevant.

- compute the derivative h_i' of h to determine $t^\star$, i.e.,

$$-P_{ii}t+\left(q_i-\sum_{k=1}^{p}P_{ik}\mu_i^{(n)}\right)=0 \implies t^\star=\frac{q_i-\sum_{k=1}^{p}P_{ik}\mu_i^{(n)}}{p_{ii}}$$

- so the ith coordinate of current $\mu^{(n)}$ is updated as

$$\begin{aligned}\mu_i^{(n)} &:= \mu_i^{(n)} + t^\star \\ &= \mu_i^{(n)} + \frac{1}{p_{ii}}\left(q_i-\sum_{k=1}^{p}P_{ik}\mu_i^{(n)}\right)\end{aligned}$$

- $\mu \succeq 0$ can be assimilated as?

$$\mu_i^{(n)} := \max\left\{0, \mu_i^{(n)} + \frac{1}{p_{ii}}\left(q_i-\sum_{k=1}^{p}P_{ik}\mu_i^{(n)}\right)\right\}$$

- iterate from $i = 1$ to $i = p$ and cycles back to $i = 1$

Algorithm 1 Coordinate Decent

Input: $\mu^{(0)} \succeq 0$, $n = 0$

1: **while** `a stopping criterion true` **do**
2: **for** $i \leftarrow 1$ to p **do**
3: $\mu_i^{(n)} \leftarrow \max \left\{ 0, \mu_i^{(n)} + \frac{1}{p_{ii}} \left(q_i - \sum_{k=1}^{p} P_{ik} \mu_i^{(n)} \right) \right\}$
4: **end for**
5: $\mu^{(n+1)} = \mu^{(n)}$ and $n \leftarrow n + 1$
6: **end while**
7: **return** $\mu^{(n+1)}$

- then the solution is given by (35) with $\mu = \mu^{(n+1)}$

- recall the objective function [7]

$$
\begin{aligned}
h(\mu) = & -\tfrac{1}{2}\textstyle\sum_{i=1}^{N} P_{ii}\mu_i^2 + \sum_{i=1}^{N} q_i\mu_i \\
& - \tfrac{1}{2}\textstyle\sum_{\{i,j|i\neq j, P_{ij}\geq 0\}} P_{ij}\mu_i\mu_j - \tfrac{1}{2}\sum_{\{i,j|i\neq j, P_{ij}<0\}} P_{ij}\mu_i\mu_j \\
= & -\tfrac{1}{2}\textstyle\sum_{i=1}^{N} P_{ii}\mu_i^2 + \sum_{i=1}^{N} q_i\mu_i \\
& - \tfrac{1}{2}\textstyle\sum_{\{i,j|i\neq j, P_{ij}\geq 0\}} P_{ij}\mu_i\mu_j + \tfrac{1}{2}\sum_{\{i,j|i\neq j, P_{ij}<0\}} |P_{ij}|\mu_i\mu_j \\
\geq & -\tfrac{1}{2}\textstyle\sum_{i=1}^{N} P_{ii}\mu_i^2 + \sum_{i=1}^{N} q_i\mu_i \\
& - \tfrac{1}{2}\sum_{\{i,j|i\neq j, P_{ij}\geq 0\}} P_{ij} \left[\frac{\mu_i^{(n)}}{2\mu_j^{(n)}}\,\mu_j^2 + \frac{\mu_j^{(n)}}{2\mu_i^{(n)}}\,\mu_i^2\right] \\
& + \tfrac{1}{2}\sum_{\{i,j|i\neq j, P_{ij}<0\}} |P_{ij}|\,\mu_i^{(n)}\mu_j^{(n)} \left[1 + \ln\left(\frac{\mu_i}{\mu_i^{(n)}}\right) + \ln\left(\frac{\mu_j}{\mu_j^{(n)}}\right)\right] \\
= & \; g\big(\mu|\mu^{(n)}\big)
\end{aligned}
$$

[7] The constant r is dropped since it is irrelevant.

- here the last inequality follows from

$$\mu_i \mu_j \leq \frac{\mu_i^{(n)}}{2\mu_j^{(n)}} \; \mu_j^2 + \frac{\mu_j^{(n)}}{2\mu_i^{(n)}} \; \mu_i^2$$

and

$$-\mu_i \mu_j \leq -\mu_i^{(n)} \mu_j^{(n)} \left[1 + \ln\left(\frac{\mu_i}{\mu_i^{(n)}} \right) + \ln\left(\frac{\mu_j}{\mu_j^{(n)}} \right) \right]$$

- compute the derivative $g'(\,\cdot\,|\mu^{(n)})$ of $g'(\,\cdot\,|\mu^{(n)})$ to yield

$$\begin{aligned}
&\left[\sum_{\{i|P_{ki}>0\}} \big(\mu_i^{(n)}/\mu_k^{(n)}\big) P_{ki}\right] \mu_k \\
&\qquad\qquad\qquad - \left[\sum_{\{i|P_{ki}<0\}} \mu_i^{(n)} \mu_k^{(n)} |P_{ki}|\right] \frac{1}{\mu_k} - q_k = 0 \\
&\implies \alpha\mu_k^2 + \beta\mu_k + \gamma \quad \text{form} \\
&\implies \text{take the positive root as } \mu_k^{(n+1)}
\end{aligned}$$

- in particular we get

$$\mu_k^{(n+1)} = \frac{q_k + \sqrt{q_k^2 + 4\left[\sum_{\{i|P_{ki}>0\}} \mu_i^{(n)} P_{ki}\right]\left[\sum_{\{i|P_{ki}<0\}} \mu_i^{(n)} |P_{ki}|\right]}}{\left[\sum_{\{i|P_{ki}>0\}} \mu_i^{(n)} P_{ki}\right]} \tag{37}$$

- iterates can be perform in parallel

Algorithm 2 MM Principle

Input: $\mu^{(0)} \succeq 0$, $n = 0$

while `a stopping criterion true` **do**
    $\forall\ k,\ \ \mu_k^{(n+1)}$ is computed from (37) and $n \leftarrow n + 1$
end while
return $\mu^{(n+1)}$

- then the solution is given by (35) with $\mu = \mu^{(n+1)}$
- main differences between MM based algorithm and the coordinate descent?

Arithmetic-Geometric Mean Inequality

A Majorization to Monomials

- weighted arithmetic-geometric mean inequality

$$\prod_{i=1}^{p} x_i^{\alpha_i} \leq \sum_{i=1}^{p} \alpha_i x_i \quad \text{for all} \quad x \succeq 0 \tag{38}$$

- α_i are given, $\alpha_i > 0$ [8] and $\sum_i \alpha_i = 1$
- (38): a majorization to $\prod_{i=1}^{p} x_i^{\alpha_i}$ at $\{\gamma 1 \in \mathbb{R}^p \mid \gamma \in \mathbb{R}_+\}$
- a majorization function to $\prod_{i=1}^{p} x_i^{\beta_i}$ at
  - arbitrary $y \succeq 0$ when $\beta \succ 0$?

[8] If $\alpha_i = 0$, the corresponding x_i is irrelevant.

A General Majorization Function

- let $\beta_{\texttt{sum}} = \sum_i \beta_i$
- substitute $x_i \leftarrow (x_i/y_i)^{\beta_{\texttt{sum}}}$ and $\alpha_i \leftarrow \beta_i/\beta_{\texttt{sum}}$ in (38)
- thus, we get

$$\begin{aligned}\prod_{i=1}^{p} x_i^{\beta_i} &\leq \left[\prod_{i=1}^{p} y_i^{\beta_i}\right] \left[\sum_{i=1}^{p} \frac{\beta_i}{\beta_{\texttt{sum}}} \left(\frac{x_i}{y_i}\right)^{\beta_{\texttt{sum}}}\right] \quad \text{for all} \quad x \succeq 0 \\ &= g(x|y) \quad \text{for all} \quad x \succeq 0\end{aligned}$$

A Minorization to Monomials

- we rely on the supporting hyperplane inequality

$$\log z \leq z - 1 \quad \text{for all} \quad z \in \mathbb{R}_{++} \tag{39}$$

- suppose $\beta \succeq 0$ is given, $x_i > 0$
- substitute $z = \prod_{i=1}^{p} (x_i / y_i)^{\beta_i}$ in (39), i.e.,

$$\begin{aligned} \prod_{i=1}^{p} x_i^{\beta_i} &\geq \prod_{i=1}^{p} y_i^{\beta_i} \left[1 + \sum_{i=1}^{p} \beta_i \ln (x_i / y_i)\right] \quad \text{for all} \quad x \succ 0 \\ &= \prod_{i=1}^{p} y_i^{\beta_i} \left[1 + \sum_{i=1}^{p} \beta_i \ln x_i - \sum_{i=1}^{p} \beta_i \ln y_i\right] \\ &= g(x|y) \end{aligned}$$

Examples

A Majorization to A Signomial

- consider the signomial f

$$f(x) = \frac{1}{x_1^3} + \frac{3}{x_1 x_2^2} + x_1 x_2 - \sqrt{x_1 x_2} \tag{40}$$

- majorization function to f at y?
  - $1/(x_1 x_2^2) \le y_1^2/(3y_2^2 x_1^3) + (2y_2)/(3y_1 x_2^3)$
  - $x_1 x_2 \le (y_2 x_1^2)/(2y_1) + (y_1 x_2^2)/(2y_2)$
  - $\sqrt{x_1 x_2} \ge (1/2)\sqrt{y_1 y_2}\left(2 + \ln x_1 + \ln x_2 - \ln y_1 - \ln y_2\right)$

Appendices

Composition with Affine Function

- suppose $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable
- then define $\tilde{f} : \mathbb{R} \to \mathbb{R}$ by

  $$\tilde{f}(\tau) = f(a + \tau b)$$

  is differentiable and [9]

  $$\tilde{f}'(\tau) = \frac{d\tilde{f}(\tau)}{d\tau} = \nabla f(a + \tau b)^{\mathsf{T}} b \qquad (41)$$

[9] see § A.4.2 of Convex Optimization by S. Boyd and L. Vandenberghe, 2004.

Composition with Affine Function

- suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable
- then $\tilde{f}$ [*cf.* (41)] is twice differentiable and [10]

$$\tilde{f}''(\tau) = \frac{d^2 \tilde{f}(\tau)}{d\tau^2} = b^\mathsf{T} \nabla^2 f(a + \tau b) b \tag{42}$$

[10] see § A.4.4 of Convex Optimization by S. Boyd and L. Vandenberghe, 2004.

Newton-Leibniz Formula

- recall $\tilde{f}(\tau) = f(a + \tau b)$
- let us apply Newton-Leibniz formula [11]

$$\tilde{f}(1) = \tilde{f}(0) + \int_0^1 \tilde{f}'(t)\ dt \tag{43}$$

- thus from (41), (43) becomes

$$f(a+b) = f(a) + \nabla f(a)^{\mathsf{T}} b + \int_0^1 \nabla f(a+tb)^{\mathsf{T}} b\ dt \tag{44}$$

[11] Based on elementary classical analysis.

Taylor with the Integral Remainder

- Taylor formula with the integral remainder
- recall $\tilde{f}(\tau) = f(a + \tau b)$
- we have [12]

$$\tilde{f}(1) = \tilde{f}(0) + \tilde{f}'(0) + \int_0^1 \int_0^t \tilde{f}''(\tau)\ d\tau dt \qquad (45)$$

- thus from (41) and (42), (45) becomes

$$f(a+b) = f(a)+\nabla f(a)^{\mathsf{T}} b+ \int_0^1 \int_0^t b^{\mathsf{T}} \nabla^2 f(a+\tau b) b\, d\tau dt \qquad (46)$$

[12] Based on elementary classical analysis.