Majorization-Minimization Algorithms

Pradip Sasmal

Indian Institute of Science, Bangalore

September 22, 2018

Organization

- Basics of MM algorithm
- Convergence of MM algorithm
- Surrogate function construction
- Application

Overview of the MM Algorithm

- The MM algorithm is not an algorithm, but a prescription for constructing optimization algorithms.
- The EM algorithm from statistics is a special case.
- An MM algorithm operates by creating a surrogate function that minorizes or majorizes the objective function. When the surrogate function is optimized, the objective function is driven uphill or downhill as needed.
- In minimization MM stands for majorize/minimize, and in maximization MM stands for minorize/maximize.

MM Principle

- generate an algorithm that avoids matrix inversion
- separate the parameters of a problem
- linearize an optimization problem
- deal with equality and inequality constraints
- turn a non-differentiable problem into a smooth problem
- the existence of a closed-form optimizer

MM Algorithm

- Optimization prob:

$$\min_{x \in \chi} f(x)$$

  where χ :nonempty closed set in $\mathbb{R}^n$ and $f : \chi \to \mathbb{R}$ continuous function.
- Initialized as $x_0 \in \chi$, MM generates a sequence of feasible points $(x_t)_{t \in \mathbb{N}}$
- $g(x|x_t) : \chi \to \mathbb{R}$ is said to majorize the function $f(x)$ at x_t provided $f(x_t) = g(x_t|x_t)$ and$f(x) \leq g(x|x_t) \quad \forall x \in \chi$
- The majorization relation between functions is closed under the formation of sums, nonnegative products, limits, and composition with an increasing function.
- $x_{t+1} = \arg\min_{x \in \chi} g(x|x_{t+1})$
- A function $g(x|x_t)$ is said to minorize the function $f(x)$ at x_t provided $-g(x|x_t)$ majorizes $-f(x)$

Convergence of MM

- $f(x_{t+1}) \leq g(x_{t+1}|x_t) \leq g(x_t|x_t) = f(x_t) \Rightarrow (f(x_t))$ is nonincreasing and converges to a limit f^* by the assumption that f is bounded below.
- establish the conditions that guarantee f^* being a stationary value and also the convergence of the sequence $(x_t)_{t \in \mathbb{N}}$.
- The convexity of χ and continuity of f are minimum assumptions for a unified study of algorithm convergence.

Unconstrained Optimization:Assumptions

- $(A1)$ The sublevel set $\text{lev}_{\leq f(x_0)} f := \{x \in \chi | f(x) \leq f(x_0)\}$ is compact given that $f(x_0) < \infty$
- $(A2.1)$ $f(x)$ and $g(x|x_t)$ are continuously differentiable with respect to x
- $(A3.1)$ $g(x|x_t)$ is continuous in x and x_t .
- the set of stationary points of f is defined as

$$\chi^* = \{x \in \chi | \nabla f(x) = 0\}$$

Unconstrained Optimization:Results

- $(C1)$ Any limit point x_∞ of (x_t) is a stationary point of f
- $(C2)$ $f(x_t) \downarrow f^*$ monotonically and $f(x) = f^*$ with $x \in \chi^*$
- $M : \mathbb{R}^n \to \mathbb{R}^n \Rightarrow x_t \longmapsto x_{t+1}$
- $(C3)$ If $f(M(x)) = f(x)$, then $x \in \chi^*$ and $x \in \arg\min g(\cdot|x)$
- $(C4)$ If x is a fixed point of M, then x is a convergent point of MM and belongs to χ^*

Assumptions

- convergence of sequence $(x_t)_{t\in\mathbb{N}}$ to a stationary point
- $(A4.1)$ Set χ^* is a singleton;
- $(A4.2)$ Set is χ^* discrete and $\|x_{t+1} - x_t\| \to 0$
- $(A4.3)$ Set χ^* is discrete, and $g(\cdot|x)$ has a unique global minimum for all $x \in \chi^*$

Constrained Optimization with Smooth Objective Function:

- With χ convex and f continuously differentiable, the set of stationary points is defined as

$$\chi^* = \{x | \nabla f(x)^T (y - x) \geq 0, \, y \in \chi\}$$

- Conclusions $(C1) - (C4)$ still hold under Assumptions $(A1), (A2.1)$ and $(A3.1)$
- $(A3.1)$ can be replaced by $(A3.2)$ For all x_t generated by the algorithm, there exists $\gamma > 0$ such that $\forall x$

$$(\nabla g(x|x_t) - \nabla g(x_t|x_t))^T (x - x_t) \leq \gamma \|x - x_t\|^2$$

- Assumption $(A3.2)$ is equivalent to stating that $g(x|x_t)$ can be uniformly upperbounded by a quadratic function with the Hessian matrix being γI, which is easier to verify than $(A3.1)$ when $g(x|x_t)$ has a complicated form3.
- Convergence of sequence $(x_t)_{t \in \mathbb{N}}$ to a stationary point can be proved by further requiring $(A4.1)$ or $(A4.2)$.

Constrained Optimization With Non-Smooth Objective

- f and $g(\cdot|x)$ are nonsmooth, but their directional derivatives exist for all feasible directions.
- The set of stationary points is defined as

$$\chi^* = \{x | f^{'}(x; d) \geq 0, \forall\, x + d \in \chi\}$$

where

$$f^{'}(x_t; d) := \liminf_{\lambda \downarrow 0} \frac{f(x_t + \lambda d) - f(x_t)}{\lambda}$$

is the directional derivative of f at x_t in direction d.

- $(A2.2)$ $f^{'}(x_t; d) = g^{'}(x_t; d|x_t)$
- Under Assumptions $(A1), (A2.2), (A3.1)$, the sequence $(x_t)_{t \in \mathbb{N}}$ converges to χ^* , i.e.,

$$\lim_{t \to \infty} \inf_{x \in \chi^*} \|x - x_t\|_2 = 0$$

First Order Taylor Expansion

- $f(x) = f_0(x) + f_{ccv}(x)$ where f_{ccv} is a differentiable concave function.
- Linearizing f_{ccv} at $x = x_t$ yields the following inequality:

$$f_{ccv}(x) \leq f_{ccv}(x_t) + \nabla f_{ccv}(x_t)^T (x - x_t)$$

- $f(x) \leq f_0(x) + \nabla f_{ccv}(x_t)^T x +$ constant

Example and application

Example: $\log(x) \leq \log(x_t) + \frac{1}{x_t}(x - x_t)$ with equality achieved at $x = x_t$

Reweighted l_1-norm Minimization:

$$\min_x \sum_{i=1}^{n} \log(\epsilon + |x_i|) \quad \text{sub to} \quad y = Ax,\ \epsilon > 0$$

The reweighted l_1-normminimization algorithm solves the above problem by solving

$$\min_x \sum_{i=1}^{n} \frac{|x_i|}{\epsilon + |x_i^t|} \quad \text{sub to} \quad y = Ax$$

at the t-th iteration, which is an MMstep by applying the above inequality to the objective function.

Example

- Given a convex, a linear, and a concave function, f_{cvx}, f_{lin} and f_{ccv} respectively, if their values and gradients are equal at some x_t, then, for any x,

$$f_{cvx} \leq f_{lin} \leq f_{ccv}$$

- **Example:** Function $|x|^p$, $0 < p < 1$, which is concave on $(-\infty, 0]$ and $[0, \infty)$, can be upperbounded as

$$|x|^p = |x_t|^{p-2}x^2 + \text{constant}$$

providing that $x_t \neq 0$.

l_p−Norm Minimization:

- $\min_x \|Ax - b\|_p^p$, where $b \in \mathbb{R}^m$. Construct a quadratic surrogate function:

$$g(x|x_t) = \sum_{i=1}^{m} w_i^t (b_i - A_{i,:}x)^2$$

  where w_i^t is given by

$$w_i^t = |b_i - A_{i,:}x|^{p-2}$$

- Function $g(x|x_t)$ admits a closed-form minimizer

$$x_{t+1} = (A^T W_t A)^{-1} A^T W_t b$$

- A similar idea has been applied in solving the sparse representation problems

$$\min_x \|Ax - b\|_2^2 + \lambda \|x\|_1$$

  and

$$\min_x \|x\|_1 \quad \text{sub to} \quad b = Ax$$

Arithmetic-Geometric Mean Inequality

- **Example:**

$$\prod_{i=1}^{n} x_i^{\alpha_i} \geq \prod_{i=1}^{n} (x_i^t)^{\alpha_i} (1 + \sum_{i=1}^{n} \alpha_i \log x_i - \sum_{i=1}^{n} \alpha_i \log x_i^t)$$

- The arithmetic-geometric mean inequality states that

$$\prod_{i=1}^{n} z_i^{\alpha_i} \leq \sum_{i=1}^{n} \frac{\alpha_i}{\|\alpha\|_1} z_i^{\|\alpha\|_1},$$

  where $z_i, \alpha_i \geq 0$. Equality is achieved when the z_i' are equal.
- Let $z_i = \frac{x_i}{x_i^t}$ for $\alpha_i > 0$ and $z_i = (\frac{x_i}{x_i^t})^{-1}$ for $\alpha_i < 0$

$$\prod_{i=1}^{n} x_i^{\alpha_i} \geq \prod_{i=1}^{n} (x_i^t)^{\alpha_i} \sum_{i=1}^{n} \frac{\alpha_i}{\|\alpha\|_1} (\frac{x_i}{x_i^t})^{\|\alpha\|_1}$$

  Equality is achieved at $x_i = x_i^t \, \forall \, i = 1, \ldots, n$
- Upperbound and lowerbound serve as the basic ingredients for deriving MM algorithms for signomial programming.

Example

- **Example:** A posynomial $\sum_{i=1}^{n} u_i(x)$, where $u_i(x)$ is monomial

$$\sum_{i=1}^{n} u_i(x) \geq \prod_{i=1}^{n} \left(\frac{u_i(x)}{\alpha_i} \right)^{\alpha_i}$$

where $\alpha_i = \frac{u_i(x_t)}{\prod_{i=1}^{n} u_i(x_t)}$. Equality is achieved at $x = x_t$

- **Example**

$$\|x\|_2 \leq \frac{1}{2} \left(\|x_t\|_2 + \frac{\|x\|_2^2}{\|x_t\|_2^2} \right)$$

given that $\|x_t\|_2 \neq 0$. Equality holds at $x = x_t$.

Cauchy-Schwartz inequality

- Cauchy-Schwartz inequality states that

$$x^T y \leq \|x\|_2 \|y\|_2$$

Equality is achieved when x and y are collinear.

- **Example**

$$a^H x \geq \frac{\mathrm{Re}(x_t^H a a^H x)}{|a^H x_t|}$$

given that $a^H x_t \neq 0$. Equality is achieved at $x = x_t$

- **Example**

$$\|x\|_2 \geq \frac{x^T x_t}{\|x_t\|_2}$$

given that $\|x_t\|_2 \neq 0$. Equality is achieved at $x = x_t$

Convexity Inequality

- For a convex function f_{cvx}, we have the following inequality:

$$f_{cvx}(\sum_{i=1}^{n} w_i x_i) \leq \sum_{i=1}^{n} w_i f_{cvx}(x_i)$$

  where $\sum_{i=1}^{n} w_i$, $w_i \geq 0$. Equality can achieved if the x_i's are equal, or for different x_i's if f_{cvx} is not strictly convex.

- **Jensens Inequality:** Let $f : \chi \rightarrow \mathbb{R}$ be a convex function and x be a random variable that take values in χ. Assuming that $\mathbb{E}(x)$ and $\mathbb{E}(f(x))$ are finite, then

$$\mathbb{E}(f(x)) \geq f(\mathbb{E}(x)).$$

- With Jensens inequality we can show that EM is a special case of MM

Example

- $$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \log f_i(x_t) + (\sum_{i=1}^{n} \alpha_i) \log \left(\frac{\sum_{i=1}^{n} \alpha_i \frac{f_i(x)}{f_i(x_t)}}{\sum_{i=1}^{n} \alpha_i} \right)$$

  where $f_i(x) > 0, \alpha_i > 0 \, \forall i$. Equality is achieved at $x = x_t$.

- $$\sum_{i=1}^{n} \alpha_i \log f_i(x) \leq \sum_{i=1}^{n} \alpha_i \Big(\log f_i(x_t) + \frac{1}{f_i(x_t)} (f_i(x) - f_i(x_t)) \Big)$$

- The concave upperbound is tighter, thus is preferred for a faster convergence rate

Construction by Second Order Taylor Expansion

- **Descent Lemma:** Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function with a Lipschitz continuous gradient and Lipschitz constant L. Then, for all $x, y \in \mathbb{R}^n$

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{L}{2}\|x - y\|^2$$

- More generally, if f has bounded curvature, i.e., there exists a matrix M such that $M \geq \nabla^2 f(x)$, $x \in \chi$, then the following inequality implied by Taylors theorem holds:

$$f(x) \leq f(y) + \nabla f(y)^T (x - y) + \frac{1}{2}(x - y)^T M (x - y)$$

- **Example:** For $f(x) = x^H L x$, the following inequality holds

$$x^H L x \leq x^H M x + \mathrm{Re}(x^H (L - M) x_t) + x_t^H (M - L) x_t,$$

where $M \geq L$. Equality holds at $x = x_t$.

Y. Sun, P. Babu, and D. P. Palomar, "Majorization-Minimization Algorithms in Signal Processing, Communications, and Machine Learning," IEEE TRANSACTIONS ON SIGNAL PROCESSING, VOL. 65, NO. 3, FEBRUARY 1, 2017.

Thank You