Majorization Minimization - the Technique of Surrogate

Andersen Ang

Mathématique et de Recherche opérationnelle
Faculté polytechnique de Mons
UMONS
Mons, Belgium

email: manshun.ang@umons.ac.be
homepage: *angms.science*

June 6, 2017

Overview

1. Introduction to Majorization Minimization
   - What is Majorization Minimization
   - Definition of surrogate
2. How to construct surrogate
   - By quadratic upper bound of convex smooth function
   - By quadratic lower bound of strongly convex function
   - By Taylor expasion : 1st and 2nd order
   - By inequalities

What is Majorization Minimization

Majorization Minimization (MM) is an optimization algorithm.

More accurately, MM itself is not an algorithm, but a framework on how to construct an optimization algorithm.

Example of MM : Expectation Minimization (EM-Algorithm).

Another name of MM is "Successive upper bound minimization method".

The idea of MM: Successive upper bound minimization

Want to solve $\min_{x \in \mathcal{Q}} f(x)$

How to solve : construct an iterative algorithm that produces a sequence $\{x_k\}$ such taht the objective function is non-increasing: $f(x_{k+1}) \leq f(x_k)$

Problem: if f is **complicated** $\implies$ cannot handle the problem *directly*

Idea: attack the problem *indirectly*
Generate the sequence $\{x_k\}$ to minimize f by another **simpler** function g such that minimizing g 'helps' minimizing f.

g is called *surrogate function* / *auxiliary function*

How minimizing g 'helps' minimizing f : if g is the upper bound of f

The idea of MM - Successive upper bound minimization

In other words, the idea of MM is:

1. [Original problem is too complicated]. Want to solve $\min_{x \in \mathcal{Q}} f(x)$, but f is too complicated or solving $\min_x f(x)$ directly is too expensive

2a. [Indirect attack of the problem via surrogate]. Finds/constructs a simpler function g that solving $\min_{x \in \mathcal{Q}} g(x)$ is cheaper
2b. Finally, use the solution of $\min_x g(x)$ to solve $\min_{x \in \mathcal{Q}} f(x)$

Questions:

- how to find g?
- how to use the information on g to minimize f?

The surrogate g

Surrogate function $g(x)$ can be defined as a parametric function with the form

$$g(x|\theta)$$

where θ is the parameter

In MM for minimization, θ can be defined as x_k.
i.e. the information about the variable x at the current iteration is used to construct g.

The surrogate helps to minimize f by finding the variable in the next iteration as the minimizer of the current surrogate:

$$x_{k+1} = \arg\min_{x \in \mathcal{Q}} g_k \Big(x | x_k \Big)$$

Overall framework of MM

To solve

$$\min_{x \in \mathcal{Q}} f(x)$$

Use the following surrogate scheme:

---

(1) Initialize x_0
(2) Construct a surrogate function at x_k as $g_k(x|x_k)$
(3) Updating : $x_{k+1} = \arg\min_{x \in \mathcal{Q}} g_k(x|x_k)$
(4) Repeat (2)-(3) until converge

Note that the surrogate function is changing in each iteration, as $g_k(x|x_k)$ depends on the changing variable x_k

Also notice that the process of minimizing g will help minizing f as the sequence $\{x_k\}$ produced satisfies $f(x_{k+1}) \leq f(x_k)$

More questions

But, more exactly:

(1) how to constrcut g?

(2) what is the condition of g ?

(3) how to optimize g ?

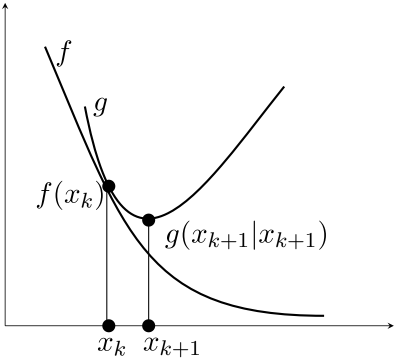
(4) how to know that optimizing g is cheaper than optimizing f?

Definition of surrogate

Two key conditions for surrogate g are:

1. g *majorizes* the original function f at x_k for all other points.
Mathematically, $g_k\left(x|x_k\right) \geq f(x)$, $\forall x, x^k \in \mathcal{Q}$.

2. g *touches* the original function f at x_k at the point $x = x_k$.
Mathematically, $g_k\left(x_k|x_k\right) = f(x_k)$, $\forall x_k \in \mathcal{Q}$



The convergence theorem of MM

Theorem. If the surrogate g satisfies the two conditions :
1. $g(x|x_k) \geq f(x), \forall x$.
2. $g(x_k|x_k) = f(x_k), \forall x_k$.

Then the iterative method $x_{k+1} = \arg\min_{x \in \mathcal{Q}} g_k(x|x_k)$ will produce a sequence $f(x_k)$ that converge to a local optimum. i.e.

$$f(x_{k+1}) \leq f(x_k)$$

Proof.

$$\begin{array}{rcll} f(x_{k+1}) & \leq & g(x_{k+1}|x_k) & \text{by condition 1} \\ & \leq & g(x_k|x_k) & x_{k+1} \text{ minimizes } g \\ & = & f(x_k) & \text{by condition 2} \end{array}$$

□

Construct surrogate by quadratic upper bound of smooth convex function

Suppose f is convex (both f and dom$f$) and β-smooth (∇f is Lipschitz continuous with parameter β).

Then f is bounded above by the following at $x_0 \in$ dom$f$ for all $x \in$ dom f

$$f(x) \leq f(x_0) + \nabla f(x_0)^T(x - x_0) + \frac{\beta}{2}\|x - x_0\|_2^2$$

Surrogate can be defined as such upper bound.

$$g_k(x|x_k) \leq f(x_k) + \nabla f(x_0)^T(x - x_k) + \frac{\beta}{2}\|x - x_k\|_2^2$$

Pros: simple construction of f
Cons: need the knowledge of β

Construct surrogate by quadratic lower bound of strongly convex function

Suppose we want to do Minorization Maximization (the opposite).

Suppose f is α-strongly convex. Then f is bounded below by the following at $x_0 \in \mathrm{dom} f$ for all $x \in \mathrm{dom}\ f$

$$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{\alpha}{2} \|x - x_0\|_2^2$$

Surrogate can be defined as such upper bound.

$$g_k(x|x_k) \geq f(x_k) + \nabla f(x_0)^T (x - x_k) + \frac{\alpha}{2} \|x - x_k\|_2^2$$

Pros: simple construction of f
Cons: need the knowledge of α

Construct surrogate by first order Taylor expansion

Taylor Expansion. Taylor expansion of a differentiable f at a point x_0 is

$$f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + \mathcal{O}$$

where $\mathcal{O}$ is higher order term.

If f is convex, the first order Taylor approximation is a global underestimator of f. i.e. $f(x) \geq f(x_0) + \nabla f(x_0)^T(x - x_0)$. This is useful for Minorization Maximization.

If f is concave, the first order Taylor approximation is a global overestimator of f. i.e. $f(x) \leq f(x_0) + \nabla f(x_0)^T(x - x_0)$. This is useful for Majorization Minimization.

Construct surrogate by majorizing the second order Taylor expansion

Consider the Taylor expansion at x_0 again

$$f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + \mathcal{O}$$

Suppose f is twice differentiable. Now express explicitly the higher order term $\mathcal{O}$ in Lagrangian form

$$\mathcal{O} = \frac{1}{2}(x - x_0)^T \nabla^2 f(\xi)(x - x_0)$$

where ξ is some constant (by mean value theorem).

Taylor expansion of f at point x_0 becomes

$$f(x) = f(x_0) + \nabla f^T(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T \nabla^2 f(\xi)(x - x_0)$$

Construct surrogate by majorizing the second order Taylor expansion

One can construct a surrogate in the following form

$$g(x|x_0) = f(x_0) + \nabla f^T(x_0)(x - x_0) + \frac{1}{2}(x - x_0)^T M(x - x_0)$$

if $M \succeq \nabla^2 f(x), \forall x$

The key is $M \succeq \nabla^2 f(x)$, so if $M - \nabla^2 f(x)$ is positive semi-definite $\forall x$ (including the case $x = \xi$), then

$$g(x|x_0) - f(x) = \frac{1}{2}(x - x_0)^T \Big(M - \nabla^2 f(\xi)\Big)(x - x_0) \geq 0$$

Hence g majorizes f.

How to form M: $M = \nabla^2 f + \delta I$

Construct surrogate by majorizing the second order Taylor expansion - Least Sqaure Example

Consider $f(x) = \|Ax - b\|^2$ (F-norm or 2-norm)

$$
\begin{aligned}
\|Ax - b\|^2 &= (Ax - b)^T (Ax - b) \\
&= x^T A^T Ax - x^T A^T b - b^T Ax + b^T b \\
&= x^T A^T Ax - 2x^T A^T b + b^T b \\
\nabla_x \|Ax - b\|^2 &= 2A^T Ax - 2A^T b \\
&= 2A^T (Ax - b) \\
\nabla_x^2 \|Ax - b\|^2 &= 2A^T A
\end{aligned}
$$

2nd order Taylor expansion of $f(x) = \|Ax - b\|^2$ is

$$f(x) = f(x_0) + 2A^T \left(Ax_0 - b\right) \left(x - x_0\right) + 2(x - x_0)^T A^T A(x - x_0)$$

Thus the following g majorizes f

$$g(x|x_0) = f(x_0) + 2A^T \left(Ax_0 - b\right) \left(x - x_0\right) + 2(x - x_0)^T M(x - x_0)$$

where $M \succeq A^T A$ is a diagonal matrix. A simple way to construct M is $M = A^T A + \delta I$ with $\delta > 0$

Construct surrogate by majorizing the second order Taylor expansion - NMF Example

In Non-negative Matrix Factorization, we have

$$f(W,h) = \|Wh - x\|_2^2$$

where x is given and W, h are variable.
2nd order Taylor expansion of $f(W,h)$ is

$$f(W,h) = f(h_0) + 2W^T\left(Wh_0 - x\right)\left(h - h_0\right) + 2(h-h_0)^T W^T W (h - h_0)$$

We can construct M as $M = \text{Diag}\left(\frac{[W^T W h]_i}{[h]_i}\right)$, then $M \succeq W^T W$ and

$$g(W,h) = f(h_0) + 2W^T\left(Wh_0 - x\right)\left(h - h_0\right) + 2(h-h_0)^T M (h - h_0)$$

For detail: see the slides "Convergence analysis of NMF algorithm", and the original paper by Lee and Seung 2001

Construct surrogate by inequalities

Jensen's inequality. If f is convex, then

$$f\Big(\sum_i \lambda_i t_i\Big) \le \sum_i \lambda_i f(t_i)$$

where $\lambda_i \ge 0$ and $\sum_i \lambda_i = 1$.

Example. Let $t_i = \frac{c_i}{\lambda_i}(x_i - y_i) + c^T y$, then

$$\begin{aligned}
\lambda^T t &= \sum_i \lambda_i t_i \\
&= \sum_i \big(c_i x_i - c_i y_i + \lambda_i c^T y\big) \\
&= \sum_i c_i x_i - \sum_i c_i y_i + \big(\sum_i \lambda_i\big) c^T y \\
&= c^T x - c^T y + c^T y \\
&= c^T x
\end{aligned}$$

Construct surrogate by inequalities

By Jensen's inequality $f\Big(\sum_i \lambda_i t_i\Big) \leq \sum_i \lambda_i f(t_i)$

$$\begin{aligned} f\Big(\lambda^T t\Big) &\leq \sum_i \lambda_i f(t_i) \\ &= \sum_i \lambda_i f\Big(\frac{c_i}{\lambda_i}(x_i - y_i) + c^T y\Big) \end{aligned}$$

As $\lambda^T t = c^T x$ thus

$$f\big(c^T x\big) \leq \sum_i \lambda_i f\Big(\frac{c_i}{\lambda_i}(x_i - y_i) + c^T y\Big)$$

So for a convex function f, the surrogate function of $f(c^T x)$ is $g(x|y) = \sum_i \lambda_i f\Big(\frac{c_i}{\lambda_i}(x_i - y_i) + c^T y\Big)$, with $\sum_i \lambda_i = 1$

Construct surrogate by inequalities

Example. If c, x and y are all positive, let $t_i = \frac{x_i}{y_i} c^T y$ and $\lambda_i = \frac{c_i y_i}{c^T y}$ (hence again $\lambda^T t = c^T x$) then by Jensen's inequality

$$\begin{aligned} f\left(\lambda^T t\right) &\leq \sum_i \lambda_i f(t_i) \\ &= \sum_i \frac{c_i y_i}{c^T y} f\left(\frac{x_i}{y_i} c^T y\right) \end{aligned}$$

Hence

$$f(c^T x) \leq \sum_i \frac{c_i y_i}{c^T y} f\left(\frac{x_i}{y_i} c^T y\right)$$

So for a convex function f, the surrogate function of $f(c^T x)$, with positive c and x is $g(x|y) = \sum_i \lambda_i f\left(\frac{c_i}{\lambda_i}(x_i - y_i) + c^T y\right)$, where $\sum_i \lambda_i = 1$ and y has to be positive.

Advantage of the surrogate in the two examples : g are separable and thus parallelizable.

Construct surrogate by inequalities

Other inequalities :

Cauchy-Schwartz Inequality

$$|u^T v| \leq \|u\| \|v\|$$

Arithmetic-Geometric Mean

$$\Big(\prod_i^n x_i \Big)^{\frac{1}{n}} \leq \frac{1}{n} \sum_i^n x_i$$

Chebyshev, Hölder, and so on...

Last page - summary

- Introduction of Majorization Minimization
  (1) Initialize x_0
  (2) Construct a surrogate function at x_k as $g_k\left(x|x_k\right)$
  (3) Updating : $x_{k+1} = \arg\min_{x\in\mathcal{Q}} g_k\left(x|x_k\right)$
  (4) Repeat (2)-(3) until converge
- The surrogate funcion
  1. g *majorizes* the original function f at x_k for all other points. Mathematically, $g_k\left(x|x_k\right) \geq f(x)$, $\forall x, x^k \in \mathcal{Q}$.
  2. g *touches* the original function f at x_k at the point $x = x_k$.
- Construction of surrogate function via various methods

End of document