Bundle Method for Minimizing a Nonsmooth Function

James Cameron

Department of Mathematical Sciences
George Mason University
Fairfax, Virginia 22030, USA

Presentation of work by
Helga Schramm and Jochem Zowe

17 October 2013

Purpose of the Talk

Given a nonsmooth functional, i.e, $f \notin C^1$, the goal is:

$$\min_{\mathbb{R}^n} f(x) \text{ where } f : \mathbb{R}^n \mapsto \mathbb{R}$$

Additional constraints, if necessary, can theoretically be added as penalty terms.

Develop an algorithm by adding trust region ideas to the bundle idea.

Motivation

- Compared to earlier bundle method algorithms, this method is more stable.
- Finding lower bounds to symmetric traveling salesman problems.
- Finding the maximum eigenvalue of a real symmetric matrix, if the multiplicity of the maximal eigenvalue is greater than 1.
- Maximization of the contact area between a clamped beam and a rigid obstacle.

Definitions

$\partial f(x) = conv\{g \in \mathbb{R}^n : g = \lim_{i \mapsto \infty} \nabla f(x_i), x_i \mapsto x, \nabla f(x_i)$ exists/converges $\}$

is the subdifferential of f, elements are called subgradients.

From convex analysis, this set is a well-defined, nonempty, convex, and compact subset of $\mathbb{R}^n$

It is assumed f is locally Lipschitz.

Subgradiant Method

$x_{k+1} = x_k + \lambda_k d_k$

where $d_k = -\frac{g_k}{||g_k||}$ and λ_k is a scalar

Problems: no guaranteed descent, no stopping critera, terrible convergence

Bundle Method

Let $d = x - x_k$, then the cutting plane approximation (CP) of $f(x_k)$ is

$$\max_{i \in [1,k]} \{ g_i^T d + g_i^T (x_k - x_i) + f(x_i) \}$$

Problems: poor approximation if $f(x_i)$ is far from x_k

Solution: Introduce t_k such that $CP + \frac{1}{2t_k}||d||^2$ is a descent direction on d_k

First iteration idea

Iterating from $x_k \mapsto x_{k+1}$

1. Compute $d_k = \min_{d \in \mathbb{R}^n} \{CP + \frac{1}{2t_k} ||d||^2\}$
2. Set $\lambda > 0$, line search on $x_k + \lambda d_k$
    1. If the line search "sufficently decreased" f, make a "serious step": $x_{k+1} = x_k + \lambda d_k$ and and $g_{k+1} \in \partial f(x_{k+1})$
    2. If not, make a "null step": $x_{k+1} = x_k$ and and $g_{k+1} \in \partial f(x_{k+1})$
3. Stop if d_k is sufficently close to zero.

Problems: Bad choice of t_k leads to a breakdown, and we only have linear convergence.

Trust Region

The idea behind the trust region concept is to change t_k systematically.

Improve CP by null steps only.

Conclusion: t_k-ball and CP so that we always have sufficent decreases.

Note that this gets rid of the line search step.

New iteration idea

Iterating from $x_k \mapsto x_{k+1}$

1. Compute $d_k = \min_{d \in \mathbb{R}^n} \{CP + \frac{1}{2t_k}||d||^2\}$
2. If $f(x_k + d_k)$ is sufficently smaller than $f(x_k)$, then
   1. increase t_k, return to 1, or
   2. make a serious step
3. If $f(x_k + d_k)$ is not sufficently smaller than $f(x_k)$, then
   1. decrease t_k, return to 1, or
   2. make a null step
4. Stop if d_k is sufficently close to zero.

Convex Case

If f is convex then

$$g \in \partial f(x) \iff g^T(y - x) \leq f(y) - f(x) \;\; \forall y \in \mathbb{R}^n$$

Need exact formula for CP, exact conditions for our iteration idea, and an overall algorithm.

Exact Cutting Plane Approximation

At x_k we have all previous x-values, assign $y_i = x_i$ and $g_i \in \partial f(y_i)$ for $1 \leq i \leq k$.

Then the CP model is $max_{1 \leq i \leq k}\{g_i^T(x - y_i + f(y_i))\}$ with error

$\alpha_{i,k} = f(x_k) - (g_i^T(x - y_i + f(y_i))$

Let $d = x - x_k$, then CP is $max_{1 \leq i \leq k}\{g_i^T d - \alpha_{k,i}\}$

So step one of the iteration idea is $d = \min_{d \in \mathbb{R}^n}\{CP + \frac{1}{2t_k}||d||^2\}$.

Which is the same as $d = \min_{d \in \mathbb{R}^n}\{v + \frac{1}{2t_k}||d||^2 : v \geq CP\}$

Remarks

This is a strictly convex problem, so

- There are unique minimal d(t), v(t) such that they depend continuously on t.
- $\alpha_{k,i} \geq 0$
- $\alpha_{k,i}$ in some sense "measures" how well g_i satisfies the convexity inequality.

For notation, J_k is an index set that represents how many subgradients are being carried along, in the overall algorithm we will want to reset this occasionaly due to memory concerns.

Inner Iteration

1. Fix an upper bound T for t, stopping parameter $\varepsilon \geq 0$, $t_1 = t_{k-1}$, $0 < m_1 < m_2 < 1$, $0 < m_3 < 1$.
2. Compute v_j and d_j from CP, where j is an index starting at 1.
3. Two possible outcomes:
    1. $\frac{1}{t_j}||d_j|| \leq \varepsilon$ and $-\frac{1}{t_j}||d_j||^2 - v_j \leq \varepsilon$, stop
    2. Else, $y_j = x_k + d_k$ and find $g_k \in \partial f(y_j)$
        1. If SS1 and SS2 hold, serious step and stop
        2. If SS1 and not SS2 hold, iterate $t_{j+1} = \frac{1}{2}(T + t_j)$ and recompute, increase j
        3. If NS1 and NS2 hold, null step and stop
        4. If NS1 and not NS2 hold, iterate $t_{j+1} = \frac{1}{2}(T + t_j)$ and recompute, increase j

Conditions

SS1 is $f(y_j) - f(x_k) \leq m_1 v_j$

SS2 is $g_j^T d_j \geq m_2 v_j$

NS1 is $f(y_j) - f(x_k) \geq m_1 v_j$

NS2 is $\alpha_{k,j} \leq m_3$

By theorem, the inner iteration algorithm is finite and ends with either ε-optimal solution, a serious step, or a null step.

Overall Algorithm

1. Choose x_1, $T > 0$, stopping parameter $\varepsilon \geq 0$, $0 < m_1 < m_2 < 1$, $0 < m_3 < 1$, upper bound $J_{max} \geq 3$ for $|J_k|$.
2. Find $f(x_1)$, $g_1 \in \partial f(x_1)$, set $y_1 = x_1$, set $J_1 = \{1\}$, set $k = 1$.
3. Run Inner Interation, if ε-optimal, stop.
4. If $|J_k| = J_{max}$:
    1. Reset J by choosing a subset of J_k with 2 less elements.
5. Else:
    1. If serious step: $\alpha_{k+1,i} = \alpha_{k,i} + f(x_{k+1}) - f(x_k) - g_i^T d_k$
    2. If null step: $\alpha_{k+1,i} = \alpha_{k,i}$
    3. Either way, set $J_{k+1} = J_k \cup \{k+1\}$

Remarks

Convergence is proved thorugh analysis on 6 different sequences based off of total number of serious steps taken, omitted for brevity.

For f non-convex, our convexity inequality doesn't hold and $\alpha < 0$ is possible.

Repace $\alpha_{k,i}$ wth $\beta_{k,i} = \max\{\alpha_{k,i}, c_0||x_k - y_i||^2\}$ for some $c_0 > 0$ small.

Remarks

Replace SS2 with NS3, which is $g_j^T d_j - \beta_{k,j} \geq m_2 v_j$.

SS1 implies a serious step.
NS1 and NS2 and NS3 implies a null step.
NS1 and NS2 and not NS3 implies a line search update.
NS1 and not NS2 implies a t_j update, as before.

Everything else remains the same, and simliar convergence holds.

Bibliography

[1] Schramm, Helga and Zowe, Jochem "A verison of the bundle idea for minimizing a nonsmooth functional: conceptual idea, convergence analysis, numerical results." SIAM J. Optimization, 2 (1992), 121-152