A Nonmonotone Version of Bundle Trust Region Method with Linear Subproblems

Shuai Liu

RMIT University

Workshop on Optimization, Nonlinear Analysis, Randomness and Risk
12 July 2014

Consider the following unconstrained nonsmooth optimization (NSO) problem:

$$\min_{x\in\mathbb{R}^n} \quad f(x)$$

$f: \ \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is locally Lipschitz continuous (LLC).

Consider the following unconstrained nonsmooth optimization (NSO) problem:

$$\min_{x \in \mathbb{R}^n} \quad f(x)$$

$$f: \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \text{ is locally Lipschitz continuous (LLC).}$$

Some common assumptions in literature: quasi-differentiable, semi-smooth, prox-regular, etc.
Nonsmooth first order necessary optimality condition $0 \in \partial f(x^*)$.

$$\begin{aligned}\partial_C f(x) &:= \{s \in \mathbb{R}^n | s^T d \leq f^\circ(x; d) \text{ for all } d \in \mathbb{R}^n\} (f \text{ is LLC}) \\ &= \text{co}\{\lim_j \nabla f(y^j) | y^j \to x, \nabla f(y^j) \text{ exists and converges}\}. \quad (1)\end{aligned}$$

$$\partial_M f(x) := \{s \in \mathbb{R}^n | (s, -1) \in N((x, f(x)); \text{epi} f)\} = \text{limsup}_{y \xrightarrow{f} x} \hat{\partial} f(y) \ (f \text{ is l.s.c.})$$

$$\hat{\partial} f(y) := \{s \in \mathbb{R}^n | \liminf_{z \to y} \frac{f(z) - f(y) - s^T(z - y)}{\|z - y\|} \geq 0\}.$$

Subdifferentials are not inner semicontinuous, i.e. $\partial f(\bar{x}) \not\subseteq \liminf_{x \to \bar{x}} \partial f(x)$.

- In 1975, Philip Wolfe and Claude Lemaréchal separately developed bundle method.

- In 1975, Philip Wolfe and Claude Lemaréchal separately developed bundle method.
- The name *bundle* was born in an IIASA workshop in 1977 and it means $\{(y_i, f(y_i), s_i) | i \in I\}$ where y_i are previous iteration points, $s_i \in \partial f(y_i)$, and I is index set.

- In 1975, Philip Wolfe and Claude Lemaréchal separately developed bundle method.
- The name *bundle* was born in an IIASA workshop in 1977 and it means $\{(y_i, f(y_i), s_i) | i \in I\}$ where y_i are previous iteration points, $s_i \in \partial f(y_i)$, and I is index set.
- Based on a cutting plane model $m(x^k; d) = \max\left\{f(y_i) + s_i^T(x^k + d - y_i) | i = 1, ...l_k,\ s_i \in \partial f(y_i)\right\}$ bundle methods use search direction $d(\lambda^k) = \arg\min_{d \in \mathbf{R}^n} \left\{m(x^k; d) + \frac{1}{2\lambda^k}||d||^2\right\}$ by solving a quadratic subproblem (QP) and then perform a line search on the direction.

- In 1975, Philip Wolfe and Claude Lemaréchal separately developed bundle method.
- The name *bundle* was born in an IIASA workshop in 1977 and it means $\{(y_i, f(y_i), s_i) | i \in I\}$ where y_i are previous iteration points, $s_i \in \partial f(y_i)$, and I is index set.
- Based on a cutting plane model $m(x^k; d) = \max\left\{f(y_i) + s_i^T(x^k + d - y_i) | i = 1, ...l_k,\ s_i \in \partial f(y_i)\right\}$ bundle methods use search direction $d(\lambda^k) = \arg\min\limits_{d \in \mathbf{R}^n} \left\{m(x^k; d) + \frac{1}{2\lambda^k}||d||^2\right\}$ by solving a quadratic subproblem (QP) and then perform a line search on the direction.
- Solving large-scale QP is time consuming. Although it has some advantages: the optimal solution $d^k(\lambda^k) = \arg\min\limits_{d \in \mathbf{R}^n} \left\{m(x^k; d) + \frac{1}{2\lambda^k}||d||^2\right\}$ is unique and expressible. (But is this necessary?)

In the following reference,

Linderoth, Jeff and Wright, Stephen.
Decomposition Algorithms for Stochastic Programming on a Computational Grid.
Computational Optimization and Applications, 24(2-3):207–250, 2003.

a special bundle trust-region method for a two-stage stochastic linear programming problem, with linear subproblem (LP) was proposed. We generalized this method so that it can solve any convex unconstrained optimization problems. Replace the QP in original bundle method with the following LP.

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & m(x) = \max_{i \in I} \{ f(y_i) + \langle s_i, x - y_i \rangle \} \\ \text{subject to} \quad & ||x - \bar{x}||_\infty \le \Delta, \end{aligned} \tag{3}$$

where $\bar{x}$ is the current best candidate for a minimizer of f and Δ is the trust region radius. The optimal solution is not necessarily unique, often on a line or in the corner of the box trust region. Instead of line search we use trust region method.

Solve the linear programming subproblem and obtain an optimal solution x^{kl};

if $f(x^k) - m_l^k(x^{kl}) \leq (1 + |f(x^k)|)\epsilon_{\mathrm{tol}}$ **then**

  STOP

end

if $\rho_l^k = \frac{f(x^k) - f(x^{kl})}{f(x^k) - m_l^k(x^{kl})} \geq \eta_1$ **then**

  $x^{k+1} = x^{kl}$;

  update trust region;

  $k = k + 1$, continue to next major iteration

else

  add the cutting plane $f(x_l^k) + {s_l^k}^T(x - x_l^k)$ to the model m_{l+1}^k;

  update trust region;

  set $l = l + 1$ and continue to next minor iteration

end

Let x^{kl} be inner iteration points. The following is trust region updating procedure. Define

$$\rho_l^k = \frac{f(x^k) - f(x^{kl})}{f(x^k) - m_l^k(x^{kl})} \tag{2}$$

if $\rho_l^k > \eta_3$ **and** $\|\boldsymbol{x}^{kl} - \boldsymbol{x}^k\|_\infty = \Delta_l^k$ **then**

$\quad \Delta_1^{k+1} = \min(2\Delta_l^k, \Delta_{\max})$

else if $\rho_l^k < -\frac{1}{\min(1,\Delta_l^k)}$ **then**

$\quad \Delta_{l+1}^k = \frac{1}{\min(-\min(1,\Delta_l^k)\rho_l^k,4)}\Delta_l^k$;

$\quad$ **else**

$\quad\quad \Delta_{l+1}^k = \Delta_l^k$

$\quad$ **end**

end

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
2. prove trust region radius Δ is bounded below;

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
2. prove trust region radius Δ is bounded below;
3. prove model reduction decreases to 0;

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
2. prove trust region radius Δ is bounded below;
3. prove model reduction decreases to 0;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
2. prove trust region radius Δ is bounded below;
3. prove model reduction decreases to 0;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

1. Prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
2. prove trust region radius Δ is bounded below;
3. prove model reduction decreases to 0;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

Let f^* be the minimum value and $P(\cdot)$ is the projection onto the optimal solution set. From the subgradient inequality one can get

$$f(x^k) - f(P(x)) \leq s^T(x^k - P(x)) \leq ||s||_1 ||x^k - P(x)||_\infty, \quad \forall\, x,\ \forall\, s \in \partial f(x^k).$$

Hence $\frac{f(x^k)-f^*}{||x^k - P(x^k)||_\infty} \leq ||s||_1, \quad \forall\, s \in \partial f(x^k)$. Observe the quantity $\frac{f(x^k)-f^*}{||x^k - P(x^k)||_\infty}$ is an underestimate of the minimal norm of subgradient $||g(x^k)||_1 = \min\limits_{g \in \partial f(x^k)} ||g||_1$.

Lemma For all x^{kl} generated by algorithm 2 such that x^k is not a minimizer of f, the model reduction satisfies

$$m_l^k(x^k) - m_l^k(x^{kl}) \geq [f(x^k) - f^*] \min\left(\frac{\Delta_l^k}{||x^k - P(x^k)||_\infty}, 1\right), \tag{4}$$

Lemma For all x^{kl} generated by algorithm 2 such that x^k is not a minimizer of f, the model reduction satisfies

$$m_l^k(x^k) - m_l^k(x^{kl}) \geq [f(x^k) - f^*] \min\left(\frac{\Delta_l^k}{\|x^k - P(x^k)\|_\infty}, 1\right), \tag{4}$$

Lemma The trust region radius Δ_l^k in the previous algorithm satisfies

$$\Delta_l^k \geq \frac{1}{4} \min\left(\min_{1\leq i\leq k} \|x^i - P(x^i)\|_\infty, \min_{\substack{1\leq i\leq k \\ x^i \notin S}} \frac{f(x^i) - f^*}{L\|x^i - P(x^i)\|_\infty}\right). \tag{5}$$

Lemma For all x^{kl} generated by algorithm 2 such that x^k is not a minimizer of f, the model reduction satisfies

$$m_l^k(x^k) - m_l^k(x^{kl}) \geq [f(x^k) - f^*] \min\left(\frac{\Delta_l^k}{\|x^k - P(x^k)\|_\infty}, 1\right), \tag{4}$$

Lemma The trust region radius Δ_l^k in the previous algorithm satisfies

$$\Delta_l^k \geq \frac{1}{4} \min\left(\min_{1\leq i\leq k} \|x^i - P(x^i)\|_\infty, \min_{\substack{1\leq i\leq k \\ x^i \notin S}} \frac{f(x^i) - f^*}{L\|x^i - P(x^i)\|_\infty}\right). \tag{5}$$

Model reduction indeed decreases to 0 if there is an infinite sequence.

Lemma For all x^{kl} generated by algorithm 2 such that x^k is not a minimizer of f, the model reduction satisfies

$$m_l^k(x^k) - m_l^k(x^{kl}) \geq [f(x^k) - f^*] \min\left(\frac{\Delta_l^k}{\|x^k - P(x^k)\|_\infty}, 1\right), \tag{4}$$

Lemma The trust region radius Δ_l^k in the previous algorithm satisfies

$$\Delta_l^k \geq \frac{1}{4} \min\left(\min_{1\leq i\leq k} \|x^i - P(x^i)\|_\infty, \min_{\substack{1\leq i\leq k \\ x^i \notin S}} \frac{f(x^i) - f^*}{L\|x^i - P(x^i)\|_\infty}\right). \tag{5}$$

Model reduction indeed decreases to 0 if there is an infinite sequence.

Theorem Suppose that $\epsilon_{\mathrm{tol}} = 0$. $P(x^k)$ is the projection of x^k on the optimal solution set.

(i) If the algorithm terminates at x^{kl}, then $\boldsymbol{x}^k$ is a minimizer of f with $x^k = P(x^k)$;

(ii) if there is an infinite number of minor iterations during the kth major iteration, then $\boldsymbol{x}^k$ is a minimizer of f with $x^k = P(x^k)$ and $\lim_{l\to\infty} m_l^k(x^{kl}) - f(x^k) = 0$;

(iii) if the sequence of major iterations $\{x^k\}$ is infinite then $\lim_{k\to\infty} \|x^k - P(x^k)\|_\infty = 0$.

Nonconvex case - properties of our functions

How do we generalize this method so that it can solve nonconvex problems? It is because of convexity that the previous algorithm can keep obtaining function reduction. This motivates us to consider the para-convex functions.

How do we generalize this method so that it can solve nonconvex problems? It is because of convexity that the previous algorithm can keep obtaining function reduction. This motivates us to consider the para-convex functions.

Definition (para-convexity)

Given a point $\bar{x} \in \mathbb{R}^n$ and a real number $\epsilon > 0$, a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *para-convex* on $B(\bar{x}, \epsilon)$ with respect to a if there exists $a \geq 0$ such that the function $f(\cdot) + \frac{a}{2}||\cdot||^2$ is convex on $B(\bar{x}, \epsilon)$.

How do we generalize this method so that it can solve nonconvex problems? It is because of convexity that the previous algorithm can keep obtaining function reduction. This motivates us to consider the para-convex functions.

Definition (para-convexity)

Given a point $\bar{x} \in \mathbb{R}^n$ and a real number $\epsilon > 0$, a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *para-convex* on $B(\bar{x}, \epsilon)$ with respect to a if there exists $a \geq 0$ such that the function $f(\cdot) + \frac{a}{2}||\cdot||^2$ is convex on $B(\bar{x}, \epsilon)$.

Definition (prox-regularity)

A function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is *prox-regular* at $\bar{x}$ for $\bar{v}$ with respect to ϵ and a if f is finite and locally lower semicontinuous at $\bar{x}$ with $\bar{v} \in \partial f(\bar{x})$, and there exist $\epsilon > 0$ and $a \geq 0$ such that

$$f(x') \geq f(x) + \langle v, x' - x \rangle - \frac{a}{2}||x' - x||^2 \; \forall \; x' \in B(\bar{x}, \epsilon) \tag{6}$$

when $||x - \bar{x}|| < \epsilon$, $v \in \partial f(x)$, $||v - \bar{v}|| < \epsilon$, $f(x) < f(\bar{x}) + \epsilon$. When this holds for all $\bar{v} \in \partial f(\bar{x})$, f is said to be prox-regular at $\bar{x}$.

Assumption 1 The objective function f is locally Lipschitz continuous and bounded below; given $x^0 \in \mathbb{R}^n$ f is prox-regular on bounded level set $\mathrm{lev}_{x^0} f$.

Assumption 1 The objective function f is locally Lipschitz continuous and bounded below; given $x^0 \in \mathbb{R}^n$ f is prox-regular on bounded level set $\mathrm{lev}_{x^0} f$.

Proposition (Eberhard 2001) If a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is locally Lipschitz continuous and prox-regular at $\bar{x}$ then there exist ϵ and a such that f is para-convex on $B(\bar{x}, \epsilon)$ with respect to a.

Assumption 1 The objective function f is locally Lipschitz continuous and bounded below; given $x^0 \in \mathbb{R}^n$ f is prox-regular on bounded level set $\mathrm{lev}_{x^0} f$.
Proposition (Eberhard 2001) If a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is locally Lipschitz continuous and prox-regular at $\bar{x}$ then there exist ϵ and a such that f is para-convex on $B(\bar{x}, \epsilon)$ with respect to a.
The proximal point mapping

$$x \mapsto P_a(x) := \arg\min_y \left\{ f(y) + \frac{a}{2} \|y - x\|^2 \right\} \tag{7}$$

satisfies if $x = P_a(x)$ then $0 \in \partial f(x)$ for suitable a.

Assumption 1 The objective function f is locally Lipschitz continuous and bounded below; given $x^0 \in \mathbb{R}^n$ f is prox-regular on bounded level set $\mathrm{lev}_{x^0} f$.
Proposition (Eberhard 2001) If a function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is locally Lipschitz continuous and prox-regular at $\bar{x}$ then there exist ϵ and a such that f is para-convex on $B(\bar{x}, \epsilon)$ with respect to a.
The proximal point mapping

$$x \mapsto P_a(x) := \arg\min_y \left\{ f(y) + \frac{a}{2} \|y - x\|^2 \right\} \tag{7}$$

satisfies if $x = P_a(x)$ then $0 \in \partial f(x)$ for suitable a. Our goal is to find the global minimizer of

$$g(y) := g(y; x, a) : y \mapsto f(y) + \frac{a}{2} ||y - x||^2 \tag{8}$$

i.e. minimize a sequence of functions $\{g(y; x^n, a^n)\}$ such that $\lim_{n\to\infty} \|\boldsymbol{x}^n - P_{a^n}(\boldsymbol{x}^n)\|_\infty = 0$.

Consider

$$g(y) := g(y; x, a) : y \mapsto f(y) + \frac{a}{2}||y - x||^2 \tag{9}$$

with x and a as parameters.

- If f is para-convex, then according to Definition 1, $g(y)$ is convex on some neighborhood $B_b(x)$.
- Clearly, for different x and b, to make $g(y)$ convex with respect to y, there exists a threshold for the value of a.
- Suppose we have some sequences $x^k \to x'$, $a^k \to a'$, and $b^k \to b'$ such that $g(y; x^k, a^k)$ is convex with respect to y on $B_{b^k}(x^k)$ for all k. Then we can use a cutting-planes model for $g(y; x^k, a^k)$ with box trust region to generate descent.

In order to find the global minimizer of $g(y; x, a)$ we want to show that it is convexifiable at least on $\mathrm{lev}_{x^0} f$.

Theorem

Suppose f is prox-regular and locally Lipschitz on a bounded level set $lev_{x^0} f$ with int $lev_{x^0} f \neq \emptyset$. Let $g\,(y; x, a)$ be defined in (9) with $a \geq 0$. There exists an a^{th} such that $g(y; x, a)$ is the restriction to $lev_{x^0} f$ of a convex function $H(y; x, a)$ for $a \geq a^{th}$ and for any $x \in \mathbb{R}^n$.

Then we can define $P(x, g(y; x^k, a^k))$ as the projection of $x \in B_{b^k}(x^k)$ onto the optimal solution set of $g(y; x^k, a^k)$ over the neighbourhood $B_{b^k}(x^k)$. If we can show $||x^k - P(x^k, g(y; x^k, a^k))||_\infty \to 0$ when $k \to \infty$, then we get $x' = P(x', g(y; x', a'))$. That is, x' is an optimal solution of $g(y; x', a')$ over $B_{b'}(x')$.

On-the-fly convexification

In 2010 Hare and Sagastizábal proposed redistributed bundle method with an on-the-fly convexification technique (still with QP and line search).

In 2010 Hare and Sagastizábal proposed redistributed bundle method with an on-the-fly convexification technique (still with QP and line search). What is the necessary condition for a function to be convex? If f is convex then

$$f(x) = \sup\left\{f(y) + s^T(x - y) | y \in \mathbb{R}^n,\ s \in \partial f(y)\right\}.$$

In 2010 Hare and Sagastizábal proposed redistributed bundle method with an on-the-fly convexification technique (still with QP and line search). What is the necessary condition for a function to be convex? If f is convex then

$$f(x) = \sup\left\{ f(y) + s^T(x-y) | y \in \mathbb{R}^n,\ s \in \partial f(y) \right\}.$$

The linearization error of any cutting plane at $\bar{x}$

$$\tilde{e}_i := f(\bar{x}) - [f(y_i) + \langle s_i, \bar{x} - y_i \rangle]. \tag{10}$$

should be positive unless $\bar{x}$ is minimizer. We are minimizing $g(x; \bar{x}, a)$ so the linearization error of its cutting plane at $\bar{x}$

$$E_i := f(\bar{x}) - [f(y_i) + \frac{a}{2}\|y_i - \bar{x}\| + \langle s_i + a(y_i - \bar{x}), \bar{x} - y_i \rangle] \tag{11}$$

should be positive.

In 2010 Hare and Sagastizábal proposed redistributed bundle method with an on-the-fly convexification technique (still with QP and line search). What is the necessary condition for a function to be convex? If f is convex then

$$f(x) = \sup\left\{f(y) + s^T(x-y) | y \in \mathbb{R}^n,\ s \in \partial f(y)\right\}.$$

The linearization error of any cutting plane at $\bar{x}$

$$\tilde{e}_i := f(\bar{x}) - [f(y_i) + \langle s_i, \bar{x} - y_i \rangle]. \tag{10}$$

should be positive unless $\bar{x}$ is minimizer. We are minimizing $g(x; \bar{x}, a)$ so the linearization error of its cutting plane at $\bar{x}$

$$E_i := f(\bar{x}) - [f(y_i) + \frac{a}{2}\|y_i - \bar{x}\| + \langle s_i + a(y_i - \bar{x}), \bar{x} - y_i \rangle] \tag{11}$$

should be positive. This can be done by manually manipulating a when some linearization errors at such points are negative.

For example, to make sure all the current cutting planes are below the graph of g at $\bar{x}$, we can set $E_i \geq 0$ for all $i \in I$. Suppose $\bar{x}$, $y_i,\ s_i$ for $i \in I$ are given and a is not fixed. Then we can deduce an inequality of a:

$$a \geq \max_{i \in I} \left\{ \frac{f(y_i) - f(\bar{x}) + \langle s_i, \bar{x} - y_i \rangle}{\frac{1}{2}||y_i - \bar{x}||^2} \right\} =: a^{\min}. \tag{12}$$

Note that $a^{\min}$ is dependent on the prox-center $\bar{x}$ and the cutting-plane index set I.

For example, to make sure all the current cutting planes are below the graph of g at $\bar{x}$, we can set $E_i \geq 0$ for all $i \in I$. Suppose $\bar{x}$, y_i, s_i for $i \in I$ are given and a is not fixed. Then we can deduce an inequality of a:

$$a \geq \max_{i \in I} \left\{ \frac{f(y_i) - f(\bar{x}) + \langle s_i, \bar{x} - y_i \rangle}{\frac{1}{2}||y_i - \bar{x}||^2} \right\} =: a^{\min}. \tag{12}$$

Note that $a^{\min}$ is dependent on the prox-center $\bar{x}$ and the cutting-plane index set I. We keep adding new cutting planes and updating $a^{\min}$. We make sure the used a in the LP is not less than $a^{\min}$.

We solve the linear subproblem

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & m(x;\bar{x},a,I) := \max\{\textit{cutting planes of } g(x;\bar{x},a)\} \\ \text{subject to} \quad & \|x-\bar{x}\|_\infty \leq \Delta \end{aligned} \tag{13}$$

which is equivalent to the following problem

$$\min_{(x,z)\in\mathbb{R}^{n+1}} \quad z \tag{14a}$$

$$\text{subject to} \quad f(y_i) + \frac{a}{2}\|y_i-\bar{x}\|^2 + \langle s_i + a(y_i-\bar{x}), x-y_i\rangle \leq z, \;\; i \in I, \tag{14b}$$

$$\|x-\bar{x}\|_\infty \leq \Delta. \tag{14c}$$

We relax the rule for accepting major iteration points to allow $f(x^{k+1}) > f(x^k)$. Compute the largest index $m(k)$ such that

$$f(x^{m(k)}) = \max_{i=\max\{k-p,0\},\cdots,k} f(x^i). \tag{15}$$

$$\hat{\rho}_l^k = \frac{f(x^{m(k)}) - f(x^{kl})}{\sum\limits_{i=m(k)}^{k} f(x^i) - z^{i,l_i}}, \ \rho_l^k = \frac{f(x^k) - f(x^{kl})}{f(x^k) - z^{kl}}, \tag{16}$$

$$\bar{\rho}_l^k = \max\{\rho_l^k, \hat{\rho}_l^k\}. \tag{17}$$

We replace the ρ_l^k in previous version with $\bar{\rho}_l^k$ and then follow the same trust region update scheme as before.

Solve the LP (14) and obtain an optimal solution (x^{kl}, z^{kl});
if $f(x^k) - z^{kl} \leq (1 + |f(x^k)|)\epsilon_{\mathrm{tol}}$ **then**
  STOP; $\boldsymbol{x}^k$ is an approximate stationary point
end
if $\bar{\rho}_l^k \geq \eta_1$ **then**
  declare a serious step; $x^{k+1} \leftarrow x^{kl}$;
  update trust region radius (same as in convex case);
  check E_i and update a^{min};
  $k = k + 1$, continue to next major iteration
else
  update trust region radius;
  check E_i and update a^{min};
  set x^{kl} as a new auxiliary point y_i and add the cutting-plane function to the model m_{l+1}^k;
  set $l = l + 1$ and continue to next minor iteration;
end

1. Prove model reduction decreases to 0;

1. Prove model reduction decreases to 0;
2. prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);

1. Prove model reduction decreases to 0;
2. prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
3. prove trust region radius Δ is bounded below;

1. Prove model reduction decreases to 0;
2. prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
3. prove trust region radius Δ is bounded below;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

1. Prove model reduction decreases to 0;
2. prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
3. prove trust region radius Δ is bounded below;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

1. Prove model reduction decreases to 0;
2. prove model reduction $\geq A \cdot B$, where $A \approx$ the subgradient with minimum norm, $B \approx$ trust region radius, ($A \cdot B \approx$ directional derivative);
3. prove trust region radius Δ is bounded below;
4. find a contradiction with A decreases to 0 but x^* is not an optimal solution.

The orders are changed; 2 requires convexity and 3 requires 2.

Lemma The model reduction of LP Bundle Nonconvex converges to 0. Specifically, (i) if in iteration k there is an infinite sequence of minor iterations then $\lim_{l \to \infty} [m_l^k(x^k) - m_l^k(x^{kl})] = 0$; (ii) if the sequence of major iteration points $\{x^k\}$ is infinite then

$$\lim_{k \to \infty} [m_{l_k}^k(x^k) - m_{l_k}^k(x^{k+1})] = 0. \tag{18}$$

where l_k is the last minor iteration in k-th major iteration so that $x^{k+1} = x^{kl_k}$. Note this lemma should be slightly changed in nonmonotone version.

A major iterate occurs when we get sufficient descent at the l_k minor iteration and then we place $x^{k+1} = x^{kl_k}$. If don't distinguishes these we can place $n \equiv \{k, l\}$

Lemma

If the objective function $f(x)$ satisfies Assumption 1, then there exists $\bar{n} > 1$ such that the sequence $\{a^n\}$ stabilizes:

$$a^n = a^{\bar{n}}, \ \forall \ n \geq \bar{n}.$$

Assumption 2 There exists a certain iteration $\tilde{n} = (\bar{k}, \bar{l})$, where the actual value of parameter a used in subproblem (14) is updated such that $a^{\tilde{n}} \geq a^{th}$ and the algorithm does not stop before reaching such iteration $\tilde{n}$.

A major iterate occurs when we get sufficient descent at the l_k minor iteration and then we place $x^{k+1} = x^{kl_k}$. If don't distinguishes these we can place $n \equiv \{k, l\}$

Lemma

If the objective function $f(x)$ satisfies Assumption 1, then there exists $\bar{n} > 1$ such that the sequence $\{a^n\}$ stabilizes:

$$a^n = a^{\bar{n}},\ \forall\ n \geq \bar{n}.$$

Assumption 2 There exists a certain iteration $\tilde{n} = (\bar{k}, \bar{l})$, where the actual value of parameter a used in subproblem (14) is updated such that $a^{\tilde{n}} \geq a^{th}$ and the algorithm does not stop before reaching such iteration $\tilde{n}$.

This means $g(x; x^k, a)$ is the restriction of a convex function $H(x; x^k, a)$ to the whole level set $\text{lev}_{x^0} f$ for $a \geq a^{th}$ and for any $x^k \in \mathbb{R}^n$.

Lemma Suppose Assumption 2 holds true. Then for all (k, l) after $(\bar{k}, \bar{l})$ we have $a_l^k \equiv a^{\tilde{n}}$ and $P_{a_l^k}(\boldsymbol{x}^k)$ is well defined and single valued; further more, if $x^k \neq P_{a_l^k}(\boldsymbol{x}^k)$, then the model reduction satisfies

$$m_l^k(\boldsymbol{x}^k) - m_l^k(x^{kl}) \geq \left[f(x^k) - g\left(P_{a_l^k}(\boldsymbol{x}^k); x^k, a_l^k \right) \right] \min \left(\frac{\Delta_l^k}{\|x^k - P_{a_l^k}(\boldsymbol{x}^k)\|_\infty}, 1 \right).$$

Lemma Suppose Assumption 2 holds true. Then for all (k, l) after $(\bar{k}, \bar{l})$, Δ_l^k satisfies

$$\Delta_l^k \geq \alpha_1 \min\left\{ \min_{\bar{k} \leq i \leq k} \|x^i - P_{a^{\tilde{n}}}(\boldsymbol{x}^i)\|_\infty, \min_{\substack{\bar{k} \leq i \leq k \\ x^i \neq P_{a^{\tilde{n}}}(\boldsymbol{x}^i)}} \frac{f(x^i) - g\left(P_{a^{\tilde{n}}}(\boldsymbol{x}^i); x^i, a^{\tilde{n}} \right)}{\bar{L}\|x^i - P_{a^{\tilde{n}}}(\boldsymbol{x}^i)\|_\infty} \right\}, \text{ or } \Delta_l^k \geq \Delta_{\bar{l}}^{\bar{k}}.$$

Theorem Let Assumption 2 hold true and $\epsilon_{\text{tol}} = 0$. Let iteration $(\bar{k}, \bar{l})$ correspond to $\tilde{n}$ in sequences and (k, l) be not before $(\bar{k}, \bar{l})$.
(i) If the algorithm terminates at x^{kl}, then $x^k = P_{a^{\tilde{n}}}(\boldsymbol{x}^k)$;
(ii) if there is an infinite number of minor iterations after the kth major iteration, then $x^k = P_{a^{\tilde{n}}}(\boldsymbol{x}^k)$;
(iii) if the sequence of major iterations $\{x^k\}$ is infinite, then $\lim_{k \geq \bar{k},\ k \to \infty} ||x^k - P_{a^{\tilde{n}}}(\boldsymbol{x}^k)||_\infty = 0$.

Comparison of performance on standard testing problems with the splitting bundle method appeared in 2013 shows that LP bundle method is comparable with QP bundle method.

No.	Problem	f_opt	f*(LPB)	fn(LPB)	f*(ncvx)	fn(ncvx)
2	Crescent	0	5.14E-08	29	7.55E-06	27
3	CB2	1.9522245	1.952225451	17	1.9522246	20
4	CB3	2	2.000000000	4	2.0000000	19
5	DEM	-3	-3.000000000	9	-3.000000000	10
6	QL	7.2	7.200000668	18	7.2000001	21
7	LQ	-1.4142136	-1.41421274	19	-1.4142136	8
8	Mifflin1	-1	-0.999999683	29	-0.9999839	99
9	Mifflin2	-1	-1.000000000	21	-0.9999999	10

No.	Problem	f_opt	f*(LPB)	fn(LPB)	f*(ncvx)	fn(ncvx)
10	Wolfe	-8	-8.000000000	6	-7.9999999	26
11	Rosen	-44	-43.99998585	55	-44.000000000	36
12	Shor	22.60016	22.60018019	56	22.600212	48
13	Colville 1	-32.3487	-32.34867809	43	–32.348678	51
14	HS78	-2.9197	-2.918896332	191		
15	El-Attar	0.5598131	0.55981318	95	0.55981572	287
16	Maxquad	-0.84141	-0.841407525	223	-0.8414077	61
17	Gill	9.7857721	9.785976382	185	9.7860516	253
18	Steiner 2	16.703838	16.70384921	107	16.703839	173
19	Maxq	0	1.94176E-07	342	2.77E-05	324
20	Maxl	0	0.000000000	216	3.40E-07	22
22	Goffin	0	0.000000000	52	6.06E-08	57
23	MXHILB	0	1.53885E-06	18	2.936E -07	13
24	L1HILB	0	0.000000000	3	8.47E-07	33

Discussion

- We generalized a special bundle trust region method so that it can solve generic convex problems.
- Under convexification we extended the method by minimizing a sequence of locally convex functions and showed that the iteration sequence converges to a fixed point of the proximal point mapping.
- We think this fixed point is already a local minimizer of the objective function, instead of a stationary point, if the convexification is successful. But we are yet to prove it.
- When the parameter a is too big the function $g(x; x^k, a)$ is too steep and the real reduction of f can be very small; in the next stage of this project, we would like to allow a to be decreased in the numerical experiments.

- We generalized a special bundle trust region method so that it can solve generic convex problems.
- Under convexification we extended the method by minimizing a sequence of locally convex functions and showed that the iteration sequence converges to a fixed point of the proximal point mapping.
- We think this fixed point is already a local minimizer of the objective function, instead of a stationary point, if the convexification is successful. But we are yet to prove it.
- When the parameter a is too big the function $g(x; x^k, a)$ is too steep and the real reduction of f can be very small; in the next stage of this project, we would like to allow a to be decreased in the numerical experiments.

Thank you!