Trust Region Methods

Jubayer Ibn Hamid

1 Introduction

These are notes on trust region optimization methods that I took as I read [1]. Intuitively, line search methods first make a quadratic model of the function f to generate a step direction and calculate the step length (preferably one that satisfies the Wolfe Conditions). In contrast, trust region methods also generate a model of the function f - but they define a region around where we are currently such that inside that region we believe our model is more or less the same as function f and then step to the minimizer of the model. Therefore, we would want our model to be one that we *can* in fact solve.

2 Model

Let us first discuss how to model the function f in the first place. This is simply done through Taylor expansion. Suppose, we are currently at x_k and we choose to take a step in direction p. Then, if we expand f, we get:

$$f(x_k + p) = f_k + g_k^T p + \frac{1}{2} p^T \nabla^2 f(x_k + tp)p$$

where $t \in (0, 1)$, $f_k := f(x_k)$ and $g_k := \nabla f(x_k)$. Now, suppose we approximate the matrix $\nabla^2 f(x_k)$ with B_k, such that B_k is a symmetric matrix. Then, we model the function to be:

$$m_k(p) := f_k + g_k^T p + \frac{1}{2} p^T B_k p.$$

Therefore, at each step, we define the trust region to be a sphere of radius Δ_k around x_k. Therefore, from x_k, we will take a step p_k such that $||p_k|| \leq \Delta_k$ and it solves the model m_k. So, we are searching for the solution:

$$\arg\min_{p \in \mathbb{R}^n} m_k(p)$$

To emphasize again, note that within the trust region, we are solving the *model*, not the function f. Given a step p_k, we may *think* that we have reduced the function f by a certain amount, but may end up reducing the function by a different amount. To quantify this, we define:

$$\rho_k := \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}.$$

Note that the numerator is the the amount by which f got reduced after we took the step p_k from x_k. The denominator is the amount by which we reduced the *model*. This ratio therefore is also a measure of how good our model is - if it is close to 1, it means our model represented the function well and therefore the step p_k reduced f by the same amount by which it reduced m_k.

Generally, the way trust region methods work is that we start at x_k, fix a radius Δ_k to define the trust region, make a model m_k of the function inside this trust region and then take a step p_k that reduces the model m_k and **stays inside the trust region**. Then, we calculate ρ_k to get an estimation of how good our model is. If it is close to 1, we expand our trust region radius (because our model is good we can take a step that's longer). On the other hand, if ρ_k is close to 0, we realise that our model is not too good and so we choose to not take the step p_k, we make the trust region smaller to try and find a more accurate model and then try again.

Trust Region Algorithm Outline:

Hyperparameters: $\Delta_{max} > 0, \Delta_0 \in (0, \Delta_{max}), \eta \in [0, \frac{1}{4})$:

For k=0,1, 2,...:

  Solve m_k to find p_k

  Calculate ρ_k

  If $\rho_k < \frac{1}{4}$:

    $\Delta_k := \frac{1}{4}\Delta_k$

  Else:

    If $\rho_k > \frac{3}{4}$ and $||p_k|| = \Delta_k$:

      $\Delta_{k+1} := \min(2\Delta_k, \Delta_{max})$

    Else:

      $\Delta_{k+1} := \Delta_k$

  if $\rho_k > \eta$:

    $x_{k+1} := x_k + p_k$

  Else:

    $x_{k+1} := x_k$

3 How to solve the model m_k

3.1 Solving model using Cauchy Point:

We define the solutions of two problems:

$$p_k^s := \arg\min_{p \in \mathbb{R}^n} f_k + g_k^T p \quad \text{s.t} \; ||p|| \leq \Delta_k$$

$$\tau_k^s := \arg\min_{\tau \geq 0} m_k(\tau p_k^s) = f_k + g_k^T(\tau p_k^s) + \frac{1}{2}\tau^2 (p_k^s)^T B_k p_k^s \quad \text{s.t} \; ||p|| \leq \Delta_k$$

Then, we define:

$$p_k^C := \tau_k p_k^s.$$

Now, the solution to the first problem is simply the same as what we saw in line search methods:

$$p_k^s = -\frac{\Delta_k}{|g_k||} g_k.$$

Solving the second problem is slightly trickier.

First, consider the case where $g_k^T B_k g_k \leq 0$: Then, $m_k(\tau p_k^s)$ decreases monotonically with τ whenever $g_k \neq 0$. So, τ_k is just the largest value such that we are inside the trust region, so $\tau_k = 1$.

Next, consider the case where $g_k^T B_k g_k > 0$: Then, $m_k(\tau p_k^s)$ is convex quadratic in τ. So, either τ_k is going to be the value such that $\tau_k p_k^s$ minimizes m_k which is $\frac{||g_k||^3}{\Delta_k g_k^T B_k g_k}$ (differentiate $m_k(\tau p_k^s)$ with respect to τ and use the definition of p_k^s to get this) or it is going to be 1.

Putting these together, we get the following solution:

$$p_k^C := -\tau_k \frac{\Delta_k}{||g_k||} g_k$$

where

$$\tau_k := \begin{cases} 1 & \text{if } \; g_k^T B_k g_k \leq 0 \\ \min(\frac{||g_k||^3}{\Delta_k g_k^T B_k g_k}, 1) & \text{otherwise} \end{cases} \tag{1}$$

While this method seems concrete, we can improve on it. For starter, notice that our step direction p_k^s is not affected by B_k at all, meaning that the second order terms are not used to generate the step direction at all. It only affects the step length τ_k. We could try and use information from B_k to determine the step direction too.

3.2 Solving model using Dogleg Method:

Note 1: We use this method for only when B_k is positive definite. There are some fairly simple and nice ways to *make* B_k positive definite without compromising on accuracy much.

Note 2: We will drop the k in subscripts for easier notation.

When B is positive definite, from our study of line search methods, we know that the solution is just $p^B := -B^{-1}g$. So when this is a step that keeps us inside the trust region, we will choose this as our solution. Therefore,

$$p^*(\Delta) = p^B = -B^{-1}g$$

when $\Delta \geq \left| |p^B| \right|$. We wrote p^* as a function of the trust region radius to emphasize the fact that our step will depend on the radius. Of course, the other case is where $\Delta < \left| |p^B| \right|$. In this case, the quadratic term does not contribute much to our model, so we only keep the linear terms to get

$$p^*(\Delta) \approx -\Delta \frac{g}{||g||}.$$

Dogleg method essentially gives us a solution that interlaces these two solutions. The solution is:

$$\tilde{p}(\tau) := \begin{cases} \tau p^U, & 0 \leq \tau \leq 1 \\ p^U + (\tau - 1)(p^B - p^U), & 1 \leq \tau \leq 1 \end{cases} \tag{2}$$

where $p^U := -\frac{g^T g}{g^T B g} g$ and $p^B := -B^{-1} g$.

Now, we will prove a theorem that shows that the path we described parametrised by τ does not overlap and that along this path we do minimise our model:

Theorem 1. *Let B be positive definite. Then,*
(i) $||\tilde{p}(\tau)||$ is an increasing function of τ and
(ii) $m(\tilde{p}(\tau))$ is a decreasing function of τ.

Proof. It is very easy to see that the theorem is true for τ. We focus on $\tau \in [1, 2]$. We prove (i) first. Define $S(\alpha) := \frac{1}{2} ||\tilde{p}(1 + \alpha)||^2$. Expanding using (2), we get $S(\alpha) = \frac{1}{2} \left| |p^U + \alpha(p^B - p^U)| \right|^2 = \frac{1}{2} \left| |p^U| \right|^2 + \alpha (p^U)^T (p^B - p^U) + \frac{1}{2}\alpha^2 \left| |p^B - p^U| \right|^2$. Then, for $\alpha \in (0, 1)$,

$$\begin{aligned} S'(\alpha) &= -(p^U)^T(p^U - p^B) + \alpha \left| |p^U - p^B| \right|^2 \\ &\geq -(p^U)^T(p^U - p^B) \\ &= \frac{g^T g}{g^T B g} g^T \left(-\frac{g^T g}{g^T B g} g + B^{-1} g\right) \\ &= g^T g \frac{g^T B^{-1} g}{g^T B g} \left(1 - \frac{(g^T g)^2}{(g^T B g)(g^T B^{-1} g)}\right) \end{aligned}$$

Now, we show that $\frac{(g^Tg)^2}{(g^TBg)(g^TB^{-1}g)} \geq 1$:

$$\begin{aligned}
\frac{(g^Tg)^2}{(g^TBg)(g^TB^{-1}g)} &= \frac{(g^Tg)^2}{(g^TBB^Tg)(B^Tg)^{-1}g(g^TB^{-1}(B^{-1})^Tg)((B^{-1})^Tg)^{-1}g} \\
&= \frac{(g^Tg)^2}{(B^Tg \cdot B^Tg)(B^Tg)^{-1}g((B^{-1})^Tg \cdot (B^{-1})^Tg)g^{-1}B^Tg} \\
&\geq \frac{(g^Tg)^2}{|B^Tg \cdot (B^{-1})^Tg|^2} \quad \text{Using Cauchy-Schwarz Inequality} \\
&= \frac{(g^Tg)^2}{|g^TB^T(B^{-1})^Tg|^2} \\
&= \frac{(g^Tg)^2}{(g^Tg)^2} \\
&= 1.
\end{aligned}$$

Using this, we get $S'(\alpha) \geq 0$ proving (i).

Now, we prove (ii) (for $\tau \in [1, 2]$):

Define $\hat{S}(\alpha) := m(\tilde{p}(1 + \alpha))$. We will show $\hat{S}'(\alpha) \leq 0$ for $\alpha \in (0, 1)$.

$$\begin{aligned}
\hat{S}(\alpha) &= (p^B - p^U)^T(g + Bp^U) + \alpha(p^B - p^U)^TB(p^B - p^U) \\
&\leq (p^B - p^U)^T(g + Bp^U + B(p^B - p^U)) \\
&= (p^B - p^U)^T(g + Bp^B) \\
&= 0.
\end{aligned}$$

□

4 Convergence Properties

First, we prove the following lemma:

Lemma 2. *The Cauchy point p_k^C satisfies*

$$m_k(0) - m_k(p_k) \geq \frac{1}{2}\,||g_k||\min(\Delta, \frac{||g_k||}{||B_k||})$$

where we are using the Frobenius norm of matrices.

Proof. We drop the k's in subscripts for easier notation.

Case 1: $g^T Bg \leq 0$
In this case, we have $\tau = 1$, so

$$\begin{aligned} m(p^C) - m(0) &= m(-\frac{\Delta}{||g||}g) - f \\ &= -\frac{\Delta}{||g||}\,||g||^2 + \frac{1}{2}\frac{\Delta^2}{||g||^2}g^T Bg \\ &\leq -\Delta\,||g|| \\ &\leq -\,||g||\min(\Delta, \frac{||g||}{||B||}) \end{aligned}$$

Case 2: $g^T Bg > 0$ and $\frac{||g||^3}{\Delta g^T Bg} \leq 1$
Then,

$$\tau := \frac{||g||^3}{\Delta g^T Bg}.$$

So,

$$\begin{aligned} m(p^C) - m(0) &= -\frac{||g||^4}{g^T Bg} + \frac{1}{2}g^T Bg\frac{||g||^4}{(g^T Bg)^2} \\ &= -\frac{1}{2}\frac{||g||^4}{g^T Bg} \\ &\leq -\frac{1}{2}\frac{||g||^4}{||B||\,||g||^2} \\ &= -\frac{1}{2}\frac{||g||^2}{||B||} \\ &\leq -\frac{1}{2}\,||g||\min(\Delta, \frac{||g||}{||B||}) \end{aligned}$$

Case 3: $g^T Bg > 0$ and $\frac{||g||^3}{\Delta g^T Bg} > 1$

Then, $\tau = 1$. So,

$$\begin{aligned} m(p^C) - m(0) &= -\frac{\Delta}{||g||}\,||g||^2 + \frac{1}{2}\frac{\Delta^2}{||g||^2} g^T B g \\ &\leq -\Delta\,||g|| + \frac{1}{2}\frac{\Delta^2}{||g||^2}\frac{||g||^3}{\Delta} \\ &= -\frac{1}{2}\Delta\,||g|| \\ &\leq -\frac{1}{2}\,||g||\min(\Delta, \frac{||g||}{||B||} \end{aligned}$$

□

Using this, we can now prove the following theorem which shows that we are minimising our model m:

Theorem 3. *Let p_k be any vector such that $||p_k|| \leq \Delta_k$ and $m_k(0) - m_k(p_k) \geq c_2(m_k(0) - m_k(p_k^C))$. Then, p_k satisfies:*

$$m_k(0) - m_k(p_k) \geq \frac{c_2}{2}\,||g||\min(\Delta_k, \frac{||g_k||}{||B_k||}).$$

In particular, if p_k is the exact solution $p_k^* = \arg\min_{p \in \mathbb{R}^n} m_k(p) := f_k + g_k^T p + \frac{1}{2} p^T B_k p$ (where $||p|| \leq \Delta_k$). Then, it satisfies:

$$m_k(0) - m_k(p_k) \geq \frac{1}{2}\,||g||\min(\Delta_k, \frac{||g_k||}{||B_k||}).$$

Proof. Given $||p_k|| \leq \Delta_k$, we use Lemma 2 to write:

$$m_k(0) - m_k(p_k) \geq c_2(m_k(0) - m_k(p_k^C)) \geq \frac{1}{2} c_2\,||g||\min(\Delta_k, \frac{||g_k||}{||B_k||})$$

.

Note that when $p_k = p_k^*$, the step is the Cauchy point and therefore, the first inequality becomes equality with $c_2 = 1$. From that we get the second part of the theorem. □

Next, we prove two theorems to show these methods end up converging to stationary points - which ultimately is the real goal of optimisation.

Theorem 4. *Let $\eta = 0$ in our algorithm (the pseudocode is above). Suppose, $||B_k|| \leq \beta$ for some constant β. Suppose, f is bounded on the level set $S := \{x : f(x) \leq f(x_0)\}$.*

Now, suppose f is Lipschit continuously differentiable in the neighbourhood $S(R_0) := \{x : ||x - y|| \leq R_0\}$ (for some $y \in S$) with Lipschitz constant β_1.

Finally, suppose all approximate solutions of $\min_{p\in\mathbb{R}^n} m_k(p_k) = f_k + g_k^T p + \frac{1}{2} p_k^T B_k p_k$ satisfies:

(a) $m_k(0) - m_k(p_k) \geq c_1 ||g_k|| \min(\Delta_k, \frac{||g_k||}{||B_k||})$ for some $c_1 \in (0, 1]$.

(b) $||p_k|| \leq \gamma\Delta_k$ for some $\gamma \geq 1$.

Then,

$$\liminf_{k\longrightarrow\infty} ||g_k|| = 0$$

Proof. We first bound $|\rho_k - 1|$:

$$\begin{aligned} |\rho_k - 1| &= \left| \frac{(f(x_k) - f(x_k + p_k)) - (m_k(0) - m_k(p_k))}{m_k(0) - m_k(p_k)} \right| \\ &= \left| \frac{-f(x_k + p_k) + m_k(p_k)}{m_k(0) - m_k(p_k)} \right|. \end{aligned}$$

Now, we use mean value theorem (with some algebraic manipulation - adding a zero term) to write:

$$f(x_k + p_k) = f(x_k) + g(x_k)^T p_k + \int_0^1 [g(x_k + tp_k) - g(x_k)]^T p_k dt.$$

Then, we use the definition of m_k to write:

$$|m_k(p_k) - f(x_k + p_k)| = \left| \frac{1}{2} p_k^T B_k p_k - \int_0^1 [g(x_k + tp_k) - g(x_k)]^T p_k dt \right| \quad (3)$$

$$\leq \frac{\beta}{2} ||p_k||^2 + \beta_2 ||p_k||^2 \quad (4)$$

where we got the last inequality using the Lipschitz continuity condition: $||g(x_k + tp_k) - g(x_k)|| \leq \beta_1 ||x_k + tp_k - x_k||$ and we assumed $||p_k|| \leq R_0$ to ensure both x_k and $x_k + tp_k$ are inside $S(R_0)$.

Now, we suppose, for contradiction:

Claim A: There exists $\epsilon > 0$ and $Z > 0$ such that $||g_k|| \geq \epsilon$ for any $k \geq Z$. Then, for any $k \geq Z$, we have

$$m_k(0) - m_k(p_k) \geq c_1 ||g_k|| \min(\Delta, \frac{||g_k||}{||B_k||}) \geq c_1\epsilon \min(\Delta, \frac{\epsilon}{\beta}). \quad (5)$$

Now, we use (4) and (5) to write:

$$|\rho_k - 1| \le \frac{\gamma^2 \Delta_k^2 (\frac{\beta}{2} + \beta_1)}{c_1 \epsilon \min(\Delta_k, \frac{\epsilon}{\beta})}.$$

Now, we define $\bar{\Delta} := \min(\frac{1}{2}\frac{c_1\epsilon}{\gamma^2(\frac{\beta}{2}+\beta_1)}, \frac{R_0}{\gamma})$. We consider all $\Delta_k \le \bar{\Delta}$. Note that we added R_0/γ inside the min function to ensure that $||p_k|| \le \gamma\Delta_k$ since $\gamma\Delta_k \le \gamma\bar{\Delta} \le R_0$.

With this, since $c_1 \le 1$ and $\gamma \ge 1$, we have $\bar{\Delta} \le \frac{\epsilon}{\beta}$. Thus, for any $\Delta_k \in [0, \bar{\Delta}]$, we have $\min(\Delta_k, \frac{\epsilon}{\beta}) = \Delta_k$.

$$\begin{aligned} |\rho_k - 1| &\le \frac{\gamma^2 \Delta_k^2 (\frac{\beta}{2} + \beta_1)}{c_1 \epsilon \Delta_k} \\ &= \frac{\gamma^2 \Delta_k (\frac{\beta}{2} + \beta_1)}{c_1 \epsilon} \\ &\le \frac{\gamma^2 \bar{\Delta} (\frac{\beta}{2} + \beta_1)}{c_1 \epsilon} \\ &\le \frac{1}{2} \end{aligned}$$

where we got the last inequality using the definition of $\bar{\Delta}$.

Using this, we know $\rho_k > \frac{1}{4}$ and so, in our algorithm, $\Delta_{k+1} \ge \Delta_k$ whenever $\Delta_k \le \bar{\Delta}$. So, $\Delta_{k+1} = \frac{1}{4}\Delta_k$ only if $\Delta \ge \bar{\Delta}$. Together. we have

$$\Delta_k \ge \min(\Delta_k, \bar{\Delta}/4) \tag{6}$$

for any $k \ge Z$. Now, for the sake of contradiction: *Claim B:* suppose there exists an infinite subsequence ϕ such that $\rho_k \ge \frac{1}{4}$ for any $k \in \phi$.

Then, for any $k \in \phi$ and $k \ge Z$, we have from (5):

$$f(x_k) - f(x_{k+1}) = f(x_k) - f(x_k + p_k) \ge \frac{1}{4}[m_k(0) - m_k(p_k)] \ge \frac{1}{4} c_1 \epsilon \min(\Delta_k, \frac{\epsilon}{\beta})$$

where we got the first inequality using $\rho_k \ge \frac{1}{4}$. Now, given f is bounded below, therefore, $\lim_{k \in \phi, k \to \infty} \Delta_k = 0$. This contradicts (6) meaning our assumption, *claim B*, was wrong. Therefore, $\Delta_{k+1} = \frac{1}{4}\Delta_k$ at every iteration, so $\lim_{k \to \infty} \Delta_k = 0$, which contradicts (6) again, meaning our assumption, *claim A*, was wrong. Given $||g_k||$ is bounded below by 0, this implies the theorem. □

Finally, we prove an even stronger theorem showing convergence to stationary points:

Theorem 5. *Let $\eta \in (0, 1/4)$ in our algorithm (the pseudocode is above). Suppose, $||B_k|| \leq \beta$ for some constant β. Suppose, f is bounded on the level set $S := \{x : f(x) \leq f(x_0)\}$.*

Now, suppose f is Lipschit continuously differentiable in the neighbourhood $S(R_0) := \{x : ||x - y|\,| \leq R_0\}$ (for some $y \in S$).

Finally, suppose all approximate solutions of $\min_{p \in \mathbb{R}^n} m_k(p_k) = f_k + g_k^T p + \frac{1}{2} p_k^T B_k p_k$ satisfies:

(a) $m_k(0) - m_k(p_k) \geq c_1 \, ||g_k|| \min(\Delta_k, \frac{||g_k||}{||B_k||})$ for some $c_1 \in (0, 1]$.

(b) $||p_k|| \leq \gamma \Delta_k$ for some $\gamma \geq 1$.

Then,

$$\lim_{k \longrightarrow \infty} g_k = 0$$

.

Proof. Consider some $m > 0$ such that $g_m \neq 0$. By Lipschitz continuity, we have $||g(x) - g_m|| \leq \beta_1 ||x - x_m||$ for any $x \in S(R_0)$.

Now, define $\epsilon := \frac{1}{2} ||g_m||$, $R := \min(\frac{\epsilon}{\beta_1}, R_0)$. Now, the open ball centered at x_m - $B_R(x_m) := \{x : ||x - x_m|| \leq R\}$ is contained in $S(R_0)$, meaning the Lipschitz contintuity of g holds inside this open ball.

Then, $x \in B_R(x_m)$ implies

$$\begin{aligned} ||g_m|| &\geq ||g_m|| - ||g(x) - g_m|| \\ &\geq ||g_m|| - \beta_1 R \\ &\geq ||g_m|| - \epsilon \\ &= ||g_m|| - \frac{||g_m||}{2} \\ &= \frac{1}{2} ||g_m|| \\ &= \epsilon. \end{aligned}$$

Now, if the entire sequence $\{x_k\}_{k \geq m}$ is inside $B_R(x_m)$, then $||g_m|| \geq \epsilon > 0$ for any $k \geq m$. Now, in the same way we contradicted *claim A* in theorem 4, we can show that this never occurs. So, $\{x_k\}_{k \geq m}$ will have to get outside the open ball $B_R(x_m)$ for *some* k. Let $l \geq m$ be the smallest index such that x_{l+1} is outside $B_R(x_m)$. Now, for $k \in [m, l]$, we have $||g_k|| \geq \epsilon$. Thus,

$$m_k(0) - m_k(p_k) \geq c_1 \, ||g_k|| \min(\Delta_k, \frac{||g_k||}{||B_k||}) \geq c_1 \epsilon \min(\Delta_k, \frac{\epsilon}{\beta}).$$

Thus,

$$\begin{aligned} f(x_m) - f(x_{l+1}) &= \sum_{k=m}^{l} f(x_k) - f(x_{k+1} \\ &\geq \sum_{k=m}^{l} \eta(m_k(0) - m_k(p_k)) \quad \text{Using definition of } \rho_k \\ &\geq \sum_{k=m}^{l} \eta c_1 \epsilon \min(\Delta_k, \frac{\epsilon}{\beta}) \end{aligned}$$

Now, if $\Delta_k \leq \frac{\epsilon}{\beta}$ for $k \in [m, l]$, we have

$$\begin{aligned} f(x_m) - f(x_{l+1}) &\geq \eta c_1 \epsilon \sum_{k=m}^{l} \Delta_k \\ &\geq \eta c_1 \epsilon R \quad \text{Since we are summing over } \Delta_k \text{ and } x_{l+1} \text{ is outside } B_R(x_m) \\ &= \eta c_1 \epsilon \min(\frac{\epsilon}{\beta_1}, R_0) \end{aligned}$$

So,

$$f(x_m) - f(x_{l+1}) \geq \eta c_1 \epsilon \min(\frac{\epsilon}{\beta_1}, R_0). \tag{7}$$

On the other hand, if $\Delta_k > \frac{\epsilon}{\beta}$ for some $k \in [m, l]$, we have

$$f(x_m) - f(x_{l+1}) \geq \eta c_1 \epsilon \frac{\epsilon}{\beta}. \tag{8}$$

Since the sequence $\{f(x_k)\}_{k=0}^{\infty}$ is decreasing and bounded below, therefore $\lim_{k\to\infty} f(x_k) = f^*$ for some $f^* > -\infty$.

now, we use (7) and (8) to write:

$$\begin{aligned} f(x_m) - f^* &\geq f(x_m) - f(x_{l+1}) \\ &\geq \eta c_1 \epsilon \min(\frac{\epsilon}{\beta}, \frac{\epsilon}{\beta_1}, R_0) \\ &= \frac{1}{2}\eta c_1 \, ||g_m|| \min\left(\frac{||g_m||}{2\beta}, \frac{||g_m||}{2\beta_1}, R_0\right) \\ &> 0. \end{aligned}$$

Now, since $\lim_{k\to\infty} f(x_k) - f^* = 0$, we must have $\lim_{k\to\infty} ||g_k|| = 0$ □

5 References

[1] Jorge Nocedal, Stephen J. Wright. *Numerical Optimization.* Spring Series in Operations Research