Lecture 8: CS395T Numerical Optimization for Graphics and AI — Trust Region Methods II

Qixing Huang
The University of Texas at Austin
huangqx@cs.utexas.edu

1 Disclaimer

This note is adapted from

- Section 4 of *Numerical Optimization* by Jorge Nocedal and Stephen J. Wright. Springer series in operations research and financial engineering. Springer, New York, NY, 2. ed. edition, (2006)

2 Introduction

In this lecture, we will study the global convergence of trust region methods. We are particularly interested in the following algorithm:

1. **Input**: Given $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \in [0, \frac{1}{4})$:
2. **for** $k = 0, 1, 2, \ldots$
3. Obtain $\boldsymbol{p}_k$ by approximately solving the subproblem:

$$\begin{aligned} \boldsymbol{p}_k := \underset{\boldsymbol{p}}{\operatorname{argmin}} \quad & m_k(\boldsymbol{p}) := f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T B_k \boldsymbol{p} \\ \text{subject to} \quad & \|\boldsymbol{p}\| \leq \Delta_k. \end{aligned} \tag{1}$$

4. Evaluate $\rho_k = \frac{f(\boldsymbol{x}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)}{m_k(0) - m_k(\boldsymbol{p}_k)}$;
5. **if** $\rho_k < \frac{1}{4}$
6. $\Delta_{k+1} = \frac{1}{4} \Delta_k$
7. **else**
8. **if** $\rho_k > \frac{3}{4}$ and $\|\boldsymbol{p}_k\| = \Delta_k$
9. $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta})$;
10. **else**
11. $\Delta_{k+1} = \Delta_k$;
12. **if** $\rho_k \geq \eta$
13. $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{p}_k$

14. **else**

15. $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$;

16. **end(for)**.

Last lecture, we have talked about how to solve the sub-problem (1) exactly. In this lecture, we study the global convergence of this algorithm under different strategies for solving (1).

2.1 Algorithms Based on the Cauchy Point

We have discussed line search methods can be globally convergent even when the optimal step length is not used at each iteration. In fact, the step length α_k only need to satisfy fairly loose criteria. A similar situation applies in trust-region methods. Although in principle we seek the optimal solution of the subproblem, it is enough for purposes of global convergence to find an approximate solution $\boldsymbol{p}_k$ that lies within the trust region and gives a sufficient reduction in the model. The sufficient reduction can be quantified in terms of the Cauchy point, which we denote by $\boldsymbol{p}_k^C$ and define in terms of the following simple procedure.

Cauchy point calculation. The Cauchy point is calculated by following a two step procedure. The first step determines the search direction by solving the following optimization problem:

$$\begin{aligned} \boldsymbol{p}_k := \underset{\boldsymbol{p}}{\text{argmin}} \quad & f_k + \boldsymbol{g}_k^T \boldsymbol{p} \\ \text{subject to} \quad & \|\boldsymbol{p}\| \leq \Delta_k. \end{aligned} \tag{2}$$

Given the search direction, we then optimize the best step-size τ_k by solving the reduced trust-region problem by involving B_k:

$$\begin{aligned} \tau_k := \underset{\tau}{\text{argmin}} \quad & f_k + \boldsymbol{g}_k^T (\boldsymbol{p}_k \tau) + \frac{1}{2} (\boldsymbol{p}_k \tau)^T B_k (\boldsymbol{p}_k \tau) \\ \text{subject to} \quad & \|\boldsymbol{p}_k \tau\| \leq \Delta_k. \end{aligned} \tag{3}$$

It is easy to see that

$$\boldsymbol{p}_k := -\Delta_k \frac{\boldsymbol{g}_k}{\|\boldsymbol{g}_k\|},$$

and

$$\tau_k := \begin{cases} 1 & \boldsymbol{g}_k^T B_k \boldsymbol{g}_k \leq 0 \\ \min(1, \frac{\|\boldsymbol{g}_k\|^3}{\Delta_k \boldsymbol{g}_k^T B_k \boldsymbol{g}_k}) & \boldsymbol{g}_k^T B_k \boldsymbol{g}_k > 0 \end{cases}$$

The Cauchy step $\boldsymbol{p}_k^C = \tau_k \boldsymbol{p}_k$ is inexpensive to calculate—no matrix factorizations are required—and is of crucial importance in deciding if an approximate solution of the trust-region sub-problem is acceptable. As we will see later, a trust-region method will be globally convergent if its steps $\boldsymbol{p}_k$ give a reduction in the model m_k that is at least some fixed positive multiple of the decrease attained by the Cauchy step.

Cauchy point method can be considered as a specialized version of steepest decent, which may converge poorly. The major issue is that the second order term B_k is not involved in determining the search direction. Below we study a few enhanced versions of the Cauchy point method which utilize the second order information B_k.

Dogleg method. This method is used in the case B_k is positive definite. To motivate this method, we start by examining the effect of the trust-region radius Δ on the solution $p^\star(\Delta)$ of the sub-problem. When B_k is positive definite, we have already noted that the unconstrained minimizer of m_k is $\boldsymbol{p}_k^B = -B_k^{-1} \boldsymbol{g}_k$. When this point is feasible, it is obviously a solution, so we have

$$\boldsymbol{p}_k^\star(\Delta_k) = \boldsymbol{p}^B, \qquad \text{when } \Delta_k \geq \|\boldsymbol{p}_k^B\|.$$

When Δ_k is small relative to $\boldsymbol{p}_k^B$, the restriction ensures that the quadratic term in m_k has little effect on the solution of the sub-problem. For such Δ_k, we can get an approximation to $p(\Delta_k)$ by simply omitting the quadratic term in the sub-problem and writing

$$\boldsymbol{p}_k^\star(\Delta_k) \approx -\Delta_k \frac{\boldsymbol{g}_k}{\|\boldsymbol{g}_k\|}, \qquad \text{when } \Delta_k \text{ is small.}$$

For intermediate Δ_k, $\boldsymbol{p}_k^\star(\Delta_k)$ follows a curved trajectory that interpolates $\boldsymbol{x}^k$ and $\boldsymbol{p}_k^B$. The curved trajectory is also tangent to $\boldsymbol{g}_k$.

The dogleg method approximates this trajectory by a polygonal curve with three vertices $\boldsymbol{x}_k$, $\boldsymbol{p}_k^U$ and $\boldsymbol{p}_k^B$. Here $\boldsymbol{p}_k^U$ is given by the optimal solution along the search direction (requires a bit derivation):

$$p_k^U = -\frac{\|\boldsymbol{g}_k\|^2}{\boldsymbol{g}_k^T B_k \boldsymbol{g}_k} \boldsymbol{g}_k.$$

Formally we denote this trajectory as

$$\hat{\boldsymbol{p}}_k(\tau_k) = \begin{cases} \tau_k \boldsymbol{p}_k^U, & 0 \le \tau_k \le 1, \\ \boldsymbol{p}_k^U + (\tau_k - 1)(\boldsymbol{p}_k^B - \boldsymbol{p}_k^U), & 1 \le \tau_k \le 2. \end{cases} \tag{4}$$

The dogleg method chooses $\boldsymbol{p}_k$ to minimize the model m_k along this path, subject to the trust-region bound. The following lemma shows that the minimum along the dogleg path can be found easily.

Lemma 2.1. *Let B_k be positive definite. Then*

- *$\|\hat{\boldsymbol{p}}_k(\tau_k)\|$ is an increasing function of τ_k, and*
- *$m_k(\hat{\boldsymbol{p}}_k(\tau_k))$ is a decreasing function of τ_k.*

The proof is straight-forward, we will work this in class. The Lemma also gives a way to calculate the optimal τ_k:

$$\tau_k = \begin{cases} \frac{\Delta_k}{\|\boldsymbol{p}_k^U\|} & \Delta_k \le \|\boldsymbol{p}_k^U\| \\ \|\boldsymbol{p}_k^U + (\tau_k - 1)(\boldsymbol{p}_k^B - \boldsymbol{p}_k^U)\| = \Delta_k & \|\boldsymbol{p}_k^U\| \le \Delta_k \le \|\boldsymbol{p}_k^B\| \\ 2 & \Delta_k \ge \|\boldsymbol{p}_k^B\|. \end{cases} \tag{5}$$

Two-dimensional Sub-space Minimization

$$\begin{aligned} \underset{\boldsymbol{p}}{\text{minimize}} \quad & m_k(\boldsymbol{p}) := f_k + \boldsymbol{g}_k^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T B_k \boldsymbol{p} \\ \text{subject to} \quad & \|\boldsymbol{p}\| \le \Delta_k, \ \boldsymbol{p} \in \text{span}[\boldsymbol{g}_k, B_k^{-1} \boldsymbol{g}_k]. \end{aligned} \tag{6}$$

When B_k is indefinite, we can replace $B_k^{-1}\boldsymbol{g}_k$ by $(B_k + \alpha I)^{-1}\boldsymbol{g}_k$, where $\alpha \in (-\lambda_n(B_k), -2\lambda_n(B_k))$.

3 Global Convergence

The main argument we will develop is that the dogleg and two-dimensional subspace minimization algorithms produce approximate solutions $\boldsymbol{p}_k$ of the sub-problem that satisfy the following estimate of decrease in the model function:

$$m_k(0) - m_k(\boldsymbol{p}_k) \ge c_1 \|\boldsymbol{g}_k\| \min\Big(\Delta_k, \frac{\|\boldsymbol{g}_k\|}{\|B_k\|}\Big), \tag{7}$$

for some constant $c_1 \in (0, 1]$. The usefulness of this estimate will become clear in the following two sections. For now, we note that when Δ_k is the minimum value in (7), the condition is slightly reminiscent of the first Wolfe condition: The desired reduction in the model is proportional to the gradient and the size of the step. We show now that the Cauchy point $\boldsymbol{p}_k^C$ satisfies (7), with $c_1 = \frac{1}{2}$.

Lemma 3.1. *The Cauchy point $\boldsymbol{p}_k^C$ satisfies (7) with $c_1 = \frac{1}{2}$, that is,*

$$m_k(0) - m_k(\boldsymbol{p}_k^C) \geq \frac{1}{2}\|\boldsymbol{g}_k\| \min(\Delta_k, \frac{\|\boldsymbol{g}_k\|}{\|B_k\|}). \tag{8}$$

To satisfy (8), our approximate solution $\boldsymbol{p}_k$ has only to achieve a reduction that is at least some fixed fraction c_2 of the reduction achieved by the Cauchy point. We state the observation formally as a theorem.

Theorem 3.1. *Let $\boldsymbol{p}_k$ be any vector such that $\|\boldsymbol{p}_k\| \leq \Delta_k$ and $m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k) \geq c_2(m_k(\mathbf{0}) - m_k(\boldsymbol{p}_k^C))$. Then $\boldsymbol{p}_k$ satisfies (8) with $c_1 = \frac{c_2}{2}$. In particular, if $\boldsymbol{p}_k$ is the exact solution $\boldsymbol{p}_k^\star$ of the sub-problem, then it satisfies (8) with $c_1 = \frac{1}{2}$.*

Note that the dogleg and two-dimensional subspace minimization algorithms both satisfy (8) with $c_1 = \frac{1}{2}$, because they all produce approximate solutions $\boldsymbol{p}_k$ for which $m_k(\boldsymbol{p}_k) \leq m_k(\boldsymbol{p}_k^C)$.

Convergence to Stationary Points. We make a few assumptions regarding the objective function f:

- f is bounded below on the level set

$$S := \{\boldsymbol{x} | f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)\}.$$

- We also consider an open neighborhood of this set by

$$S(R_0) := \{\boldsymbol{x} | \|\boldsymbol{x} - \boldsymbol{y}\| < R_0 \text{ for some } \boldsymbol{y} \in S\}.$$

- We also allow the length of the approximate solution $\boldsymbol{p}_k$ of the sub-problem to exceed the trust-region bound, provided that it stays within some fixed multiple of the bound; that is, for some constant $\gamma \geq 1$,

$$\|\boldsymbol{p}_k\| \leq \gamma \Delta_k. \tag{9}$$

The first result deals with the case $\gamma = 0$.

Theorem 3.2. *Let $\gamma = 0$. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is bounded below on the level set S and Lipschitz continuously differentiable in the neighborhood $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions $\boldsymbol{p}_k$ of the sub-problem satisfy the inequalities (8) and (9) for some positive constants c_1 and γ. We then have*

$$\liminf \|\boldsymbol{g}_k\| = 0.$$

Sketch proof: First of all, we can obtain

$$|\rho_k - 1| = |\frac{m_k(\boldsymbol{p}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)}{m_k(0) - m_k(\boldsymbol{p}_k)}|.$$

Using the bound on B_k and the Lipschitz continuity condition, we have

$$|m_k(\boldsymbol{p}_k) - f(\boldsymbol{x}_k + \boldsymbol{p}_k)| \leq (\frac{\beta}{2})\|\boldsymbol{p}_k\|^2 + \beta_1\|\boldsymbol{p}_k\|^2.$$

Show that the following argument leads to a contradiction:

$$\|\boldsymbol{g}_k\| \geq \epsilon, \qquad \text{for all } k \geq K.$$

A similar analysis leads to

Theorem 3.3. *Let $\gamma \in (0, \frac{1}{4})$. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is bounded below on the level set S and Lipschitz continuously differentiable in $S(R_0)$ for some $R_0 > 0$, and that all approximate solutions $\boldsymbol{p}_k$ of the sub-problem satisfy the inequalities (8) and (9) for some positive constants c_1 and γ. We then have*

$$\lim g_k = 0.$$