Lecture 7: CS395T Numerical Optimization for Graphics and AI — Trust Region Methods

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

Line search methods and trust-region methods both generate steps with the help of a quadratic model of the objective function, but they use this model in different ways. Line search methods use it to generate a search direction, and then focus their efforts on finding a suitable step length along this direction. Trust-region methods define a region around the current iterate within which they trust the model to be an adequate representation of the objective function, and then choose the step to be the approximate minimizer of the model in this region. In effect, they choose the direction and length of the step simultaneously. If a step is not acceptable, they reduce the size of the region and find a new minimizer. In general, the direction of the step changes whenever the size of the trust region is altered.

Remark 2.1. *Trust region methods have proven to be very effective on various applications. However, it seems to be less used compared to line search methods, partly because it is more complicated to understand and implement. Nevertheless, below is a list of recent papers:*

- *Trust Region Policy Optimization. John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan and Philipp Moritz. Proceedings of the 32nd International Conference on Machine Learning (ICML-15). Pages: 1889-1897*
- *Fast Trust Region for Segmentation. Lena Gorelick, Frank R. Schmidt, Yuri Boykov; The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1714-1721*
- *Discriminative Clustering for Image Co-segmentation Armand Joulin, Francis Bach and Jean Ponce. CVPR, 2010.*

The size of the trust region is critical to the effectiveness of each step. If the region is too small, the algorithm misses an opportunity to take a substantial step that will move it much closer to the minimizer of the objective function. If too large, the minimizer of the model may be far from the minimizer of the objective function in the region, so we may have to reduce the size of the region and try again. In practical algorithms, we choose the size of the region according to the performance of the algorithm during previous iterations. If the model is consistently reliable, producing good steps and accurately predicting the behavior of the objective function along these steps, the size of the trust region may be increased to allow longer, more

ambitious, steps to be taken. A failed step is an indication that our model is an inadequate representation of the objective function over the current trust region. After such a step, we reduce the size of the region and try again. The difference between trust region methods and Newton methods is that in many instances the second order approximation is accurate within a certain region, while the optimal solution to the Newton method may be out of this region.

In this section, we will assume that the model function m_k that is used at each iterate x_k is quadratic. Moreover, m_k is based on the Taylor-series expansion of f around x_k, which is

$$f(\boldsymbol{x}_k + \boldsymbol{p}) = f(\boldsymbol{x}_k) + (\nabla f(\boldsymbol{x}_k))^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T \nabla^2 f(\boldsymbol{x}_k + t\boldsymbol{p})\boldsymbol{p},$$

where t is some scalar in the interval $(0, 1)$. By using an approximation B_k to the Hessian in the second-order term, m_k is defined as follows:

$$m_k(\boldsymbol{p}) = f(\boldsymbol{x}_k) + (\nabla f(\boldsymbol{x}_k))^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p},$$

where B_k is some symmetric matrix. For example, For example, the choice $B_k = \nabla^2 f(\boldsymbol{x}_k)$ leads to the trust-region Newton method. In this lecture, we will carry out the discussion by assuming B_k is general except that it is symmetric and has uniform boundedness.

More precisely, we will consider the following sub-optimization problem:

$$\min_{\boldsymbol{p}\in\mathbb{R}^n} \; m_k(\boldsymbol{p}) = f(\boldsymbol{x}_k) + (\nabla f(\boldsymbol{x}_k))^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B_k \boldsymbol{p} \quad s.t. \|\boldsymbol{p}\| \leq \Delta_k, \tag{1}$$

where $\Delta_k > 0$ is the trust-region radius. As we will see later, solving (1) can be down without too much computational cost. Moreover, in most cases, we only need an approximate solution to obtain convergence and good practical behavior.

Outline Of The Trust-Region Approach

One of the key ingredients in a trust-region algorithm is the strategy for choosing the trust-region radius Δ_k at each iteration. We base this choice on the agreement between the model function m_k and the objective function f at previous iterations. Given a step p_k we define the ratio

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

The numerator is called the actual reduction, and the denominator is the predicted reduction (that is, the reduction in f predicted by the model function). Note that since the step $\boldsymbol{p}_k$ is obtained by minimizing the model m_k over a region that includes $\boldsymbol{p} \geq 0$, the predicted reduction will always be nonnegative. Hence, if ρ_k is negative, the new objective value $f(\boldsymbol{x}_k + \boldsymbol{p}_k)$ is greater than the current value $f(\boldsymbol{x}_k)$, so the step must be rejected. On the other hand, if ρ_k is close to 1, there is good agreement between the model m_k and the function f over this step, so it is safe to expand the trust region for the next iteration. If ρ_k is positive but significantly smaller than 1, we do not alter the trust region, but if it is close to zero or negative, we shrink the trust region by reducing Δ_k at the next iteration.

Exercise 1. Convert the paragraph above into a formal algorithm.

3 Main Theorem Regarding the Sub-program

Trust region amounts to solve the following optimization problem at each iteration:

$$\min_{\boldsymbol{p}} \; f + \boldsymbol{g}^T \boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B \boldsymbol{p} \quad s.t. \quad \|\boldsymbol{p}\| \leq \Delta. \tag{2}$$

The following theorem characterizes the optimal solution to this

Theorem 3.1. *Vector $\boldsymbol{p}^\star$ is the global solution of (2) if and only if $\boldsymbol{p}^\star$ is feasible and there is a scalar $\lambda \geq 0$ such that the following conditions are satisfied:*

$$\begin{aligned} (B + \lambda I)\boldsymbol{p}^\star &= -\boldsymbol{g}, \\ \lambda(\Delta - \|\boldsymbol{p}^\star\|) &= 0, \\ (B + \lambda I) \quad &\text{is positive semidefinite} \end{aligned} \tag{3}$$

Proof Sketch. The first step is to make simplify the problem by converting into the following simplied form:

$$\begin{aligned} \min_{y_1,\cdots,y_n} \quad & \sum_{i=1}^{n} \frac{1}{2}\lambda_i (y_i - o_i)^2 \\ \text{subject to} \quad & \sum_{i=1}^{n} y_i^2 \leq \Delta^2. \end{aligned} \tag{4}$$

Without losing generality, we can assume $o_i \geq 0$, since otherwise we have flip the sign of the corresponding y_i. So the global solution either happens to be $y_i = o_i, 1 \leq i \leq n$, which means $\lambda_i > 0, 1 \leq i \leq n$ and $\sum_{i=1}^{n} o_i^2 \leq \Delta^2$, or along the bound of $\|\boldsymbol{p}\| \leq \Delta$. If it is on the boundary, it must be the case that the isolines of $\sum_i y_i^2 = \Delta^2$ and $\sum_i \lambda_i (p_i - o_i)^2$ must have the same tangent line, and the gradient is in opposite directions. In other words, there exists a positive number λ so that

$$-\lambda y_i = \lambda_i (y_i - o_i), \quad 1 \leq i \leq n,$$

which means

$$y_i = \frac{\lambda_i}{\lambda + \lambda_i} o_i, \quad 1 \leq i \leq n.$$

Now consider the objective function

$$f(\boldsymbol{y}) = \sum_{i=1}^{n} \lambda y_i^2 + \sum_{i=1}^{n} \lambda_i (y_i - o_i)^2.$$

For any other $\overline{\boldsymbol{y}}$ that $\Delta^2 = \overline{\boldsymbol{y}}^T \overline{\boldsymbol{y}} = \boldsymbol{y}^T \boldsymbol{y}$, we have

$$f(\overline{\boldsymbol{y}}) \geq f(\boldsymbol{y}).$$

This means

$$\begin{aligned} f(\overline{\boldsymbol{y}}) - f(\boldsymbol{y}) &= \sum_{i=1}^{n} \lambda_i \big((\overline{y}_i - o_i)^2 - (y_i - o_i)^2 \big) \\ &= \sum_{i=1}^{n} (\lambda + \lambda_i)(\overline{y}_i - y_i)^2 \geq 0. \end{aligned}$$

So $B + \lambda I$ must be SDP. □

Formal Proof. The proof relies on the following technical lemma, which deals with the unconstrained minimizers of quadratics and is particularly interesting in the case where the Hessian is positive semidefinite.

Lemma 3.1. *Let m be the quadratic function defined by*

$$m(\boldsymbol{p}) = \boldsymbol{g}^T \boldsymbol{p} + \frac{1}{2} \boldsymbol{p}^T B \boldsymbol{p},$$

where B is any symmetric matrix. Then the following statements are true.

- *m attains a minimum if and only if B is positive semi-definite and is in the range of B. If B is positive semi-definite, then every $\boldsymbol{p}$ satisfying $B\boldsymbol{p} = -\boldsymbol{g}$ is a global minimizer of m.*
- *m has a unique minimizer if and only if B is positive definite.*

Proof. We prove each of the three claims in turn.

- We start by proving the "if" part. Since $\boldsymbol{g}$ is in the range of B, there is a $\boldsymbol{p}$ with $B\boldsymbol{p} = -\boldsymbol{g}$. For all $\boldsymbol{w} \in \mathbb{R}^n$, we have

$$\begin{aligned} m(\boldsymbol{p}+\boldsymbol{w}) &= \boldsymbol{g}^T(\boldsymbol{p}+\boldsymbol{w}) + \frac{1}{1}(\boldsymbol{p}+\boldsymbol{w})^T B(\boldsymbol{p}+\boldsymbol{w}) \\ &= (\boldsymbol{g}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T B\boldsymbol{p}) + \boldsymbol{g}^T\boldsymbol{w} + (B\boldsymbol{p})^T\boldsymbol{w} + \frac{1}{2}\boldsymbol{w}^T B\boldsymbol{w} \\ &= m(\boldsymbol{p}) + \frac{1}{2}\boldsymbol{w}^T B\boldsymbol{w} \\ &\geq m(\boldsymbol{p}), \end{aligned}$$

  since B is positive semidefinite. Hence, $\boldsymbol{p}$ is a minimizer of m. For the "only if" part, let $\boldsymbol{p}$ be a minimizer of m. Since $\nabla m(\boldsymbol{p}) = B\boldsymbol{p} + \boldsymbol{g} = 0$, we have that $\boldsymbol{g}$ is in the range of B. Also, we have $\nabla^2 m(\boldsymbol{p}) = B$ positive semidefinite, giving the result.

- For the "if" part, the same argument above suffices with the additional point that $\boldsymbol{w}^T B\boldsymbol{w} > 0$ whenever $\boldsymbol{w} \neq 0$. For the "only if" part, we proceed as in (i) to deduce that B is positive semidefinite. If B is not positive definite, there is a vector $\boldsymbol{w} \neq 0$ such that $B\boldsymbol{w} = 0$. Hence, we have $m(\boldsymbol{p}+\boldsymbol{w}) = m(\boldsymbol{p})$, so the minimizer is not unique, giving a contradiction.

□

Now let us go back to the prove. Assume first that there is $\lambda \geq 0$ such that the conditions in the theorem are satisfied. The lemma above implies that $\boldsymbol{p}^\star$ is a global minimum of the quadratic function

$$\hat{m}(\boldsymbol{p}) = \boldsymbol{g}^T\boldsymbol{p} + \frac{1}{2}\boldsymbol{p}^T(B+\lambda I)\boldsymbol{p} = m(\boldsymbol{p}) + \frac{\lambda}{2}\boldsymbol{p}^T\boldsymbol{p}.$$

Since $\hat{m}(\boldsymbol{p}) \geq \hat{m}(\boldsymbol{p}^\star)$, we have

$$m(\boldsymbol{p}) \geq m(\boldsymbol{p}^\star) + \frac{\lambda}{2}((\boldsymbol{p}^\star)^T\boldsymbol{p}^\star - \boldsymbol{p}^T\boldsymbol{p}). \tag{5}$$

Because $\lambda(\Delta - \|p^\star\|) = 0$ and therefore $\lambda(\Delta^2 - (\boldsymbol{p}^\star)^T\boldsymbol{p}^\star) = 0$, we have

$$m(\boldsymbol{p}) \geq m(\boldsymbol{p}^\star) + \frac{\lambda}{2}(\Delta^2 - \boldsymbol{p}^T\boldsymbol{p}).$$

Hence, from $\lambda \geq 0$, we have $m(\boldsymbol{p}) \geq m(\boldsymbol{p}^\star)$ for all $\boldsymbol{p}$ with $\|\boldsymbol{p}\| \leq \Delta$. Therefore, $\boldsymbol{p}^\star$ is a global minimizer of the sub-problem.

For the converse, we assume that $\boldsymbol{p}^\star$ is a global solution of (2) and show that there is a $\lambda \geq 0$ that satisfies (3). In the case $\|\boldsymbol{p}^\star\| < \Delta$, $\boldsymbol{p}^\star$ is an unconstrained minimizer of m, and so

$$\nabla m(\boldsymbol{p}^\star) = B\boldsymbol{p}^\star + \boldsymbol{g} = 0, \quad \nabla^2 m(\boldsymbol{p}^\star) = B\text{positive semidefinite},$$

and so the properties (3) hold for $\lambda \geq 0$.

Assume for the remainder of the proof that $\|\boldsymbol{p}^\star\| = \Delta$. Then (3)(b) is immediately satisfied, and $\boldsymbol{p}^\star$ also solves the constrained problem

$$\min m(\boldsymbol{p}) \quad \text{subject to} \|\boldsymbol{p}\| = \Delta.$$

By applying optimality conditions for constrained optimization to this problem (a simpler version is what we have discussed above), we find that there is a λ such that the Lagrangian function defined by

$$L(\boldsymbol{p}, \lambda) := m(\boldsymbol{p}) + \frac{\lambda}{2}(\boldsymbol{p}^T\boldsymbol{p} - \Delta^2)$$

has a stationary point at $\boldsymbol{p}^\star$. By setting $\nabla_{\boldsymbol{p}} L(\boldsymbol{p}^\star, \lambda) = 0$, we obtain

$$B\boldsymbol{p}^\star + \boldsymbol{g} + \lambda \boldsymbol{p}^\star = 0 \rightarrow (B + \lambda I)\boldsymbol{p}^\star = -\boldsymbol{g},$$

so that (3)(a) holds. Since $\hat{m}(\boldsymbol{p}) \geq \hat{m}(\boldsymbol{p}^\star)$ for any $\boldsymbol{p}$ with $\boldsymbol{p}^T\boldsymbol{p} = (\boldsymbol{p}^\star)^T\boldsymbol{p}^\star = \Delta^2$, we have for such vectors $\boldsymbol{p}$ that

$$m(\boldsymbol{p}) \geq m(\boldsymbol{p}^\star) + \frac{\lambda}{2}((\boldsymbol{p}^\star)^T\boldsymbol{p}^\star - \boldsymbol{p}^T\boldsymbol{p}).$$

If we substitute the expression for $\boldsymbol{p}^\star = (B+\lambda I)^{-1}\boldsymbol{g}$ into this expression, we obtain after some rearrangement that

$$\frac{1}{2}(\boldsymbol{p} - \boldsymbol{p}^\star)^T(B + \lambda I)(\boldsymbol{p} - \boldsymbol{p}^\star) \geq 0.$$

Since the set of directions $\boldsymbol{p} - \boldsymbol{p}^\star$ is dense, so $B + \lambda I$ must be positive semidefinite.

It remains to show that $\lambda \geq 0$. Because (3)(a) and (3)(c) are satisfied by $\boldsymbol{p}^\star$, we have from the Lemma that $\boldsymbol{p}^\star$ minimizes $L(\boldsymbol{p}, \lambda)$, so (5) holds. Suppose that there are only negative values of λ that satisfy (3)(a) and (3)(c). Then we have from (5) that $m(\boldsymbol{p}) \geq m(\boldsymbol{p}^\star)$ whenever $\|\boldsymbol{p}\| \geq \|\boldsymbol{p}^\star\|$. Since we already know that $\boldsymbol{p}^\star$ minimizes m for $\|\boldsymbol{p}\| \leq \Delta$, it follows that m is in fact a global, unconstrained minimizer of m. From Lemma it follows that $B\boldsymbol{p} = -\boldsymbol{g}$ and B is positive semidefinite. Therefore conditions 3(a) and 3(c) are satisfied by $\lambda \geq 0$, which contradicts our assumption that only negative values of $\lambda \geq 0$ can satisfy the conditions. We conclude that $\lambda \geq 0$, completing the proof.

□

This theorem suggests that the sub-problem can be reformulated as defining

$$p(\lambda) = -(B + \lambda I)^{-1}\boldsymbol{g}$$

for λ sufficiently large that $B + \lambda I$ is positive definite and seek a value $\lambda > 0$ such that

$$\|p(\lambda)\| = \Delta.$$

This problem is one-dimensional root-finding problem in the variable λ.

It is easy to formulate (i.e., by using the eigenvector decomposition of B)

$$\|p(\lambda)\|^2 = \sum_{j=1}^{n} \frac{(\boldsymbol{u}_i^T\boldsymbol{g})^2}{(\lambda_j + \lambda)^2},$$

where $\boldsymbol{u}_i$ denotes the eigenvectors of B. Let λ_n be the smallest eigenvector of B, then it is clear that $\|p(\lambda)\|$ is a continuous, non-increasing function of λ on the interval $(-\lambda_n, \infty)$. In fact, we have that

$$\lim_{\lambda \rightarrow +\infty} \|p(\lambda)\| = 0.$$

When $\boldsymbol{u}_n^T\boldsymbol{g} \neq 0$, the root $\|p(\lambda)\| - \Delta = 0$ can be found using Newton's method. A numerically more stable strategy is find the root for $\frac{1}{\Delta} - \frac{1}{\|p(\lambda)\|} = 0$.

When $\boldsymbol{u}_n^T\boldsymbol{g} = 0$, it can be shown that the optimal $\lambda = -\lambda_n$.