PART I. OPTIMIZATION: CLASSICAL APPROACHES
(LECTURE 3)

Shpilev Petr Valerievich

Faculty of Mathematics and Mechanics, SPbU

September, 2025

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Санкт-Петербургский государственный университет

Comments

In this lecture, we will consider a different approach to finding the minimum: an approach based on the use of a trust region. The idea is that instead of searching for the minimum of the objective function itself, we search for the minimum of a model—a quadratic approximation of the objective function. We will explore the question of the existence of a solution and general fundamental approaches such as the Cauchy point and the dogleg method.

- Minimizes quadratic model $m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p$ within trust region $\|p\| \le \Delta_k$.
- Chooses step direction and length together, unlike line search.

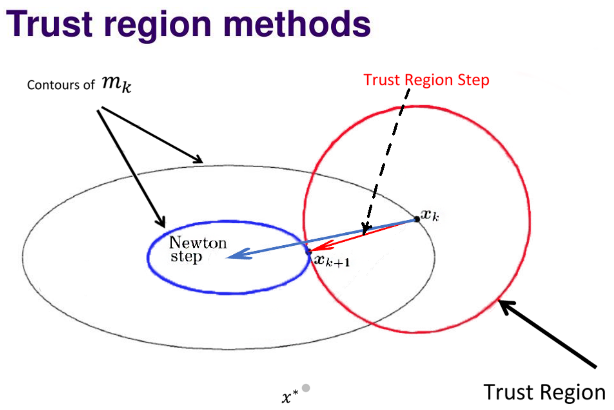

Figure: Trust-region step (red arrow within circle) vs. line search direction (blue arrow) for x_k to x^*.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Trust-region methods minimize a quadratic model m_k within a trust region of radius Δ_k, determining the step's direction and length simultaneously. This differs from line search methods, which first choose a direction and then adjust the step length α. In the figure, the blue arrow represents the Newton step, the unconstrained minimizer of m_k, which exceeds Δ_k.

The red arrow shows the Trust-Region Step, the minimizer of m_k within the trust region, staying inside the red circle. These steps are distinct: the Newton step ignores the trust region constraint, while the Trust-Region Step adheres to it. The size of Δ_k is key—too small slows progress toward x^*, too large risks model inaccuracy, prompting adjustments based on the step's success.

Basically, the core concept of trust-region methods lies in the fact that the quadratic model approximating the objective function within the trust region does not diverge significantly from the actual function, thereby allowing the iterative step to closely approach the true minimum of the function within that region.

- Quadratic model m_k at x_k based on Taylor expansion:

$$f(x_k+p) = f_k+g_k^T p+\frac{1}{2}p^T\nabla^2 f(x_k+tp)p,\ \text{where } f_k = f(x_k),\ g_k = \nabla f(x_k),\ t \in (0,1)$$

- Approximated model using $B_k \approx \nabla^2 f(x_k)$, symmetric:

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

- Subproblem: Minimize m_k within trust region:

$$\min_{p\in\mathbb{R}^n} m_k(p) \quad \text{s.t. } \|p\| \le \Delta_k \tag{3a}$$

Note: Error $m_k(p) - f(x_k + p) = O(\|p\|^2)$, or $O(\|p\|^3)$ if $B_k = \nabla^2 f(x_k)$. If B_k positive definite and $\|B_k^{-1}g_k\| \le \Delta_k$, solution is $p_k^B = -B_k^{-1}g_k$. In this case, we call p_k^B the full step.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Consider the Taylor expansion of the objective function around the point x_k up to the second-order term included. Let f_k be the value of the function f at the point x_k, and let g_k be the gradient of f at that point. The expansion takes the form $f(x_k + p) = f_k + g_k^T p + \frac{1}{2}p^T\nabla^2 f(x_k + tp)p$, where t is a scalar between zero and one.

Approximating the Hessian at $x_k + tp$ with a symmetric matrix B_k, the quadratic model is defined as $m_k(p) = f_k + g_k^T p + \frac{1}{2}p^T B_k p$. The error between $m_k(p)$ and $f(x_k + p)$ is of the order of the square of the norm of p, or of the order of the cube of the norm of p if $B_k = \nabla^2 f(x_k)$.

The trust-region subproblem minimizes $m_k(p)$ subject to the norm of p less than or equal to Δ_k, using the Euclidean norm. When B_k is positive definite and the norm of the inverse of B_k times g_k is less than or equal to Δ_k, the unconstrained minimizer $p_k^B = -B_k^{-1}g_k$ becomes the solution, known as the full step. Otherwise, an approximate solution ensures convergence.

Strategy: Choose trust-region radius Δ_k based on agreement between model m_k and objective f, using ratio

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

where the numerator is called the actual reduction, and the denominator is the predicted reduction.

Principle:

- Predicted reduction is always nonnegative since p_k minimizes m_k.
- If $\rho_k \approx 1$, model m_k aligns well with f, so expand Δ_k.
- If ρ_k is small or negative, shrink Δ_k to improve model accuracy.

Comments

Consider the strategy for selecting the trust-region radius at each iteration, denoted as Δ_k, which depends on the agreement between the model function m_k and the objective function f from previous steps. Define the ratio ρ_k as the actual reduction in the function value divided by the reduction predicted by the model.

The actual reduction is the difference between the function value at x_k and at x_k+p_k, while the predicted reduction is the difference between $m_k(0)$ and $m_k(p_k)$. Since p_k minimizes m_k over a region including zero, the predicted reduction is always nonnegative. When ρ_k is close to one, the model m_k aligns well with the function f, indicating that the model is a good approximation, so the radius Δ_k can be safely expanded for the next iteration to allow for larger steps. Conversely, if ρ_k is small or negative, the model does not accurately represent the function, so the radius Δ_k should be reduced to improve the model's accuracy and ensure better progress.

Algorithm 6 (Trust Region):

Require: $\hat{\Delta} > 0$, $\Delta_0 \in (0, \hat{\Delta})$, and $\eta \in [0, \frac{1}{4})$
for $k = 0, 1, 2, \ldots$ do
    Obtain p_k by (approximately) solving the subproblem, evaluate ρ_k.
    if $\rho_k < \frac{1}{4}$ then $\Delta_{k+1} = \frac{1}{4}\Delta_k$
    else
        if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$ then $\Delta_{k+1} = \min(2\Delta_k, \hat{\Delta})$
        else
            $\Delta_{k+1} = \Delta_k$
        end if
    end if
    if $\rho_k > \eta$ then $x_{k+1} = x_k + p_k$
    else
        $x_{k+1} = x_k$
    end if
end for

Note: $\hat{\Delta}$ is the upper bound on step length. Radius increases only if step reaches region boundary; otherwise, Δ_k remains unchanged.

Comments

Consider the process of iteratively adjusting the trust-region radius to balance the accuracy of the model with the progress of the optimization. The algorithm evaluates how well the model predicts the function's behavior by comparing actual and predicted reductions, using this to decide whether to accept a step or adjust the radius. If the model's prediction is poor, the radius shrinks to ensure the model becomes more accurate locally, preventing steps that might worsen the objective function. If the model performs well and the step fully utilizes the current radius, the radius expands to allow larger steps, accelerating convergence. Steps are accepted only when they provide sufficient improvement, ensuring steady progress toward the minimum.

Subproblem: Restate the trust-region subproblem as

$$\min_{p \in \mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p \quad \text{s.t. } \|p\| \le \Delta. \tag{3b}$$

Theorem 12 (Moré-Sorensen)

The vector p^* is a global solution of the trust-region subproblem

$$\min_{p \in \mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2} p^T B p \quad \text{s.t. } \|p\| \le \Delta,$$

if and only if p^* is feasible and there exists a scalar $\lambda \ge 0$ such that:

- $(B + \lambda I)p^* = -g,$ (4a)
- $\lambda(\Delta - \|p^*\|) = 0,$ (4b)
- $B + \lambda I$ is positive semidefinite. (4c)

Comments

To make the algorithm we discussed practical, we need to address the subproblem of finding the global minimum of the model m within the trust region defined by Δ. The following theorem provides the necessary and sufficient conditions for a given vector p^* to be the global minimizer of this subproblem. To achieve this, an auxiliary parameter $\lambda \ge 0$ is introduced. The theorem establishes that p^* is the global minimizer if and only if it satisfies a specific system of equations involving λ. This approach not only guarantees the optimality of the solution but also offers a computationally efficient method to determine p^*, thereby making the trust-region framework both theoretically robust and practically applicable.

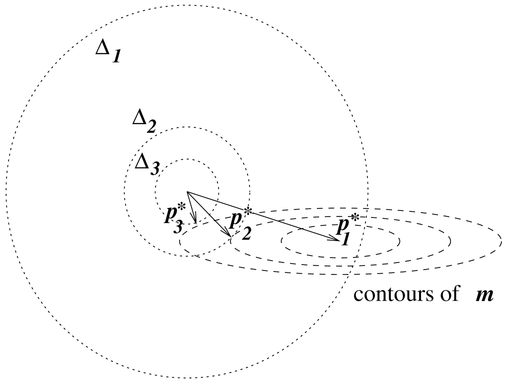

Figure: Solution of trust-region subproblem for different radii $\Delta_1, \Delta_2, \Delta_3$.

Key Features:

- Complementarity condition: $\lambda(\Delta - \|p^*\|) = 0$, where $\lambda \geq 0$.
- For $\Delta = \Delta_1$: p_1^* inside trust region, $\lambda = 0$, $Bp^* = -g$.
- For $\Delta = \Delta_2, \Delta_3$: p_2^*, p_3^* on boundary, $\lambda > 0$, p^* collinear with $-\nabla m(p^*)$.

Comments

The trust-region subproblem focuses on minimizing a quadratic model subject to the constraint that the norm of the solution step is less than or equal to the radius Δ. The image shows solutions, labeled as p_1^*, p_2^*, and p_3^*, for three different radii, called Δ_1, Δ_2, and Δ_3.

The complementarity condition states that the product of a scalar parameter λ and the difference between the radius and the norm of the solution step must equal zero, meaning either λ is zero or the norm of the solution step equals the radius.

For a large radius, Δ_1, the first optimal step lies strictly inside the trust region, with its norm less than Δ_1, so λ is zero, and the optimality condition—where the matrix B times the solution step equals the negative gradient—implies the first optimal step is the unconstrained minimizer, provided B allows it.

For smaller radii, Δ_2 and Δ_3, the second and third optimal steps lie on the boundaries, with their norms equal to Δ_2 and Δ_3, respectively, so λ is allowed to take a positive value. Here, the condition that λ times the solution step equals the negative of B times the solution step minus the gradient, which is also the negative gradient of the model at the solution step, implies that the solution step is collinear with the negative gradient of the model and normal to its contours, as seen in the alignment of the second and third optimal steps with the gradient direction in the figure.

Lemma 2

Let $m(p) = g^T p + \frac{1}{2} p^T B p$, where B is any symmetric matrix. Then the following statements are true:

(i) m attains a minimum if and only if B is positive semidefinite and g is in the range of B.
If B is positive semidefinite, then every p satisfying $Bp = -g$ is a global minimizer of m.

(ii) m has a unique minimizer if and only if B is positive definite.

Proof: (i) We start by proving the "if" part. Let B be positive semidefinite, then, since g is in the range of B, there is a p with $Bp = -g$. For all $w \in \mathbb{R}^n$, we have:

$$\begin{aligned} m(p+w) &= g^T(p+w) + \frac{1}{2}(p+w)^T B(p+w) \\ &= \left(g^T p + \frac{1}{2} p^T B p\right) + g^T w + (Bp)^T w + \frac{1}{2} w^T B w = \end{aligned}$$

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Now let's move on to the proof of the Theorem 12, which provides the conditions for identifying the exact solution to the trust-region subproblem, a critical step in optimizing a function within a constrained region. The proof builds on Lemma 2, which addresses the minimization of a quadratic function without constraints. The lemma establishes that a minimum exists when the matrix representing the function's curvature is positive semidefinite and the gradient lies within its range, allowing multiple solutions satisfying the gradient equation to be global minimizers. If the matrix is positive definite, the minimum is unique, ensuring a single optimal point. This foundation is essential for understanding how the trust-region constraint modifies the solution process.

We prove each of the claims in turn. We start by proving the "if" direction of part (i). Let B be positive semidefinite, then, since g is in the range of B, there is a p such that $Bp = -g$. For all $w \in \mathbb{R}^n$, we have the value of m at $p + w$ equals the inner product of g with $p + w$ plus one-half times the inner product of $p + w$ with B times $p + w$, which expands to the inner product of g with p plus one-half times the inner product of p with B times p plus the inner product of g with w plus the inner product of Bp with w plus one-half times the inner product of w with B times w is equal to

$$= m(p) + g^T w - g^T w + \frac{1}{2} w^T B w = m(p) + \frac{1}{2} w^T B w.$$

As B is positive semidefinite, $w^T B w \geq 0$, so $m(p+w) \geq m(p)$, proving p is a global minimizer.

- (i) For the “only if” direction we have If p minimizes m, then $\nabla m(p) = Bp + g = 0$, so g is in B’s range. The Hessian $\nabla^2 m(p) = B$ is positive semidefinite (by Theorem 3 (Second-Order Necessary Condition)).
- (ii) For the “if” part, the same argument as in (i) suffices with the additional point that $w^T B w > 0$ whenever $w \neq 0$.
- (ii) For the “only if” part, we proceed as in (i) to deduce that B is positive definite. If B is not positive definite, there is a vector $w \neq 0$ such that $Bw = 0$. Hence, we have $m(p+w) = m(p)$, so the minimizer is not unique, giving a contradiction. □

Comments

Continuing the proof, we simplify the expression to show that $f(p+w) = f(p) + \frac{1}{2} w^T B w$. Since B is positive semidefinite, this additional term is non-negative, ensuring $f(p+w) \geq f(p)$, which confirms p as a global minimizer.

For the "only if" direction: if p is a minimizer, the gradient at p, $\nabla f(p) = Bp + g$, must be zero, meaning $g \in \mathrm{col}(B)$, and the Hessian $\nabla^2 f(p) = B$ must be positive semidefinite.

For the second claim: if B is positive definite, then $w^T B w > 0$ for all $w \neq 0$, so $f(p+w) > f(p)$ for any perturbation, ensuring uniqueness of the minimizer. Conversely, if B is positive semidefinite but not positive definite, there exists some $w \neq 0$ such that $Bw = 0$, leading to $f(p+w) = f(p)$, which implies multiple minimizers and contradicts uniqueness.

Example for Lemma 2

Consider $B = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix}$, with eigenvalues 2, 0, 3, so B is positive semidefinite but singular. Let $m(p) = g^T p + \frac{1}{2} p^T B p$.

Case 1: $g = (4,\ 0,\ -6)^T$, second component zero, so g is in B's column space. Solve $Bp = -g$:

$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix} = \begin{bmatrix} -4 \\ 0 \\ 6 \end{bmatrix} \implies p_1 = -2,\ p_3 = 2,\ p_2 \text{ arbitrary.}$$

Thus, $p = (-2,\ p_2,\ 2)^T$ minimizes m. For $p = (-2,\ 0,\ 2)^T$, $m(p) = -8 - 12 + 10 = -10$.

Case 2: $g = (4,\ 1,\ -6)^T$, second component nonzero, so g is not in B's column space. Along direction $d = (0,\ -1,\ 0)^T$, compute the directional derivative: $\nabla m(p)^T d = g^T d = 1(-1) = -1 < 0$, and $d^T B d = 0$, so $m(p + td) = m(p) + t(-1) \to -\infty$ as $t \to \infty$, meaning no minimum exists.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

The example illustrates Lemma 2 by showing how the existence of a minimum for a quadratic function depends on the gradient's position relative to the matrix's column space. In the first case, the gradient has a zero second component, so it lies in the matrix's column space, allowing a solution to the gradient equation and yielding a minimum, as the function value remains constant along certain directions. In the second case, the gradient's second component is nonzero, placing it outside the matrix's column space, so the function decreases indefinitely along a specific direction where the matrix has no effect, demonstrating the absence of a minimum.

Proof of Theorem 12

Proof of Theorem 12: Assume $\lambda \geq 0$ satisfies conditions (4, i.e. (4a-4c)). By Lemma 2, p^* is a global minimizer of the modified quadratic:

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

Since $\hat{m}(p) \geq \hat{m}(p^*)$, we have:

$$m(p) \geq m(p^*) + \frac{\lambda}{2} \left((p^*)^T p^* - p^T p \right).$$

Given $\lambda(\Delta - \|p^*\|) = 0$, so $\lambda(\Delta^2 - (p^*)^T p^*) = 0$, this becomes:

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

Since $\lambda \geq 0$, $m(p) \geq m(p^*)$ for all p with $\|p\| \leq \Delta$, so p^* is a global minimizer of (3b).

For the converse, assume p^* is a global solution of (3b) and show that there is a $\lambda \geq 0$ that satisfies (4). If $\|p^*\| < \Delta$, then p^* is an unconstrained minimizer of m, so the gradient $Bp^* + g = 0$, and the Hessian B is positive semidefinite, satisfying (4) with $\lambda = 0$.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

We begin by assuming there exists a non-negative scalar λ satisfying the given conditions. Using Lemma 2, we show that the candidate solution is a global minimizer of a modified quadratic function, which includes an additional term involving λ times the squared norm of the step. This implies the original function's value at any point is at least the value at the candidate plus λ times half the difference of the squared norms of the candidate and the point. The complementarity condition allows us to adjust this inequality, and since λ is non-negative, we conclude the candidate minimizes the original function within the trust region.

For the converse, if the candidate is a global minimizer and its norm is less than the trust region radius, it must be an unconstrained minimizer, meaning the gradient at that point is zero and the Hessian is positive semidefinite, satisfying the conditions with $\lambda = 0$.

Assume for the remainder of the proof that $\|p^*\| = \Delta$. Then condition (4b) holds, and p^* solves the constrained problem: minimize $m(p)$ subject to $\|p\| = \Delta$. Using constrained optimization conditions (we'll discuss it later), there exists λ such that the Lagrangian:

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

is stationary at p^*. Setting the gradient of the Lagrangian with respect to p to zero we obtain:

$$Bp^* + g + \lambda p^* = 0 \implies (B + \lambda I)p^* = -g,$$

so condition (4a) holds. Since $m(p) \geq m(p^*)$ for any p with $p^T p = (p^*)^T p^* = \Delta^2$, we have:

$$m(p) \geq m(p^*) + \frac{\lambda}{2}\left((p^*)^T p^* - p^T p\right).$$

Substituting g from the previous equation, we get after rearrangement:

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

Comments

Assuming the norm of the candidate solution equals the trust region radius, the complementarity condition is immediately satisfied, and the candidate solves a constrained minimization problem where the norm of the step equals the radius. We introduce a Lagrangian function, which combines the original function with a penalty term involving the squared norm of the step minus the squared radius, scaled by half of λ.

Setting the gradient of this Lagrangian with respect to the step to zero, we derive that the matrix B times the candidate plus the gradient plus λ times the candidate equals zero, which rearranges to show the candidate satisfies the modified linear system, confirming one of the theorem's conditions. Since the function value at any point on the boundary is at least the value at the candidate, we adjust this inequality using the Lagrangian multiplier, and after substituting the gradient expression, we obtain a quadratic form involving the modified matrix.

Proof of Theorem 12 (Completed)

The set of directions $w = \pm\frac{p - p^*}{\|p - p^*\|}$, for $\|p\| = \Delta$, is dense on the unit sphere suffices to prove

$$(B + \lambda I) \text{ is positive semidefinite.}$$

To show $\lambda \geq 0$: Since: $(B + \lambda I)p^* = -g$, and $(B + \lambda I)$ is positive semidefinite, Lemma 2 implies p^* minimizes:

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

Suppose only negative λ satisfy the conditions. Then:

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

holds for $\|p\| \geq \Delta = \|p^*\|$. Since p^* minimizes m for $\|p\| \leq \Delta$, p^* is an unconstrained minimizer. By Lemma 2(i), $Bp^* = -g$ and B is positive semidefinite, so:

$$Bp^* + g = 0,$$

and $(B + 0 \cdot I)$ is positive semidefinite, satisfying the conditions with $\lambda = 0$, contradicting negative λ. Thus, $\lambda \geq 0$. □

Comments

Since the set of directions w is dense on the unit sphere, fulfilling the previous inequality suffices to prove the modified matrix is positive semidefiniteness . To prove λ is non-negative, we take advantage of the fact that the modified matrix satisfies the conditions of Lemma 2 and, therefore, by Lemma 2 the candidate minimizes a modified quadratic. Assuming only negative λ values work, the inequality implies the original function's value at points outside the trust region is at least the value at the candidate, and since the candidate minimizes within the region, it becomes an unconstrained minimizer. This leads to the gradient equation and positive semidefiniteness with $\lambda = 0$, contradicting the negative λ assumption, so λ must be non-negative. The theorem is proved.

Recall that we are looking for the optimal solution to the subproblem:

$$\min_{p \in \mathbb{R}^n} m_k(p) = f_k + g_k^T p + \frac{1}{2} p^T B_k p \quad \text{s.t. } \|p\| \le \Delta_k$$

Key Idea Global convergence of trust-region methods does not require exact minimization of the subproblem. It suffices to compute an approximate solution p_k inside the trust region that yields a sufficient reduction of the model.

Cauchy Point Algorithm:

Step 1. Solve the linearized subproblem:

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

Step 2. Find scalar $\tau_k > 0$ that minimizes $m_k(\tau p_k^s)$ subject to $\|\tau p_k^s\| \le \Delta_k$:

$$\tau_k = \arg\min_{\tau \ge 0} m_k(\tau p_k^s) \quad \text{s.t. } \|\tau p_k^s\| \le \Delta_k$$

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Although the trust-region subproblem is defined as a quadratic minimization, exact minimization is not required for global convergence. It is sufficient to compute an approximate solution that lies within the trust region and ensures a sufficient decrease in the model function. Such an approximation is provided by the so-called Cauchy point.

The first step of the algorithm is to solve a linearized minimization problem. Specifically, we minimize a linear model consisting of the current function value plus the directional derivative in the direction of the step, under the constraint that the step lies within the trust region, meaning its norm does not exceed the current trust-region radius. The solution to this constrained linear problem is referred to as the steepest descent step.

In the second step, we determine a positive scalar that minimizes the quadratic model along the direction of the steepest descent step, again under the constraint that the resulting scaled step remains within the trust region. This reduces to a one-dimensional optimization problem over the step length along that direction.

In the third step, we define the Cauchy point as the product of the computed scalar and the steepest descent direction. This point lies within the trust region and ensures sufficient decrease of the model function, which is essential for establishing convergence of the algorithm.

Step 3. Define the Cauchy point:

$$p_k^c = \tau_k p_k^s$$

Step 4. The solution of the trust-region subproblem is:

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

Step 5. For computing τ_k, consider the two cases:

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

Step 6. The Cauchy point is then:

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

Comments

It is easy to write down a closed-form definition of the Cauchy point. For a start, the solution of the trust-region subproblem is simply: $p_k^s = -\frac{\Delta_k}{\|g_k\|} g_k$.

To obtain the scalar parameter τ_k explicitly, we consider two cases. In the first case, when $g_k^T B_k g_k \le 0$, the function m_k evaluated at τ times the search direction decreases monotonically with τ, provided the gradient is nonzero. Therefore, τ_k is simply 1, the largest value that satisfies the trust-region constraint.

In the second case, when $g_k^T B_k g_k > 0$, the function $m_k(\tau p_k^s)$ is a convex quadratic in τ. In that case, τ_k is either the unconstrained minimizer of the quadratic, given by $\frac{\|g_k\|^3}{\Delta_k g_k^T B_k g_k}$, or the boundary value 1, whichever is smaller.

In summary, the Cauchy point is equal to $-\tau_k \frac{\Delta_k}{\|g_k\|} g_k$.

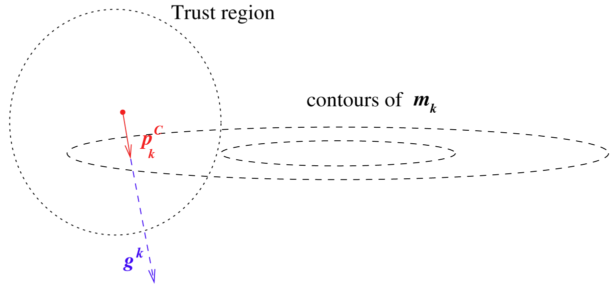

Figure: The Cauchy point p_k^c lies strictly inside the trust region. B_k is positive definite.

Key Insight: The Cauchy point p_k^c plays a central role in ensuring global convergence of trust-region methods.

Global Convergence Criterion: If the actual step p_k satisfies

$$m_k(0) - m_k(p_k) \geq \gamma\big(m_k(0) - m_k(p_k^c)\big) \quad \text{for some fixed } \gamma > 0,$$

then global convergence is guaranteed.

The Cauchy step is inexpensive to compute: it avoids matrix factorizations.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

The Cauchy point, denoted p_k^c, provides a baseline measure of progress for trust-region methods. It is defined without requiring any matrix factorizations, making it computationally inexpensive and robust. In the context of this figure, we consider the case when the matrix B_k is positive definite. This ensures that the Cauchy point lies strictly inside the trust region, that is, $\|p_k^c\| < \Delta_k$.

Crucially, the Cauchy step plays a pivotal role in verifying whether an approximate solution p_k to the trust-region subproblem is acceptable. Specifically, trust-region algorithms ensure global convergence by requiring that the actual step p_k achieves at least a constant fraction, denoted $\gamma > 0$, of the model decrease attained by the Cauchy point. This criterion enables global convergence without solving the subproblem exactly, emphasizing the importance of the Cauchy point in both theory and practice.

Beyond the Cauchy Point

Why not always use the Cauchy point?

- It ensures global convergence and is inexpensive to compute.
- However, it corresponds to the steepest descent method, which may converge slowly even with optimal step lengths.

Limitations:

- The Cauchy point is weakly dependent on B_k, which affects only the step length.
- Rapid convergence requires that B_k influence both direction and length, and that it reflect true curvature information.

Improvement strategies:

- Try to improve on the Cauchy point.
- Use $p_k^B = -B_k^{-1} g_k$ if $B_k \succ 0$ and $\|p_k^B\| \leq \Delta_k$.
- This yields superlinear convergence if $B_k = \nabla^2 f(x_k)$ or quasi-Newton.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Although the Cauchy point guarantees sufficient decrease in the model and is computationally cheap, always using it is equivalent to applying the steepest descent method with a certain step length. As we noted in the previous lecture, steepest descent converges slowly, even when optimal step lengths are used.

The Cauchy point depends only weakly on the matrix B_k, which only influences the step length. For rapid convergence, we need B_k to influence the step direction and to contain accurate curvature information.

Hence, many trust-region methods first compute the Cauchy point and then try to improve upon it. A common strategy is to take the full step $p_k^B = -B_k^{-1} g_k$ whenever B_k is positive definite and the norm of this step is less than or equal to Δ_k. If B_k is the true Hessian, or a good quasi-Newton approximation, this strategy typically leads to superlinear convergence.

Recall the trust-region subproblem:

$$\min_{p\in\mathbb{R}^n} m(p) \stackrel{\text{def}}{=} f + g^T p + \frac{1}{2}p^T Bp \quad \text{s.t. } \|p\| \le \Delta$$

Notation:

- Focus on one iteration: drop subscript k.
- Denote solution by $p^*(\Delta)$, emphasizing dependence on Δ.

Dogleg method:

- Applicable when B is positive definite.
- If $\Delta \ge \|p^B\|$, then $p^*(\Delta) = p^B$, where $p^B = -B^{-1}g$.
- If Δ is small, quadratic term in m has little effect:

  $$p^*(\Delta) \approx -\Delta \cdot \frac{g}{\|g\|}.$$
- For intermediate Δ, path of $p^*(\Delta)$ is curved.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

We begin by introducing the dogleg method — the first of three approaches to approximately solving the trust-region subproblem using curvature information. This method assumes that the symmetric matrix in the model is positive definite. To focus on a single iteration, we drop the iteration subscript and simplify the subproblem: the task is to minimize a quadratic model consisting of the function value, the inner product of the gradient and the step, and a quadratic term, subject to a bound on the step norm. If the trust-region radius exceeds the length of the unconstrained minimizer, then that minimizer becomes the solution. When the radius is very small, the optimal step approximates a scaled version of the negative gradient direction. In between, the solution follows a smooth trajectory between these two extremes, forming the characteristic path that gives the dogleg method its name.

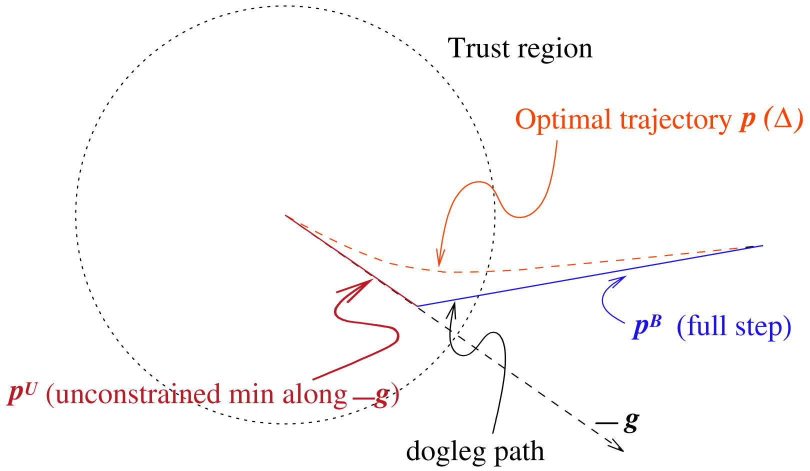

Figure: The dogleg method constructs a piecewise linear path from the origin to the Cauchy point, then toward the full Newton step.

Comments

This figure illustrates the geometry of the dogleg method, which is used in trust-region frameworks when the model matrix is positive definite. The method constructs a piecewise linear path starting from the origin in the direction of the negative gradient, leading to the so-called Cauchy point — the minimizer of the model along the steepest descent direction under the trust-region constraint. From there, the path continues toward the full Newton step, which minimizes the quadratic model without constraints. The trust-region boundary intersects this path, and the solution lies somewhere along it. This construction allows the method to account for both the linear and quadratic components of the model effectively.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Key Idea Approximate the nonlinear trajectory of the trust-region subproblem by a two-segment piecewise linear path: one segment along the steepest descent direction, followed by a segment toward the full Newton step.

Steepest descent point:

$$\mathrm{p}^{\mathrm{U}} = -\frac{\mathrm{g}^{\mathrm{T}}\mathrm{g}}{\mathrm{g}^{\mathrm{T}}\mathrm{B}\mathrm{g}}\mathrm{g}$$

Dogleg path: for $\tau \in [0, 2]$ (and $\mathrm{p}^{\mathrm{B}} = -\mathrm{B}^{-1}\mathrm{g}$

$$\tilde{\mathrm{p}}(\tau) = \begin{cases} \tau \mathrm{p}^{\mathrm{U}}, & 0 \le \tau \le 1, \\ \mathrm{p}^{\mathrm{U}} + (\tau - 1)(\mathrm{p}^{\mathrm{B}} - \mathrm{p}^{\mathrm{U}}), & 1 \le \tau \le 2. \end{cases}$$

The dogleg method minimizes the model m over this trajectory, subject to the trust-region constraint.

Comments

So, as we were saying, the dogleg method constructs an approximate solution to the trust-region subproblem using a piecewise linear trajectory. Instead of following the true curved path of the solution, it builds a path consisting of two straight-line segments. The first segment goes from the origin in the direction of steepest descent, ending at the point that minimizes the model along this direction. This point is calculated by taking the gradient, computing its squared norm, and dividing by the curvature along the gradient direction. The second segment continues from this steepest descent point toward the full Newton step, which is the unconstrained minimizer of the quadratic model.

The entire path is parametrized by a scalar variable ranging from zero to two. For parameter values between zero and one, the point lies along the steepest descent segment. For values between one and two, the point is on the segment connecting the steepest descent step to the Newton step.

The dogleg method selects the point along this path that minimizes the quadratic model, while also remaining within the trust region, which means its norm cannot exceed the current trust-region radius. In other words, the value of the parameter τ is chosen so that the norm of the vector $\tilde{\mathrm{p}}(\tau)$ is equal to Δ.

Lemma 3

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 easy to show that (i) and (ii) both hold for $\tau \in [0, 1]$, so we restrict our attention to the case of $\tau \in [1, 2]$. For (i), define $h(\alpha)$ by

$$\begin{aligned} h(\alpha) &= \frac{1}{2}\|\tilde{p}(1+\alpha)\|^2 \\ &= \frac{1}{2}\|p^U + \alpha(p^B - p^U)\|^2 \\ &= \frac{1}{2}\|p^U\|^2 + \alpha(p^U)^T(p^B - p^U) + \frac{1}{2}\alpha^2\|p^B - p^U\|^2. \end{aligned}$$

Our result is proved if we can show that $h'(\alpha) \geq 0$ for $\alpha \in (0, 1)$.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

The following lemma shows that the minimum along the dogleg path can be found easily.

There are two key monotonicity properties of the dogleg path. First, the norm of $\tilde{p}(\tau)$ grows as τ increases, and second, the model function decreases along this path.

Let us prove this. Notice that both statements are straightforward when τ is between zero and one, because in this case the step simply moves in the steepest descent direction. Therefore, we only need to consider τ between one and two.

To handle this case, define the function $h(\alpha)$ as one half the squared norm of $\tilde{p}(1+\alpha)$. Substituting the definition of the dogleg step, we obtain $h(\alpha) = \frac{1}{2}\|p^U + \alpha(p^B - p^U)\|^2$. Expanding gives a quadratic function in α.

If the derivative of this function, $h'(\alpha)$, is nonnegative for α in the interval zero to one, then the squared norm, and therefore the norm itself, is increasing. This proves part (i).

Now, compute the derivative $h'(\alpha)$:

$$h'(\alpha) = -(p^U)^T(p^U - p^B) + \alpha\|p^U - p^B\|^2 \geq -(p^U)^T(p^U - p^B) =$$

$$\frac{g^Tg}{g^TBg}g^T\left(-\frac{g^Tg}{g^TBg}g + B^{-1}g\right) = g^Tg\frac{g^TB^{-1}g}{g^TBg}\left[1 - \frac{(g^Tg)^2}{(g^TBg)(g^TB^{-1}g)}\right] \geq 0,$$

where the final inequality follows from the Cauchy-Schwarz inequality.

For (ii), we define $\hat{h}(\alpha) = m(\tilde{p}(1+\alpha)) =$

$$= f + g^T(p^U + \alpha(p^B - p^U)) + \frac{1}{2}(p^U + \alpha(p^B - p^U))^TB(p^U + \alpha(p^B - p^U))$$

and show that $\hat{h}'(\alpha) \leq 0$ for $\alpha \in (0,1)$:

$$\hat{h}'(\alpha) = (p^B - p^U)^T(g + Bp^U) + \alpha(p^B - p^U)^TB(p^B - p^U)$$

$$\leq (p^B - p^U)^T(g + Bp^U + B(p^B - p^U)) = (p^B - p^U)^T(g + Bp^B) = 0,$$

yielding the result. □

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

We now compute the derivative of $h(\alpha)$. Expanding gives $h'(\alpha) = -(p^U)^T(p^U - p^B) + \alpha\|p^U - p^B\|^2$. This is bounded below by the first term.

Substituting the definitions of p^U and p^B, we rewrite the expression in terms of the gradient g and the matrix B. After simplification, we obtain $g^Tg\frac{g^TB^{-1}g}{g^TBg}\left[1 - \frac{(g^Tg)^2}{(g^TBg)(g^TB^{-1}g)}\right]$. By the Cauchy–Schwarz inequality, this last factor is nonnegative. Therefore, $h'(\alpha) \geq 0$, which proves monotonicity of the norm. Thus the part (i) is proved.

For part (ii), we define $\hat{h}(\alpha) = m(\tilde{p}(1+\alpha))$. Expanding gives a quadratic expression in α. Its derivative is $\hat{h}'(\alpha) = (p^B - p^U)^T(g + Bp^U) + \alpha(p^B - p^U)^TB(p^B - p^U)$. Using the positive definiteness of the Hessian and the fact that α is at most 1, we bound this term above by the expression with $g + Bp^B$, which equals zero because p^B is the minimizer of the quadratic model. Hence, $\hat{h}'(\alpha) \leq 0$, and the model function decreases as τ increases. The Lemma is proved.

From Lemma 3, the path $\tilde{p}(\tau)$ intersects the trust-region boundary $\|p\| = \Delta$ at one point if $\|p^B\| \geq \Delta$, and nowhere otherwise. Since m decreases along the path, choose:

$$p = \begin{cases} p^B, & \text{if } \|p^B\| \leq \Delta, \\ \tilde{p}(\tau), & \text{at intersection } \|p^U + (\tau - 1)(p^B - p^U)\|^2 = \Delta^2. \end{cases}$$

Solve for τ: $$\|p^U + (\tau - 1)(p^B - p^U)\|^2 = \Delta^2.$$

For the model $m(p) = f + g^T p + \frac{1}{2} p^T B p, \quad \|p\| \leq \Delta$:

- If $\nabla^2 f(x_k)$ is positive definite, set $B = \nabla^2 f(x_k)$, so $p^B = -(\nabla^2 f(x_k))^{-1} g_k$.
- Otherwise, use a positive definite modified Hessian (see the previous lecture) for B.

Near a solution satisfying second-order sufficient conditions, p^B is the Newton step, enabling rapid convergence.

Note: Modified Hessians perturb $\nabla^2 f(x_k)$ arbitrarily, and trust-region solves with $B_k = \nabla^2 f(x_k)$ yield $-(\nabla^2 f(x_k) + \lambda I)^{-1} g_k$, where λ ensures positive definiteness.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

From the lemma on dogleg path properties, it follows that if the Newton step has a norm greater than the trust-region radius, then the path $\tilde{p}(\tau)$ intersects the boundary exactly once. Since the model function decreases along the path, the selected step will either be the full Newton step if it lies within the trust region, or the intersection point of the path with the boundary. In the latter case, τ is found by solving a scalar quadratic equation ensuring the step has norm equal to the trust-region radius.

When the exact Hessian is available and is positive definite, it can be directly used as B in the model, and the Newton step becomes $-(\nabla^2 f(x_k))^{-1} g_k$. Then the above procedure is followed: if this step is feasible, it is accepted; otherwise, τ is determined as described. If the Hessian is not positive definite, a positive definite modification of it is used instead, as we discussed earlier, and the same procedure is applied.

Near a solution that satisfies second-order sufficient conditions, the Hessian becomes positive definite, and the Newton step is used without modification, leading to fast local convergence.

Nevertheless, using a modified Hessian in the Newton-dogleg method is unsatisfying from an intuitive perspective. The modification perturbs the diagonals arbitrarily, which may undermine the trust-region strategy. Moreover, this modification is somewhat redundant because the trust-region approach already replaces the Hessian by the Hessian at the current iterate plus a scalar times the identity matrix, with the scalar chosen to make the matrix positive definite and depending on the trust-region radius.

Thus, the Newton-dogleg method is best suited for convex problems, where the Hessian is always positive semidefinite. In general nonconvex problems, it is preferable to use strategies that explicitly incorporate directions of negative curvature, meaning directions in which the quadratic form defined by the Hessian is negative.

For global convergence, the step p_k must achieve at least a fixed fraction of the Cauchy decrease:

$$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \tag{5}$$

for some constant $c_1 \in (0, 1]$.

This is satisfied by Dogleg methods.

When Δ_k is minimal in the expression, this condition resembles the first Wolfe condition.

Lemma 4 (Cauchy Decrease)

The Cauchy point p_k^c satisfies the bound with $c_1 = \frac{1}{2}$:

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

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

For global convergence of trust-region methods, each step p_k must guarantee a sufficient decrease of the quadratic model. This requirement is expressed as inequality (5), where the decrease must be at least a fixed fraction of the so-called Cauchy decrease. Explicitly, the model decrease is bounded below by a constant c_1 times the norm of the gradient, multiplied by the minimum of two terms: the trust-region radius Δ_k and the ratio of the gradient norm to the operator norm of B_k.

Intuitively, this condition means that the step cannot be too small in terms of model reduction. The dogleg method always satisfies this requirement, which ensures that it produces globally convergent iterations.

Lemma 4 states that the Cauchy point p_k^c, defined as the minimizer of the model along the steepest descent direction truncated by the trust region, achieves this inequality with constant $c_1 = \frac{1}{2}$. In other words, the Cauchy point guarantees at least half of the maximal possible Cauchy decrease. This result is fundamental, because it provides a uniform lower bound that can be used to prove convergence of trust-region algorithms.

Proof: Recall, that the Cauchy point is defined as:

$$p_k^c = -\tau_k \frac{\Delta_k}{\|g_k\|} g_k, \quad \text{where} \quad \tau = \begin{cases} 1 & \text{if } g_k^T B_k g_k < \frac{\|g_k\|^3}{\Delta}; \\ \frac{\|g_k\|^3}{\Delta g_k^T B_k g_k} & \text{otherwise.} \end{cases}$$

For simplicity, we omit the iteration index k during the proof.
Case 1: $g^T B g < \frac{\|g\|^3}{\Delta}$ then $\tau = 1$ and we have $m(p^c) - m(0) =$

$$m\left(-\Delta \frac{g}{\|g\|}\right) - f = -\Delta\|g\| + \frac{1}{2}\frac{\Delta^2}{\|g\|^2} g^T B g < -\frac{1}{2}\Delta\|g\| \leq -\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right)$$

Case 2: $g^T B g \geq \frac{\|g\|^3}{\Delta}$ then $\tau = \dfrac{\|g\|^3}{\Delta\, g^T B g}$ and we get

$$m(p^c) - m(0) = -\frac{\|g\|^4}{g^T B g} + \frac{1}{2}\frac{\|g\|^4}{g^T B g} = -\frac{1}{2}\frac{\|g\|^4}{g^T B g} \leq$$

$$-\frac{1}{2}\frac{\|g\|^4}{\|B\|\|g\|^2} = -\frac{1}{2}\frac{\|g\|^2}{\|B\|} \leq -\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right)$$

□

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Let's walk through this streamlined proof of Lemma 4, which demonstrates the effectiveness of the Cauchy point in trust-region optimization. The lemma shows that the Cauchy point, a strategic step we compute, reduces the model function enough to satisfy a key inequality. Recall that the Cauchy point is defined using a step size parameter that depends on the gradient, the trust-region radius, and the model's curvature, represented by a matrix B. The proof is carefully split into two cases based on the curvature condition, making it easier to follow, and we'll go through each to see why the Cauchy point is so reliable.

For simplicity, we omit the iteration index k during the proof. In the first case, when $g^T B g < \frac{\|g\|^3}{\Delta}$, the step size parameter τ equals one, meaning the Cauchy point takes the full trust-region radius in the steepest descent direction. The proof shows that this reduces the model function by greater than a half norm of the gradient times the radius, which satisfies the inequality.

In Case 2, when $g^T B g \geq \frac{\|g\|^3}{\Delta}$, τ is chosen so that the minimizer lies inside the trust region. Substituting this expression gives the decrease equal to $-\frac{1}{2}\frac{\|g\|^4}{g^T B g}$. This is then bounded by $-\frac{1}{2}\frac{\|g\|^2}{\|B\|}$. Again, this is less than or equal to $-\frac{1}{2}\|g\| \min\left(\Delta, \frac{\|g\|}{\|B\|}\right)$.

Therefore, in both cases the Cauchy point satisfies the desired decrease with constant one half.

To satisfy (5), it suffices that the approximate solution p_k achieves at least a fixed fraction c_2 of the reduction obtained by the Cauchy point:

$$m_k(0) - m_k(p_k) \geq c_2 \left(m_k(0) - m_k(p_k^c) \right).$$

Theorem 13

Let p_k satisfy $\|p_k\| \leq \Delta_k$ and $m_k(0) - m_k(p_k) \geq c_2 \left(m_k(0) - m_k(p_k^c) \right)$. Then p_k satisfies (5) with $c_1 = \frac{1}{2} c_2$. In particular, if p_k is the exact solution p_k^* of the trust-region subproblem, then $c_1 = \frac{1}{2}$.

Proof: Since $\|p_k\| \leq \Delta_k$, the previous Lemma implies

$$m_k(0) - m_k(p_k) \geq c_2 (m_k(0) - m_k(p_k^c)) \geq \frac{1}{2} c_2 \|g_k\| \min \left(\Delta_k, \frac{\|g_k\|}{\|B_k\|} \right).$$

□

Note: The dogleg algorithm satisfies (5) with $c_1 = \frac{1}{2}$, because it produces approximate solution p_k for which $m_k(p_k) \leq m_k(p_k^c)$.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Let's explore the use of the Cauchy point to ensure global convergence in trust-region optimization. The focus here is on a condition that guarantees our approximate solution is good enough to drive the optimization process toward a solution. The key idea is that any approximate step we take should reduce the model function by at least a fixed fraction of the reduction achieved by the Cauchy point, which we know is effective from our previous lemma. This slide presents a theorem that formalizes this idea, followed by a brief proof and a note about the dogleg algorithm.

The theorem states that if our step stays within the trust-region radius and achieves at least a fraction of the Cauchy point's reduction, it satisfies the required condition for convergence, with a constant that's half the fraction we choose. The proof is straightforward: it uses the lemma we proved earlier, which guarantees the Cauchy point's reduction, and scales it by the chosen fraction to show the step meets the convergence criterion.

Note that the dogleg algorithm, which constructs a path to approximate the trust-region solution, always does at least as well as the Cauchy point, ensuring the same convergence constant. This ties together the Cauchy point's reliability with practical algorithms, showing how they work together to ensure steady progress in optimization.

Global convergence results depend on the choice of the acceptance parameter η in Algorithm 6:

- For $\eta = 0$, the step is accepted if it reduces f, and the gradient sequence has a limit point at zero.
- For $\eta > 0$, actual reduction must be at least some small fraction of the predicted decrease we obtain the stronger result that $g_k \to 0$.

Assumptions for global convergence:

- The approximate Hessians B_k are uniformly bounded in norm.
- The objective function f is bounded below on the level set
$$S \stackrel{\text{def}}{=} \{x \mid f(x) \leq f(x_0)\}.$$
- We define a neighborhood of S by
$$S(R_0) \stackrel{\text{def}}{=} \{x \mid \|x - y\| < R_0 \text{ for some } y \in S\},$$
where $R_0 > 0$ is fixed.
- The step length may exceed the trust-region radius, as long as
$$\|p_k\| \leq \gamma \Delta_k, \quad \text{for some constant } \gamma \geq 1.$$

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Let's now explore the global convergence properties of trust-region methods, which tell us how the algorithm behaves over many iterations. Recall that we are talking about an algorithm that determines how the radius of the trust region changes depending on the ratio of the actual reduction to the predicted reduction.

Let's consider two scenarios based on a parameter that controls how strictly we accept a step. Next, we will prove that if we accept a step as soon as it reduces the objective function (the case $\eta = 0$, we can show that the gradients of the function at our iterates will have at least one point where they approach zero i.e. The sequence of gradients g_k has a subsequence converging to zero. However, if we require the step to achieve a small fraction of the predicted reduction, we get a stronger result: the gradients converge to zero, indicating we're approaching a stationary point.

To set the stage for the proofs, some key assumptions are introduced. The matrices approximating the curvature of the function are assumed to have bounded norms, ensuring they don't grow uncontrollably. The objective function is also assumed to be bounded below within a specific set of points (the level set) where the function value is no higher than at the starting point. Additionally, a neighborhood around this set is defined for flexibility in the analysis, and the step size is allowed to slightly exceed the trust-region radius, as long as it stays within a fixed multiple of it. These assumptions provide a solid foundation for proving that trust-region methods reliably drive the optimization process toward a solution.

Theorem 14

Let $\eta = 0$ in Algorithm 6. Suppose the following hold:

- $\|B_k\| \leq \beta$ for some constant β;
- f is bounded below on the level set S and is Lipschitz continuously differentiable in a neighborhood $S(R_0)$ for some $R_0 > 0$;
- all approximate solutions p_k of the trust-region subproblem (3a) satisfy the conditions

$$m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right), \quad \|p_k\| \leq \gamma \Delta_k,$$

for some constants $c_1 \in (0, 1]$, $\gamma \geq 1$.

Then we have $\liminf_{k\to\infty} \|g_k\| = 0.$

Comments

Let's dive into the global convergence theorem for trust-region methods when the step acceptance parameter is set to zero, meaning we accept any step that reduces the objective function. This theorem, labeled as Theorem 14, outlines conditions under which the algorithm guarantees that the norm of the gradient at some points in the iteration sequence approaches zero, a sign that we're nearing a stationary point. The theorem sets up a framework with specific assumptions to ensure this behavior.

The theorem assumes that the matrices approximating the function's curvature are uniformly bounded in norm, the objective function is bounded below on the level set defined by the initial function value and is Lipschitz continuously differentiable in an open neighborhood around this set, and the approximate solutions to the trust-region subproblem achieve a reduction in the model function at least proportional to the norm of the gradient and the trust-region radius (or a related term), while remaining within a fixed multiple of the radius.

Let me remind you that Lipschitz continuous differentiability means the norm of the difference between gradients at any two points in the neighborhood is bounded by a constant times the distance between those points, ensuring smooth behavior.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Proof: From Taylor's theorem (Theorem 1) we have:

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

Using the definition of m_k $m_k(p) = f(x_k) + g(x_k)^T p + \frac{1}{2} p^T B_k p$, it follows that

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

This gives the bound:

$$|m_k(p_k) - f(x_k + p_k)| \leq \frac{\beta}{2} \|p_k\|^2 + \beta_1 \|p_k\|^2,$$

where β_1 is the Lipschitz constant of g on $S(R_0)$. We assumed that $\|p_k\| \leq R_0$ to ensure that x_k and $x_k + tp_k$ lie in $S(R_0)$.

Comments

By Taylor's theorem, we express the value of the true objective function at the new point $x_k + p_k$ as the function value at x_k plus the scalar product of the gradient at x_k with the step p_k, plus an integral term that accounts for the change in the gradient along the step from x_k to $x_k + p_k$. This is compared to the model function, defined earlier, which includes the function value at x_k, the same gradient term, and a quadratic term involving the curvature matrix.

We then compute the absolute difference between the model function and the true function at p_k, which reduces to the difference between the quadratic term and the integral term. This difference is bounded using the norm of the curvature matrix and the Lipschitz constant of the gradient, assuming the step p_k stays within a neighborhood where the gradient is Lipschitz continuous.

Assumption for contradiction.
Suppose that there exist $\epsilon > 0$ and an index $K > 0$ such that

$$\|g_k\| \geq \epsilon, \quad \text{for all } k \geq K.$$

From the assumptions of the theorem, for all $k \geq K$ we have:

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

We analyze the acceptance ratio ρ_k:

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

Using the bound on the model accuracy and the step length, we get:

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

Comments

Next, to prove the theorem, we assume for contradiction that the norm of the gradient remains bounded below by a positive constant for all iterations beyond some point. By the theorem's assumptions, for iterations $k \geq K$, we use the bound on the model function reduction, which is at least a constant times the norm of the gradient multiplied by the minimum of the trust-region radius and a term involving the gradient norm and the curvature matrix norm. Given the contradiction assumption that the gradient norm is bounded below by a positive constant, this reduction is further bounded by a term proportional to the constant, the positive threshold ϵ, and the minimum of the trust-region radius and a scaled threshold ϵ.

We then evaluate the acceptance ratio ρ_k, which measures how close the actual reduction in the objective function is to the predicted reduction in the model. The absolute difference between this ratio and one is expressed as the ratio of the model-true function discrepancy to the model reduction. Using the earlier bounds on the model-true function difference, the model reduction, and the step size, we derive an upper bound on this difference, which depends on the square of the trust-region radius and constants from the curvature and Lipschitz properties.

Define the threshold that holds for all sufficiently small values of Δ_k, that is, for all $\Delta_k \le \bar{\Delta}$:

$$\bar{\Delta} = \min\left(\frac{1}{2}\cdot\frac{c_1\epsilon}{\gamma^2(\beta/2+\beta_1)}, \frac{R_0}{\gamma}\right).$$

The term R_0/γ in $\bar{\Delta}$ ensures that $\|p_k\| \le R_0$. Since $c_1 \le 1$ and $\gamma \ge 1$, we have $\bar{\Delta} \le \epsilon/\beta$, so

$$\min(\Delta_k, \epsilon/\beta) = \Delta_k \quad \text{for all } \Delta_k \le \bar{\Delta}.$$

Hence,

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

Therefore, $\rho_k \ge 1/2$, and Algorithm 6 guarantees that $\Delta_{k+1} \ge \Delta_k$ for all $\Delta_k \le \bar{\Delta}$. Thus, Δ_k is reduced (by a factor of $\frac{1}{4}$ only when

$$\Delta_k > \bar{\Delta},$$

which implies

$$\Delta_k \ge \min\left(\Delta_K, \bar{\Delta}/4\right) \quad \text{for all } k \ge K.$$

Comments

We now derive a bound on the right-hand side of the expression for the absolute difference between ρ_k and one, valid for all sufficiently small values of the trust-region radius, Δ_k, specifically those less than or equal to a threshold, $\bar{\Delta}$. This threshold is defined as the minimum of two terms: one proportional to the constant c_1 times the positive gradient threshold ϵ, divided by the square of γ times the sum of half the curvature bound β and the Lipschitz constant β_1, and another equal to the neighborhood radius R_0 divided by γ. The second term ensures the step's norm, at most γ times Δ_k, remains at most R_0, satisfying the step size bound.

From the definition of $\bar{\Delta}$ and the fact that $c_1 \le 1$, and $\gamma \ge 1$, it follows that $\bar{\Delta} \le \epsilon/\beta$, we simplify the denominator of the estimate for the absolute difference between ρ_k and one taking Δ_k as the smaller value for Δ_k up to $\bar{\Delta}$. Using this, we bound the absolute difference between ρ_k and one, reducing it to a term proportional to Δ_k times the sum of half β and β_1, divided by c_1 times ϵ. Substituting $\bar{\Delta}$, this difference is at most one-half, ensuring ρ_k is at least half. Thus, Algorithm 6 does not reduce Δ_k when Δ_k is at most $\bar{\Delta}$, as the step is sufficiently successful. Hence, a reduction of Δ_k by a factor of one-fourth occurs only when Δ_k is greater than $\bar{\Delta}$. Consequently, Δ_k remains at least the minimum of the initial radius at iteration K and one-fourth of $\bar{\Delta}$, ensuring the algorithm maintains a sufficiently large radius to advance the contradiction argument.

Proof of Theorem 14 (continued)

Suppose there is an infinite subsequence $\mathcal{K}$ such that $\rho_k \geq 1/4$ for $k \in \mathcal{K}$. For $k \in \mathcal{K}$ and $k \geq K$, we get

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

Since f is bounded below, this implies

$$\lim_{k \in \mathcal{K}, k \to \infty} \Delta_k = 0,$$

which contradicts the result that Δ_k is bounded from below. Hence, no such infinite subsequence $\mathcal{K}$ exists, and we must have $\rho_k < 1/4$ for all sufficiently large k. This, in turn,contradicts the result that $\rho_k \geq 1/2$ for all $\Delta_k \leq \bar{\Delta}$. Therefore, our assumption is false, and we conclude that

$$\liminf_{k \to \infty} \|g_k\| = 0.$$

□

Comments

We now suppose there exists an infinite subsequence of iterations, denoted by $\mathcal{K}$, where ρ_k is at least one-fourth for each iteration k in $\mathcal{K}$. For these iterations, when $k \geq K$, the decrease in the objective function from x_k to $x_k + p_k$ is at least one-fourth of the model function reduction, which itself is at least a constant c_1 times ϵ times the minimum of the trust-region radius Δ_k and ϵ divided by the curvature bound β. Since the objective function is bounded below, this positive decrease over an infinite subsequence implies that Δ_k must approach zero for k in $\mathcal{K}$ as k approaches infinity.

However, this contradicts the earlier result that Δ_k is bounded below by the minimum of the initial radius at iteration K and one-fourth of $\bar{\Delta}$. Thus, no such infinite subsequence $\mathcal{K}$ can exist, meaning ρ_k must be less than one-fourth for all sufficiently large k. This, in turn, contradicts the finding that ρ_k is at least one-half whenever Δ_k is at most $\bar{\Delta}$. Therefore, the assumption that the gradient norm remains bounded below by ϵ is false, leading to the conclusion that the limit inferior of the gradient norm as k approaches infinity is zero, completing the proof.

Theorem 15

Let $\eta \in \left(0, \frac{1}{2}\right]$ in Algorithm 6. Suppose that $\|B_k\| \le \beta$ 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 p_k of (3a) satisfy the inequalities $m_k(0) - m_k(p_k) \ge c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)$, $\|p_k\| \le \gamma\Delta_k$, for some constants $c_1 \in (0, 1]$, $\gamma \ge 1$. Then $\lim_{k\to\infty} g_k = 0$.

Proof: Let m be an index such that $g_m \ne 0$. Denote by β_1 the Lipschitz constant for g on $S(R_0)$ (i.e. $\|g(x) - g_m\| \le \beta_1 \|x - x_m\|$), and define

$$\epsilon = \frac{1}{2}\|g_m\|, \quad R = \min\left(\frac{\epsilon}{\beta_1}, R_0\right).$$

Then the ball $\mathcal{B}(x_m, R) = \{x \mid \|x - x_m\| \le R\}$ lies in $S(R_0)$, and Lipschitz continuity of g holds for all $x \in \mathcal{B}(x_m, R)$, thus we have

$$\|g(x)\| \ge \|g_m\| - \|g(x) - g_m\| \ge \frac{1}{2}\|g_m\| = \epsilon.$$

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

We now extend our convergence analysis to the case where the acceptance threshold η is strictly positive, with η chosen between zero and one-half. The assumptions are similar to the previous result. We suppose that the sequence of Hessian approximations, denoted by B_k, has bounded norm, specifically less than or equal to a constant β. We also assume that the objective function f is bounded below on the level set, and Lipschitz continuously differentiable on a slightly larger set with Lipschitz constant β_1. As before, every approximate solution p_k of the trust-region subproblem must satisfy two conditions: first, that the model decrease, written as $m_k(0) - m_k(p_k)$, is at least a constant c_1 times the gradient norm multiplied by the minimum of the trust-region radius Δ_k and the gradient norm divided by the norm of B_k; and second, that the step norm, the norm of p_k, does not exceed γ times Δ_k.

To start the proof, we fix an index m such that the gradient at iteration m is not zero. We then define ϵ to be one-half the gradient norm at this point, and we set a radius R equal to the minimum of ϵ divided by β_1 and R_0. The closed ball centered at x_m with radius R is fully contained in the region where Lipschitz continuity holds. Inside this ball, every point x satisfies that the gradient norm at x is at least ϵ. This construction will be used to develop the contradiction argument.

If the entire sequence $\{x_k\}_{k \geq m}$ stays inside the ball $\mathcal{B}(x_m, R)$, we would have $\|g_k\| \geq \epsilon > 0$ for all $k \geq m$. The reasoning in the proof of Theorem 14 can be used to show that this scenario does not occur. Therefore, the sequence $\{x_k\}_{k \geq m}$ eventually leaves $\mathcal{B}(x_m, R)$.

- Let $l \geq m$ be such that x_{l+1} is the first iterate after x_m outside $\mathcal{B}(x_m, R)$.
- Since $\|g_k\| \geq \epsilon$ for $k = m, \dots, l$, we use $\left(m_k(0) - m_k(p_k) \geq c_1 \|g_k\| \min\left(\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right)\right)$ to write:

$$f(x_m) - f(x_{l+1}) = \sum_{k=m}^{l} [f(x_k) - f(x_{k+1})] \geq \sum_{\substack{k=m \\ x_k \neq x_{k+1}}}^{l} \eta [m_k(0) - m_k(p_k)]$$

$$\geq \sum_{\substack{k=m \\ x_k \neq x_{k+1}}}^{l} \eta c_1 \epsilon \min\left(\Delta_k, \frac{\epsilon}{\beta}\right)$$

where we have limited the sum to the iterations k for which $x_k \neq x_{k+1}$, that is, those iterations on which a step was actually taken.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

Suppose for contradiction that the sequence of iterates remains entirely within this ball around x_m. In that case, the gradient norms would remain at least ϵ for all subsequent iterations. However, reasoning similar to that in the earlier proof shows that this cannot happen, and therefore the sequence must eventually leave the ball. Let l be the index such that the iterate x_{l+1} is the first point outside this ball. For every index from m through l, the gradient norm remains at least ϵ.

We now consider the decrease in the objective function over these iterations. By summing the reductions from x_m through x_{l+1}, and restricting attention only to those iterations where a step was actually taken, we find that the total reduction is at least η times the sum of the model decreases. Using the lower bound on the model decrease, this sum is bounded below by η times c_1 times ϵ times the minimum of Δ_k and ϵ divided by β, aggregated over the relevant iterations.

This inequality quantifies the guaranteed progress made whenever the gradient is bounded away from zero. It shows that the function value must decrease by a fixed positive amount whenever the algorithm is forced to exit the ball. This observation is a central step in reaching a contradiction, since the function is bounded below and cannot decrease indefinitely.

Global Convergence (Case $\eta > 0$, conclusion

- If $\Delta_k \le \epsilon/\beta$ for all $k = m, \ldots, l$, then
$$f(x_m) - f(x_{l+1}) \ge \eta c_1 \epsilon \sum_{\substack{k=m \\ x_k \neq x_{k+1}}}^{l} \Delta_k \ge \eta c_1 \epsilon R = \eta c_1 \epsilon \min\left(\frac{\epsilon}{\beta_1}, R_0\right)$$
- Otherwise, if $\Delta_k > \epsilon/\beta$ for some $k = m, \ldots, l$, then
$$f(x_m) - f(x_{l+1}) \ge \eta c_1 \epsilon \cdot \frac{\epsilon}{\beta}$$
- Since the sequence $\{f(x_k)\}_{k=0}^{\infty}$ is decreasing and bounded below, we have
$$f(x_k) \downarrow f^* \quad \text{for some } f^* > -\infty$$
- Therefore,
$$f(x_m) - f^* \ge f(x_m) - f(x_{l+1}) \ge \eta c_1 \epsilon \min\left(\frac{\epsilon}{\beta}, \frac{\epsilon}{\beta_1}, R_0\right)$$
$$= \frac{1}{2}\eta c_1 \|g_m\| \min\left(\frac{\|g_m\|}{2\beta}, \frac{\|g_m\|}{2\beta_1}, R_0\right) > 0$$
- Since $f(x_m) - f^* \to 0$, it must be that $\|g_m\| \to 0$.

$\square$

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

To complete the argument, we separate the analysis into two cases. First, suppose that for every relevant iteration, the trust-region radius Δ_k is less than or equal to ϵ divided by β. In this situation, the cumulative decrease in the function value from x_m to x_{l+1} is bounded below by η times c_1 times ϵ times the sum of Δ_k. Since the total step length needed to exit the ball is at least R, this bound implies that the function decreases by at least η times c_1 times ϵ times R. Substituting the definition of R, this becomes η times c_1 times ϵ times the minimum of ϵ divided by β_1 and R_0.

In the second case, suppose that for some iteration the trust-region radius Δ_k exceeds ϵ divided by β. Then the decrease in the function value from x_m to x_{l+1} is bounded below by η times c_1 times ϵ times ϵ divided by β.

Since the sequence of function values $f(x_k)$ is monotonically decreasing and bounded below, it must converge to some finite limit f^*. Combining this convergence with the inequalities we just derived, we see that $f(x_m) - f^*$ is greater than or equal to a strictly positive quantity that depends on the gradient norm at iteration m. But since $f(x_m)$ approaches f^*, this can only be true if the gradient norm itself approaches zero. This establishes the global convergence result. The theorem is proved.

Characterization of the Trust-Region Solution

The characterization of Theorem 12 suggests an algorithm for finding the solution p of the subproblem (3b):

$$\min_{p\in\mathbb{R}^n} m(p) = f + g^T p + \frac{1}{2}p^T B p \quad \text{s.t. } \|p\| \le \Delta.$$

Either $\lambda = 0$ satisfies the following conditions of Theorem 12:

- (4a): $(B + \lambda I)p^* = -g$
- (4c): $B + \lambda I$ is positive semidefinite, with $\|p\| \le \Delta$

or else for λ sufficiently large that $B + \lambda I$ is positive definite we define:

$$p(\lambda) = -(B + \lambda I)^{-1} g \quad \text{and find} \quad \lambda > 0 \text{ such that } \|p(\lambda)\| = \Delta$$

This is a one-dimensional root-finding problem in λ.

To see that a value of λ with all the desired properties exists, we appeal to the eigendecomposition of B and use it to study the properties of $\|p(\lambda)\|$.
Since B is symmetric, there is an orthogonal matrix Q and a diagonal matrix Λ such that

$$B = Q\Lambda Q^T, \quad \Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \dots, \lambda_n),$$

where $\lambda_1 \le \lambda_2 \le \dots \le \lambda_n$ are the eigenvalues of B.

Comments

We now turn to the characterization of the trust-region solution. The trust-region subproblem consists of minimizing the quadratic model $m(p) = f + g^T p + \frac{1}{2}p^T B p$, subject to the constraint that the norm of p does not exceed the trust-region radius Δ. The Moré-Sorensen theorem states that there exists a scalar λ such that the step p^* satisfies the equation $(B + \lambda I)p^* = -g$, and that the matrix $B + \lambda I$ is positive semidefinite. Furthermore, either $\lambda = 0$ and the step norm is within Δ, or else we can define $p(\lambda) = -(B + \lambda I)^{-1} g$. Then we seek a strictly positive value of λ for which the norm of $p(\lambda)$ equals Δ.

This reduces the subproblem to a one-dimensional root-finding task in the scalar variable λ. To ensure that such a λ exists, we analyze the eigenstructure of the Hessian approximation B. Since B is symmetric, it can be decomposed as $Q\Lambda Q^T$, where Q is an orthogonal matrix and Λ is diagonal with entries λ_1 through λ_n.

Behavior of $\|p(\lambda)\|$ via Eigendecomposition

Since $B + \lambda I = Q(\Lambda + \lambda I)Q^T$, we obtain for $\lambda \neq \lambda_j$:

$$p(\lambda) = -Q(\Lambda + \lambda I)^{-1} Q^T g = -\sum_{j=1}^{n} \frac{q_j^T g}{\lambda_j + \lambda} q_j,$$

where q_j is the jth column of Q.

By orthonormality of the columns of Q, the squared norm is

$$\|p(\lambda)\|^2 = \sum_{j=1}^{n} \frac{(q_j^T g)^2}{(\lambda_j + \lambda)^2}.$$

If $\lambda > -\lambda_1$, then $\lambda_j + \lambda > 0$ for all $j = 1, 2, \dots, n$, so $\|p(\lambda)\|$ is a continuous, nonincreasing function of λ on $(-\lambda_1, \infty)$.

We have:

- $\lim_{\lambda \to \infty} \|p(\lambda)\| = 0$,
- moreover, if $q_j^T g \neq 0$ then $\lim_{\lambda \to -\lambda_j} \|p(\lambda)\| = \infty$.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

This decomposition allows us to express $p(\lambda)$ explicitly as a weighted sum of eigenvectors q_j, with coefficients depending on the inner products $q_j^T g$ and the shifted eigenvalues $\lambda_j + \lambda$. As a result, the squared norm of $p(\lambda)$ can be written as a sum over j of the squared inner product divided by the squared shifted eigenvalue.

This formula has three immediate consequences. First, continuity and monotonicity: for $\lambda > -\lambda_1$, all denominators are positive; increasing λ increases every denominator, so the norm of $p(\lambda)$ is a continuous, nonincreasing function of λ on the interval $(-\lambda_1, \infty)$. Second, limits: as λ tends to infinity, every term vanishes and the norm goes to zero—large λ suppresses all directions uniformly. If for some index j the projection $q_j^T g$ is nonzero, then as λ approaches $-\lambda_j$ from the right, the corresponding denominator tends to zero and the norm blows up to infinity. Third, existence and typically uniqueness of the trust-region multiplier: because the norm of $p(\lambda)$ decreases continuously from infinity down to zero on that interval, there exists a value of $\lambda \geq 0$ such that the norm of $p(\lambda)$ equals the trust-region radius Δ. When the projections are not all zero, this solution is unique.

Intuitively, λ acts as a global damping knob: increasing λ shrinks components along "soft" eigen-directions—those with small or negative λ_j—more aggressively, yielding a step that neatly fits the prescribed radius.

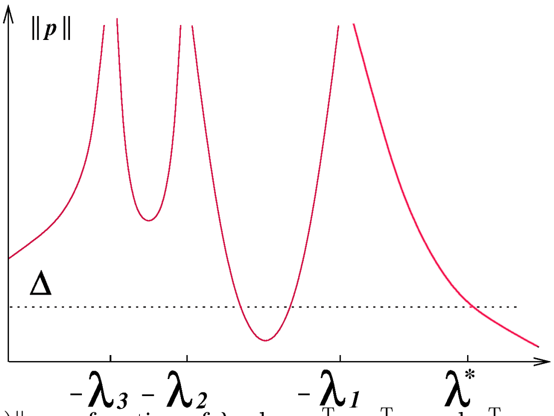

Figure: $\|p(\lambda)\|$ as a function of λ when $q_1^T g$, $q_2^T g$, and $q_3^T g$ are all nonzero.

Note: If the properties of $\|p(\lambda)\|$ derived on the previous slide hold and $\|p(\lambda)\|$ is a continuous, nonincreasing function on $(-\lambda_1, \infty)$; If $q_1^T g \neq 0$, then there exists a unique $\lambda^* \in (-\lambda_1, \infty)$ such that $\|p(\lambda^*)\| = \Delta$ (other smaller values of λ might satisfy this equation, but they violate condition (4c)).

Comments

Let us now take a closer look at the function that defines the step length as a function of the parameter λ. The key property is monotonicity. As λ increases beyond $-\lambda_1$, the smallest eigenvalue, each denominator becomes larger, and therefore the step norm decreases. Hence, the function $\|p(\lambda)\|$ is continuous and strictly decreasing on the interval from $-\lambda_1$ to infinity. Its limits are also clear: as λ tends to $-\lambda_1$ from above, the step length tends to infinity; as λ tends to infinity, the step length approaches zero.

From this, we conclude that for any trust-region radius Δ, there must exist a unique value of $\lambda^* > -\lambda_1$, such that $\|p(\lambda^*)\| = \Delta$. Graphically, we see a curve descending smoothly from infinity to zero, and the horizontal line at Δ will intersect this curve exactly once. This intersection defines λ^*. The uniqueness of this solution is crucial, because it ensures that our trust-region subproblem always has a well-defined and consistent answer. This forms the basis for the numerical procedures we develop next.

Key Idea Identify the value of $\lambda^* \in (-\lambda_1, \infty)$ for which the norm of the step equals the trust-region radius, that is, $\|p(\lambda^*)\| = \Delta$, which works when $q_1^T g \neq 0$. Use a root-finding procedure based on properties of $\|p(\lambda)\|$.

If B is positive definite and $\|B^{-1}g\| \leq \Delta$, then $\lambda^* = 0$ satisfies the conditions (4a)-(4c) of Theorem 12, so the procedure can be terminated immediately with $\lambda^* = 0$.
Otherwise (in the case $q_1^T g \neq 0$, we solve the nonlinear equation

$$\phi_1(\lambda) = \|p(\lambda)\| - \Delta = 0$$

for $\lambda > -\lambda_1$ using Newton's method.

Near the asymptote at $\lambda = -\lambda_1$, the function $\phi_1(\lambda)$ behaves like

$$\phi_1(\lambda) \approx \frac{C_1}{\lambda + \lambda_1} + C_2 \quad \text{with} \quad C_1 > 0.$$

This approximation highlights potential instability of Newton's method near $-\lambda_1$ (i.e. it may be unreliable or slow).

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

The next step is to discuss how λ^* can actually be computed. The key idea is to identify the value of λ^* within the interval from $-\lambda_1$ to infinity for which the norm of $p(\lambda)$ equals the trust-region radius Δ. If the matrix B is positive definite and the unconstrained minimizer, which is $-B^{-1}g$, already lies inside the trust region, then the trust-region constraint is inactive, and the correct value is simply $\lambda^* = 0$. In this case, no iteration is necessary. However, in the more general situation, when $q_1^T g \neq 0$ and the unconstrained minimizer lies outside the trust region, we must solve the nonlinear equation $\phi_1(\lambda) = \|p(\lambda)\| - \Delta = 0$. This is precisely a root-finding problem. Newton's method is the natural choice here because it can take advantage of the smoothness of the function. However, a difficulty arises near the singularity at $\lambda = -\lambda_1$. In this region, $\phi_1(\lambda)$ behaves approximately like $C_1/(\lambda + \lambda_1) + C_2$, with $C_1 > 0$. This means that near the singularity, Newton's method can become unstable or converge slowly. Thus, although Newton's method is powerful, care must be taken when λ is close to the negative of the smallest eigenvalue.

Improved Root-Finding via Reformulation

Key Idea Improve convergence and numerical stability by reformulating the root-finding problem using an alternative function that behaves nearly linearly near the singularity at $-\lambda_1$.

To avoid instability near $-\lambda_1$, define:

$$\phi_2(\lambda) = \frac{1}{\Delta} - \frac{1}{\|\mathrm{p}(\lambda)\|}$$

so that ϕ_2 is nearly linear near $-\lambda_1$. Using the expansion of $\|\mathrm{p}(\lambda)\|$, we get:

$$\phi_2(\lambda) \approx \frac{1}{\Delta} - \frac{\lambda + \lambda_1}{\mathrm{C}_3}, \quad \mathrm{C}_3 > 0$$

The root-finding Newton's method will perform well and it iteration becomes:

$$\lambda^{(\ell+1)} = \lambda^{(\ell)} - \frac{\phi_2(\lambda^{(\ell)})}{\phi_2'(\lambda^{(\ell)})}$$

Comments

To overcome the difficulty near the singularity, the root-finding problem is reformulated in terms of an alternative function. Instead of directly working with ϕ_1, we introduce $\phi_2(\lambda) = 1/\Delta - 1/\|\mathrm{p}(\lambda)\|$. The advantage of this formulation is that ϕ_2 behaves nearly linearly near the critical point at $-\lambda_1$. Using the expansion of the step norm, one can show that $\phi_2(\lambda) \approx 1/\Delta - (\lambda + \lambda_1)/\mathrm{C}_3$, with $\mathrm{C}_3 > 0$. This linear-like behavior makes Newton's method far more reliable. The Newton iteration is then expressed as $\lambda^{(\ell+1)} = \lambda^{(\ell)} - \phi_2(\lambda^{(\ell)})/\phi_2'(\lambda^{(\ell)})$. Thanks to the improved conditioning of the problem, convergence becomes faster and more stable, especially when λ approaches values close to $-\lambda_1$. This reformulation illustrates a general principle in numerical optimization: sometimes, by changing the function under consideration, we can transform a numerically ill-conditioned problem into one that is far more stable and efficient to solve.

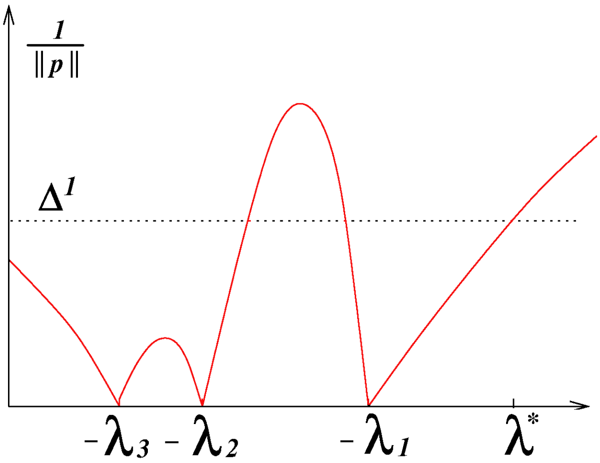

Figure: $1/\|p(\lambda)\|$ as a function of λ.

The shape of $1/\|p(\lambda)\|$ shows that $\phi_2(\lambda)$ behaves nearly linearly near $-\lambda_1$, improving stability of Newton iterations.

Comments

Finally, let us interpret the graph of the reformulated function. Instead of plotting the step norm directly, we plot its reciprocal, $1/\|p(\lambda)\|$. This transformation makes the behavior near the singularity transparent. Specifically, as λ approaches $-\lambda_1$, the reciprocal of the norm grows approximately linearly, which confirms that $\phi_2(\lambda)$ indeed behaves like a straight line in this region. As a result, when Newton's method is applied to ϕ_2, the iterations are not slowed down by the presence of the singularity but instead proceed smoothly. The graphical evidence clearly demonstrates that the reformulation is not only theoretically justified but also practically effective. Thus, by introducing the reciprocal function, we obtain a numerically stable and efficient root-finding process for determining the trust-region parameter λ^*. This completes the analysis of existence, uniqueness, and computation of λ^*, and provides the foundation for implementing robust trust-region algorithms.

Key Idea Use Newton's method on a transformed root-finding function to find $\lambda^* > -\lambda_1$ such that $\|p(\lambda^*)\| = \Delta$, computing $p(\lambda)$ via Cholesky factorization.

It easy to show by direct calculations that

$$\frac{\phi_2(\lambda^{(\ell)})}{\phi_2'(\lambda^{(\ell)})} = \frac{\frac{1}{\|p\|} - \frac{1}{\Delta}}{\frac{p^T(B+\lambda I)^{-1}p}{\|p\|^3}} = \frac{\|p\|^2}{p^T(B+\lambda I)^{-1}p)}\left(\frac{\Delta - \|p_\ell\|}{\Delta}\right)$$

Algorithm 7 (Trust Region Subproblem):

Require: Initial guess $\lambda^{(0)}$, trust region radius $\Delta > 0$

1: for $\ell = 0, 1, 2, \ldots$ do

2: Factor $B + \lambda^{(\ell)} I = R^T R$ (Cholesky)

3: Solve $R^T R p_\ell = -g$ and then $R^T q_\ell = p_\ell$

4: Update
$$\lambda^{(\ell+1)} = \lambda^{(\ell)} - \left(\frac{\|p_\ell\|}{\|q_\ell\|}\right)^2 \left(\frac{\Delta - \|p_\ell\|}{\Delta}\right)$$

5: end for

Note: When $\lambda^{(\ell)} < -\lambda_1$, Cholesky factorization fails. In practice, the algorithm with safeguards usually converges in a few steps.

Trust-Region Methods
Global solution
Cauchy Point
Dogleg Method
Global Convergence
TR Subproblem Algorithm

Comments

The presented algorithm for solving the trust-region subproblem is a practical implementation of the ideas discussed earlier. Its main goal is to find the value of the parameter λ that ensures the fulfillment of the condition $\|p(\lambda)\| = \Delta$. To achieve this, Newton's method is applied to a specially transformed function. At each step, the algorithm performs a decomposition of the matrix $B + \lambda I$ using the Cholesky factorization, which guarantees the numerical stability of the computations provided the corresponding matrix is positive definite. Next, a system of linear equations is solved to find the current approximation p_ℓ and an auxiliary vector q_ℓ. The iterative process involves adjusting the parameter λ, and the update step is designed to bring the norm of the found vector p_ℓ closer to the given trust-region radius. It is important to understand that practical implementations of this method usually do not require driving λ to high precision; a few iterations are sufficient to achieve an acceptable approximation. However, the method has a limitation: if λ becomes less than $-\lambda_1$, where λ_1 is the smallest eigenvalue of the matrix B, then the Cholesky factorization becomes impossible, and it becomes necessary to introduce special "safeguards." These protective measures prevent failures and make the algorithm applicable in practice even for ill-conditioned problems.