Lecture 25–26: Trust-Region Methods: Improving Cauchy Point; Convergence

Yudong Chen

Recall: TR sub-problem:

$$\min_{p:\|p\|\le\Delta_k} m_k(p) := f_k + \langle g_k, p\rangle + \frac{1}{2}p^\top B_k p, \tag{P_{m_k}}$$

We want to approximate the exact minimizer $p_k^*(\Delta_k)$.

1 Improving the Cauchy point

Recall Cauchy point: $p_k^{\mathrm{C}} = \tau_k p_k^{\mathrm{S}}$, where

$$p_k^{\mathrm{S}} = \operatorname*{argmin}_{p:\|p\|\le\Delta}\left\{f_k + g_k^\top p\right\} = -\frac{\Delta_k}{\|g_k\|}g_k, \tag{1a}$$

$$\tau_k = \operatorname*{argmin}_{\tau\ge 0:\left\|\tau p_k^{\mathrm{S}}\right\|\le\Delta} m_k(\tau p_k^{\mathrm{S}}) = \begin{cases} 1 & g_k^\top B_k g_k \le 0, \\ \min\left\{1, \frac{\|g_k\|^3}{\Delta g_k^\top B_k g_k}\right\}, & g_k^\top B_k g_k > 0. \end{cases} \tag{1b}$$

We discuss two ways of improving upon the Cauchy point: the dogleg method, and 2-D subspace minimization.

1.1 The dogleg method

This methods is typically used only when $B_k \succ 0$. It interpolates between two extremes:

- If Δ is small, then $m_k(p) \approx f_k + g_k^\top p$ for $\|p\| \le \Delta$, hence $p_k^* \approx p_k^{\mathrm{C}}$ (gradient descent direction).
- If Δ is large, then $\|p\| \le \Delta$ becomes irrelevant, hence $p_k^* \approx -B_k^{-1}g_k$ (unconstrained minimizer).

Formally, define

$$p_k^{\mathrm{U}} := -\frac{g_k^\top g_k}{g_k^\top B_k g_k}g_k = \text{(unconstrained) GD step with exact line search}$$
$$p_k^{\mathrm{B}} := -B_k^{-1}g_k = \text{unconstrained minimizer of } m_k$$

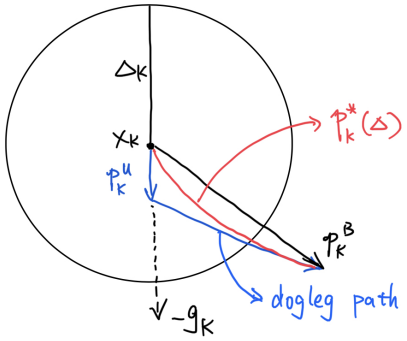
Consider the "dogleg path" defined below:

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

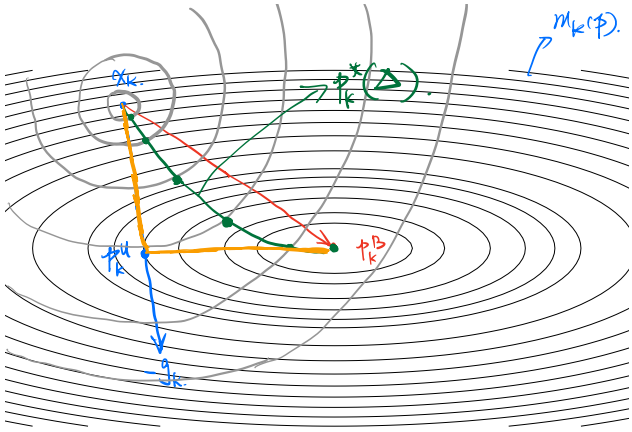
Note that $\tilde{p}_k(\tau)$ consists of two line segments and is an approximation of the optimal path $p_k^*(\Delta)$. The dogleg step is given by constrained minimizer over the path $\tilde{p}(\tau)$, i.e.,

$$p_k^{\mathrm{D}} := \min_{\substack{0\le\tau\le 2 \\ \|\tilde{p}_k(\tau)\|\le\Delta}} m_k\left(\tilde{p}_k(\tau)\right).$$

Illustration:



Another illustration:



Thanks to the following lemma, it is easy to compute the minimizer p_k^{D} along the dogleg path.

Lemma 1 (Lemma 4.2 in Nocedal-Wright). *Let B_k be positive definite. Then*

(i) $\|\tilde{p}_k(\tau)\|$ *is an increasing function of* τ*;*

(ii) $m_k\left(\tilde{p}_k(\tau)\right)$ *is a decreasing function of* τ*.*

Consequently:

- If $\left\|p^{\mathrm{B}}\right\| < \Delta$, then the dogleg path does not intersect the TR boundary $\|p\| = \Delta$. Since m_k is decreasing in τ,we have $p_k^{\mathrm{D}} = \tilde{p}_k(2) = p^{\mathrm{B}}$.
- If $\left\|p^{\mathrm{B}}\right\| \ge \Delta$, then the dogleg path intersects the boundary at one point, which is p_k^{D}. The corresponding τ can be computed by solving the scalar equation $\|\tilde{p}_k(\tau)\| = \Delta$.

1.2 Two-dimensional subspace minimization

The dogleg method minimizes over the one-dimensional path defined by p^{U} and p^{B}. This can generalized by minimizing over the 2-D subspace spanned by $p^{\mathrm{U}} \propto -g_k$ and $p^{\mathrm{B}} = -B_k^{-1} g_k$.

Formally:

$$p_k^{\mathrm{2D}} = \operatorname*{argmin}_{p \in \mathbb{R}^d} \left\{ m_k(p) : \|p\| \leq \Delta_k, p \in \operatorname{span}\{g_k, B_k^{-1} g_k\} \right\}.$$

The minimizer is relatively easy to compute (amounts to finding the roots of a fourth degree polynomial).

Unlike dogleg, 2D-subspace minimization can readily be adapted to handle indefinite B_k. In this case, there exists $\lambda > 0$ such that $p_k^* = -(B_k + \lambda I)^{-1} g_k$ (by Theorem 1 from the last lecture). Therefore, we can change the feasible 2D subspace to

$$\operatorname{span}\left\{ g_k, (B_k + \alpha_k I)^{-1} g_k \right\},$$

where $\alpha_k \in (-\lambda_{\min}(B_k), -2\lambda_{\min}(B_k))$.

2 Global convergence of TR methods

With respect to the model m_k, both dogleg and 2D subspace minimization are at least as good as taking the Cauchy point: the 2D subspace contains the dogleg path, which in turn contains the Cauchy point. Consequently, they enjoy global convergence, as we show below.

2.1 Progress made by Cauchy point

Recall that $f_k = f(x_k), g_k = \nabla f(x_k)$, and

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

The lemma below quantifies the progress on m_k made by the Cauchy point.

Lemma 2. *(Progress by Cauchy point; Lemma 4.3 in Nocedal-Wright) The Cauchy point p_k^C satisfies*

$$m_k\left(p_k^C\right) - m_k(0) \leq -\frac{1}{2} \|g_k\|_2 \min\left\{ \Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2} \right\}.$$

Before proving the lemma, we briefly mention the intuition on how to use this lemma to prove global convergence. Assume that $m_k(0) = f(x_k)$ (always true), $m_k\left(p_k^C\right) \approx f(x_{k+1})$, $B_k = \nabla^2 f(x_k) \preccurlyeq LI$ and $\frac{\|g_k\|}{\|B_k\|} \leq \Delta_k$ Then Lemma 2 implies the sufficient descent property

$$f(x_{k+1}) - f(x_k) \leq -\frac{1}{2L} \|\nabla f(x_k)\|_2^2,$$

which in turn guarantees global convergence.

Proof of Lemma 2. Note that

$$m_k(p) - m_k(0) = g_k^\top p + \frac{1}{2} p^\top B_k p.$$

Consider the three cases in the Cauchy point calculation in (1).

Case 1: $g_k^\top B_k g_k \leq 0$. Then $p_k^C = p_k^S = -\frac{\Delta_k}{\|g_k\|} g_k$. Hence

$$\begin{aligned} m_k(p_k^C) - m_k(0) &= -\|g_k\| \Delta_k + \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} \underbrace{g_k^\top B_k g_k}_{\leq 0} \\ &\leq -\|g_k\| \Delta_k \\ &\leq -\frac{1}{2} \|g_k\| \cdot \min\left\{ \Delta_k, \frac{\|g_k\|}{\|B_k\|} \right\}. \end{aligned}$$

Case 2: $g_k^\top B_k g_k > 0$ and $\frac{\|g_k\|^3}{\Delta_k g_k^\top B_k g_k} \leq 1$. Then:

$$\tau_k = \frac{\|g_k\|^3}{\Delta_k g_k^\top B_k g_k}, \qquad p_k^C = -\frac{g_k \|g_k\|^2}{g_k^\top B_k g_k}.$$

Hence

$$\begin{aligned} m_k(p_k^C) - m_k(0) &= -\frac{\|g_k\|^4}{g_k^\top B_k g_k} + \frac{1}{2} \frac{\|g_k\|^4}{g_k^\top B_k g_k} \\ &= -\frac{1}{2} \frac{\|g_k\|^4}{g_k^\top B_k g_k} \\ &\leq -\frac{1}{2} \frac{\|g_k\|^4}{\|B\|_2 \|g_k\|^2} = -\frac{1}{2} \frac{\|g_k\|^2}{\|B\|_2} \\ &\leq -\frac{1}{2} \|g_k\|_2 \cdot \min\left\{ \Delta_k, \frac{\|g_k\|_2}{\|B_k\|_2} \right\}. \end{aligned}$$

Case 3: $g_k^\top B_k g_k > 0$ and $\frac{\|g_k\|^3}{\Delta_k g_k^\top B_k g_k} > 1$. The latter implies that $g_k^\top B_k g_k < \frac{\|g_k\|^3}{\Delta_k}$ and thus $p_k^C = p_k^S$. Hence

$$\begin{aligned} m_k\left(p_k^C\right) - m_k(0) &= -\|g_k\| \Delta_k + \frac{1}{2} \frac{\Delta_k^2}{\|g_k\|^2} \underbrace{g_k^\top B_k g_k}_{< \frac{\|g_k\|^2}{\Delta_k}} \\ &\leq -\frac{1}{2} \|g_k\| \Delta_k \\ &\leq -\frac{1}{2} \|g_k\| \cdot \min\left\{ \Delta_k, \frac{\|g_k\|}{\|B_k\|} \right\}. \end{aligned}$$

In all three cases, we establish the desired inequality. □

The theorem below, which follows trivially from Lemma 2, quantifies the decrease obtained by any solution that achieves some constant fraction of the progress by the Cauchy point.

Theorem 1 (Theorem 4.4 in Nocedal-Wright). *Let p_k be a vector such that $\|p_k\| \leq \Delta_k$ and*

$$m_k(p_k) - m_k(0) \leq c\left(m_k\left(p_k^C\right) - m_k(0)\right), \tag{2}$$

where $c > 0$ is some constant. Then

$$m_k(p_k) - m_k(0) \leq -\frac{c}{2}\|g_k\| \min\left\{\Delta_k, \frac{\|g_k\|}{\|B_k\|}\right\}. \tag{3}$$

Theorem 1 can be viewed as a "descent lemma" for TR methods. The exact minimizer of the TR subproblem (P_{m_k}), the dogleg method and the 2D subspace minimization method all satisfy (2) and in turn (3) with $c = 1$.

2.2 Convergence to stationary points

Recall the generic TR algorithm.

Algorithm 1 Trust Region

Input: $\hat{\Delta} > 0$ (largest radius), $\Delta_0 \in (0, \hat{\Delta})$ (initial radius), $\eta \in [0, 1/4)$ (acceptance threshold)
for $k = 0, 1, 2, \ldots$
  $p_k = \text{argmin}_{p:\|p\| \leq \Delta_k} m_k(p)$ (or approximate minimizer)
  $\rho_k = \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)}$
  if $\rho_k < \frac{1}{4}$: \\ insufficient progress
    $\Delta_{k+1} = \frac{1}{4}\Delta_k$ \\ reduce radius
  else:
    if $\rho_k > \frac{3}{4}$ and $\|p_k\| = \Delta_k$: \\ sufficient progress, active trust region
      $\Delta_{k+1} = \min\left\{2\Delta_k, \hat{\Delta}\right\}$ \\ increase radius
    else: \\ sufficient progress, inactive trust region
      $\Delta_{k+1} = \Delta$ \\ keep radius
  if $\rho_k > \eta$: \\ sufficient progress
    $x_{k+1} = x_k + p_k$ \\ accept step
  else: \\ insufficient progress
    $x_{k+1} = x_k$ \\ reject step
end for

Two types of global convergence guarantees can be proved depending on the value of η:

1. $\eta = 0$ (always accept if there is any progress). Then $\{g_k\}$ has a limit point at zero, i.e., $\liminf_{k\to\infty} \|g_k\| = 0$.

2. $\eta > 0$ (reject steps with low progress). Then we have the stronger result that $g_k \to 0$.

Below we focus on the first case.

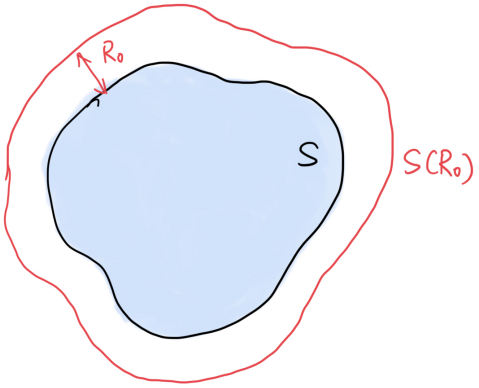
Consider the level set

$$S := \left\{ x \in \mathbb{R}^d \mid f(x) \leq f(x_0) \right\}.$$

Define an open neighborhood of S by

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

Illustration below.



Assumptions:

1. $\forall k : \|B_k\|_2 \leq \beta < \infty$.
2. f is bounded below on S.
3. f is smooth (i.e., has Lipschitz continuous gradient) on $S(R_0)$ for some $R_0 > 0$.

For generality, we allow constant-factor violations of the trust region bound; that is, we only require that

$$\|p_k\| \leq \gamma \Delta_k, \text{ for some } \gamma \geq 1. \tag{4}$$

Theorem 2 (Theorem 4.5 in Nocedal-Wright). *Let $\eta = 0$ in Algorithm 1. Suppose that the assumptions stated above are satisfied, and the step p_k satisfies the sufficient progress condition (3) and the trust region bound (4) for all k. Then*

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

Proof outline:

1. Assume for the purpose of contradiction that there exists $\epsilon > 0, K > 0$ such that for all $k \geq K$: $\|g_k\| \geq \epsilon$.
2. Then there exists $\bar{\Delta} > 0$ such that for all $k \geq K$: $\Delta_k \geq \min\left\{\Delta_K, \bar{\Delta}/4\right\} > 0$.
3. Contradict statement 2 by showing $\lim_{k \to \infty} \Delta_k = 0$. There are two cases:
   (a) $\rho_k \geq \frac{1}{4}$ for some infinite sequence of k, in which case we show that f decreases by some constant times Δ_k, so it must be that $\Delta_k \to 0$, as f is bounded below.

(b) $\rho_k < \frac{1}{4}$ for all k sufficiently large. Then by Algorithm 1, Δ_k is reduced by a factor $1/4$, so $\lim_{k\to\infty} \Delta_k = 0$.

Now the formal proof.

Proof of Theorem 2. Recall that

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

where $f_k = f(x_k)$ and $g_k = \nabla f(x_k)$, and $m_k(0) = f_k$.

Suppose that $\|p_k\| \le R_0$, so that $x_k, x_k + p_k \in S(R_0)$. By assumption, there exists β_1 such that f is β_1-smooth on $S(R_0)$. Also note that

$$|\rho_k - 1| = \left| \frac{f(x_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)} - \frac{m_k(0) - m_k(p_k)}{m_k(0) - m_k(p_k)} \right| = \left| \frac{m_k(p_k) - f(x_k + p_k)}{m_k(0) - m_k(p_k)} \right|, \tag{5}$$

where we use $f(x_k) = m_k(0)$. Let us control the numerator and denominator on the RHS.

By Taylor's Theorem, we have

$$f(x_k + p_k) = f(x_k) + \int_0^1 \langle \nabla f(x_k + t p_k), p_k \rangle \, \mathrm{d}t.$$

We can write $m_k(p_k)$ as

$$m_k(p_k) = f_k + g_k^\top p_k + \frac{1}{2} p_k^\top B_k p_k = f(x_k) + \int_0^1 \langle \nabla f(x_k), p_k \rangle \, \mathrm{d}t + \frac{1}{2} p_k^\top B_k p_k.$$

It follows that

$$\begin{aligned}
|m_k(p_k) - f(x_k + p_k)| &\le \left| \int_0^1 \langle \nabla f(x_k + t p_k) - \nabla f(x_k), p_k \rangle \, \mathrm{d}t \right| + \frac{1}{2} \left| p_k^\top B_k p_k \right| \\
&\le \frac{\beta_1}{2} \|p_k\|^2 + \frac{1}{2} \|B_k\|_2 \|p_k\|^2 && \beta_1\text{-smoothness of } f \\
&\le \frac{\beta_1 + \beta}{2} \|p_k\|^2 && \|B_k\|_2 \le \beta \\
&\le \frac{\beta_1 + \beta}{2} \gamma^2 \Delta_k^2. && \|p_k\| \le \gamma \Delta_k
\end{aligned} \tag{6}$$

For the purpose of contradiction, assume that there exists $\epsilon > 0$ and $K > 0$ such that

$$\|\nabla f(x_k)\| = \|g_k\| \ge \epsilon, \qquad \forall k \ge K.$$

In this case, the sufficient progress condition (3) implies

$$\begin{aligned}
|m_k(0) - m_k(p_k)| &\ge c_1 \|g_k\| \min \left\{ \Delta_k, \frac{\|g_k\|}{\|B_k\|} \right\} \\
&\ge c_1 \epsilon \min \left\{ \Delta_k, \frac{\epsilon}{\|B_k\|} \right\}.
\end{aligned} \tag{7}$$

Combining (5), (6) and (7), we obtain

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

We want to upper bound the RHS. Define:

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

which satisfies

$$\bar{\Delta} \leq \frac{c_1\epsilon}{\gamma^2(\beta_1+\beta)} \leq \frac{\epsilon}{\beta}$$

since $c_1 \leq 1, \gamma \geq 1, \beta_1 \geq 0$. Therefore, for all $\Delta_k \leq \bar{\Delta}$, we have $\min\left\{\Delta_k, \frac{\epsilon}{\beta}\right\} = \Delta_k$, hence

$$|\rho_k - 1| \leq \frac{\gamma^2\Delta_k^2(\beta_1+\beta)}{2c_1\epsilon\Delta_k} \leq \frac{\Delta_k}{2\bar{\Delta}} \leq \frac{1}{2}, \qquad \forall k \geq K,$$

where we use $\frac{\gamma^2(\beta_1+\beta)}{c_1\epsilon} \leq \bar{\Delta}$. It follows that $\rho_k > \frac{1}{4}$. By the workings of Algorithm 1, we have $\Delta_{k+1} \geq \Delta_k$ whenever $\Delta_k \leq \bar{\Delta}$. Thus, Δ_k can decrease (by a factor of $\frac{1}{4}$) only if $\Delta_k \geq \bar{\Delta}$, and therefore we conclude that

$$\Delta_k \geq \min\left\{\Delta_K, \frac{\bar{\Delta}}{4}\right\}, \quad \forall k \geq K. \tag{8}$$

Consider two cases:

(a) Suppose that there exists an infinite subsequence $\mathcal{K}$ of $\{K, K+1, K+2, \ldots\}$ such that $\rho_k \geq \frac{1}{4}, \forall k \in \mathcal{K}$. Then

$$\begin{aligned} f(x_{k+1}) - f(x_k) &= \rho_k\left(m_k(p_k) - m_k(0)\right) \\ &\leq -\frac{1}{4}\left(m_k(0) - m_k(p_k)\right) \\ &\leq -\frac{c_1}{4}\epsilon\min\left\{\Delta_k, \frac{\epsilon}{\beta}\right\}. \end{aligned}$$

Since f is bounded below, it must be $\lim_{k\to\infty, k\in\mathcal{K}} \Delta_k = 0$, which is a contradiction to (8).

(b) Suppose that no such $\mathcal{K}$ exists. Therefore $\rho_k < \frac{1}{4}$ must hold for all sufficiently large k. But from Algorithm 1, this means that $\Delta_{k+1} = \frac{1}{4}\Delta_k$, which implies $\lim_{k\to\infty}\Delta_k = 0$, again contradicting (8).

We conclude that the original assertion $\|g_k\| \geq \epsilon, \forall k \geq K$ must be false, hence $\liminf_{k\to\infty}\|g_k\| = 0$. □

Option reading:

- When $\eta > 0$, we have $g_k \to 0$. See Theorem 4.6 in Nocedal-Wright.
- For small-scale problems, we may solve the TR subproblem $\min_{\|p\|\leq\Delta_k} m_k(p)$ more accurately using iterative methods. See Section 4.3 in Nocedal-Wright.

3 Local convergence of TR-Newton method

The results discussed so far hold for general B_k. We now specialize to TR methods that use the true Hessian $B_k = \nabla^2 f(x_k)$ for all sufficiently large k. (We refer to these methods as TR-Newton.) In this case, we expect that the TR bound $\|p_k\| \le \Delta_k$ becomes inactive near the minimizer of f and thus approximate solution to the TR subproblem (P_{m_k}) becomes similar to the Newton step $p_k^{\mathrm{N}} := -\nabla^2 f(x_k)^{-1} \nabla f(x_k)$.

The theorem below establishes superlinear local convergence of TR-Newton.

Theorem 3 (Theorem 4.9 in Nocedal-Wright). *Let f be twice continuously differentiable (with β_1-Lipschitz gradients and L-Lipschitz Hessians) in a neighborhood of a local minimizer x^* satisfying $\nabla f(x^*) = 0, \nabla^2 f(x^*) \succ 0$. Suppose that*

1. *$\{x_k\}$ converges to x^*;*
2. *for all k sufficiently large, the TR algorithm with $B_k = \nabla^2 f(x_k)$ chooses p_k such that*
   (a) *the sufficient progress condition (3) holds, and*
   (b) *p_k is asymptotically similar to $p_k^N = -\nabla^2 f(x_k)^{-1} g_k$ whenever $\|p_k^N\| \le \frac{\Delta_k}{2}$, i.e.,*

$$\left\|p_k - p_k^N\right\| = o(\left\|p_k^N\right\|). \tag{9}$$

Then the TR bound becomes inactive for all sufficiently large k and the convergence of $\{x_k\}$ to x^* is superlinear.

Proof outline: Show that for all sufficiently large k, ρ_k is close to 1, so Algorithm 1 with keep the TR radius Δ_k large. Consequently, we have $\|p_k^{\mathrm{N}}\| \le \frac{\Delta_k}{2}$, so (9) holds, in which case we can invoke the generic result in Lecture 21, Theorem 2 to establish the superlienar convergence.

Proof of Theorem 3. We want bound

$$|\rho_k - 1| = \left|\frac{f(x_k) - f(x_k + p_k) - (m_k(0) - m_k(p_k))}{m_k(0) - m_k(p_k)}\right|. \tag{10}$$

First consider the numerator. Suppose that k is large enough.

- If $\|p_k^{\mathrm{N}}\| \le \frac{\Delta_k}{2}$,then (9) applies, hence $\|p_k\| \le \|p_k - p_k^{\mathrm{N}}\| + \|p_k^{\mathrm{N}}\| \le 2\|p_k^{\mathrm{N}}\|$.
- If $\|p_k^{\mathrm{N}}\| > \frac{\Delta_k}{2}$, then $\|p_k\| \le \Delta_k < 2\|p_k^{\mathrm{N}}\|$.

In both cases, we have

$$\|p_k\| \le 2\left\|p_k^{\mathrm{N}}\right\| = 2\left\|\nabla^2 f(x_k)^{-1} g_k\right\|_2 \le 2\left\|\nabla^2 f(x_k)^{-1}\right\|_2 \|g_k\|_2.$$

Hence $\|g_k\| \ge \frac{\|p_k\|_2}{2\|\nabla^2 f(x_k)^{-1}\|_2}$. Plugging this bound into the sufficient progress condition (3), we obtain

$$\begin{aligned}
& m_k(p_k) - m_k(0) \\
\le & -c_1 \|g_k\| \min\left\{\Delta_k, \frac{\|g_k\|}{\|\nabla^2 f(x_k)\|}\right\} \\
\le & -c_1 \frac{\|p_k\|}{2\|\nabla^2 f(x_k)^{-1}\|} \min\left\{\|p_k\|, \frac{\|p_k\|}{2\|\nabla^2 f(x_k)^{-1}\|\,\|\nabla^2 f(x_k)\|}\right\} && \because \Delta_k \ge \|p_k\| \\
= & -\frac{c_1 \|p_k\|^2}{4\|\nabla^2 f(x_k)^{-1}\|_2^2 \|\nabla^2 f(x_k)\|_2}. && \because \left\|\nabla^2 f(x_k)^{-1}\right\| \|\nabla^2 f(x_k)\| \ge 1
\end{aligned}$$

When x_k is sufficiently close to x^*, we have $\nabla^2 f(x_k) \approx \nabla^2 f(x^*)$. Hence (by continuity of Hessian)

$$\frac{c_1}{4\left\|\nabla^2 f(x_k)^{-1}\right\|^2 \left\|\nabla^2 f(x_k)\right\|} \geq \frac{c_1}{8\left\|\nabla^2 f(x^*)^{-1}\right\|^2 \left\|\nabla^2 f(x^*)\right\|} =: c_3.$$

We conclude that

$$m_k(p_k) - m_k(0) \leq -c_3 \left\|p_k\right\|^2 \qquad \text{for all sufficiently large } k.$$

We next bound the denominator of the RHS of (10):

$$\begin{aligned}
&\left|f(x_k) - f(x_k + p_k) - (m_k(0) - m_k(p_k))\right| \\
=&\left|-g_k^\top p_k - \frac{1}{2}\int_0^1 p_k^\top \nabla^2 f(x_k + tp_k) p_k \mathrm{d}t + g_k^\top p_k + \int_0^1 p_k^\top \nabla^2 f(x_k) p_k \mathrm{d}t\right| && \text{Taylor's Theorem} \\
=&\left|\frac{1}{2}\int_0^1 p_k^\top \left[\nabla^2 f(x_k + tp_k) - \nabla^2 f(x_k)\right] p_k \mathrm{d}t\right| \\
\leq&\frac{1}{2}\int_0^1 \underbrace{\left\|\nabla^2 f(x_k + tp_k) - \nabla^2 f(x_k)\right\|_2}_{\leq Lt\|p_k\|} \left\|p_k\right\|_2^2 \mathrm{d}t && \text{Lipschitzness of } \nabla^2 f \\
\leq&\frac{L}{4}\left\|p_k\right\|^3.
\end{aligned}$$

Combining the last two bounds with (10), we obtain

$$\begin{aligned}
&|\rho_k - 1| \leq \frac{(L/4)\left\|p_k\right\|^3}{c_3\left\|p_k\right\|^2} = \frac{L}{4c_3}\left\|p_k\right\| \leq \frac{L\Delta_k}{4c_3} \\
\implies &\rho_k \geq 1 - \frac{L\Delta_k}{4c_3}.
\end{aligned}$$

Therefore, Algorithm 1 will keep Δ_k bounded away from zero. On the other hand, as $\{x_k\} \to x^*$, we must have $\left\|p_k^{\mathrm{N}}\right\| = \left\|\nabla^2 f(x_k)^{-1} g_k\right\| \to 0$ since $g_k \to 0$. Thus, we have $\left\|p_k^{\mathrm{N}}\right\| \leq \frac{\Delta_k}{2}$ and thus (9) holds. In this case, we have

$$\left\|p_k\right\| \leq \left\|p_k - p_k^{\mathrm{N}}\right\| + \left\|p_k^{\mathrm{N}}\right\| \leq (1 + o(1))\left\|p_k^{\mathrm{N}}\right\| < \Delta_k,$$

so the TR constraint is eventually inactive. Moreover, Theorem 2 and Claim 1 in Lecture 21 ensures that when (9) holds, we have superlinear convergence. □

Remark 1. In fact, "reasonable" TR methods with $B_k = \nabla^2 f(x_k)$ will eventually use $p_k = p_k^{\mathrm{N}}$ and therefore converge quadratically. In particular, this holds for the dogleg method, 2D subspace minimization and other approximate algorithm for solving the TR-subproblem $\min_{\|p\| \leq \Delta_k} m_k(p)$.