Numerical Optimization

Lecture Notes #9 — Trust-Region Methods
Global Convergence and Enhancements

Fall 2019

Outline

1. Recap & Introduction
   - Recap: Iterative "Nearly Exact" Solution of the Subproblem
   - Quick Lookahead
2. Global Convergence
   - Tool #1 — A Lemma: The Cauchy Point
   - Tool #2 — A Theorem
   - Recall: The Trust Region Algorithm
3. Global Convergence...
   - Convergence to Stationary Points
4. Enhancements
   - Scaling

Recap: — Iterative "Nearly Exact" Solution of the Subproblem

Last time we looked at **nearly exact solution** of the subproblem

$$\min_{\bar{\mathbf{p}} \in T_k} m_k(\bar{\mathbf{p}}) = \min_{\bar{\mathbf{p}} \in T_k} f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2} \bar{\mathbf{p}}^T B_k \bar{\mathbf{p}}$$

This approach is viable for problems with few degrees of freedom, *e.g.* $T_k \subseteq \mathbb{R}^n$, n "small." Where "small" means that the **unitary diagonalization** $Q_k \Lambda_k Q_k^T = B_k$ is computable in a "reasonable" amount of time.

From a theoretical characterization of the exact problem, we derived an algorithm which finds a nearly exact solution at a cost per iteration approximately **three** times that of dogleg and 2D-subspace minimization.

The scheme was based on a 1-D Newton iteration (with some clever tricks), and some careful analysis of special (hard) cases.

On Today's Menu

We wrap up the first pass of Trust Region methods —

- We briefly discuss global convergence properties for trust region methods.
- We look at some theorems, but leave the proofs as "exercises."
- For second order $(B_k \neq \nabla^2 f(\bar{\mathbf{x}}_k))$ models we can show convergence to a stationary point.
- For trust-region Newton methods $(B_k = \nabla^2 f(\bar{\mathbf{x}}_k))$ models we can show convergence to a point where the second order necessary conditions hold.
- We look at modifications for poorly scaled problems, as well as the use of non-spherical trust regions.

Theorem (Second Order Necessary Conditions)

If $\bar{\mathbf{x}}^*$ is a local minimizer of f and $\nabla^2 f$ is continuous in an open neighborhood of $\bar{\mathbf{x}}^*$, then $\nabla f(\bar{\mathbf{x}}^*) = 0$ and $\nabla^2 f(\bar{\mathbf{x}}^*)$ is positive semi-definite.

Global Convergence: Tool #1 — A Lemma

Recall: The trust-region subproblem is

$$\bar{\mathbf{p}}_k = \underset{\|\bar{\mathbf{p}}\| \leq \Delta_k}{\arg\min}\, m_k(\bar{\mathbf{p}}) = \underset{\|\bar{\mathbf{p}}\| \leq \Delta_k}{\arg\min}\, f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2} \bar{\mathbf{p}}^T B_k \bar{\mathbf{p}}.$$

The following lemma gives us a lower bound for the decrease in the model at the Cauchy point:

Lemma (Cauchy point descent)

The Cauchy point $\bar{\mathbf{p}}_k^c$ satisfies

$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k^c) \geq \frac{1}{2} \|\nabla f(\bar{\mathbf{x}}_k)\| \min \left[\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|} \right].$$

Proof of Lemma

We recall the explicit expressions for the Cauchy point (from lecture 7)

$$\begin{cases} \bar{\mathbf{p}}_k^c = -\tau_k \dfrac{\Delta_k}{\|\nabla f(\bar{\mathbf{x}}_k)\|} \nabla f(\bar{\mathbf{x}}_k) \\ \text{where} \\ \tau_k = \begin{cases} 1 & \text{if } \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) \leq 0 \\ \min\left(1, \dfrac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)}\right) & \text{otherwise} \end{cases} \end{cases}$$

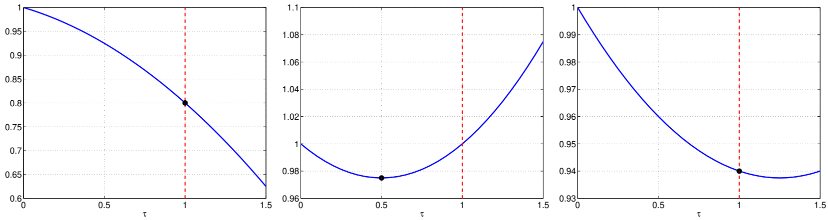


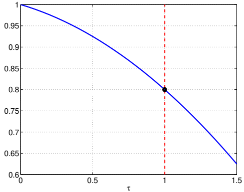
Figure: The three possible scenarios for selection of τ.

Proof of Lemma

Case#1

Case#1 $(\nabla f(\bar{\mathbf{x}}_k) B_k \nabla f(\bar{\mathbf{x}}) \le 0)$:

In this scenario $m_k(\bar{\mathbf{p}}_k^c) - m_k(\bar{\mathbf{0}}) =$



$$
\begin{aligned}
&= m_k\left(-\Delta_k \frac{\nabla f(\bar{\mathbf{x}}_k)}{\|\nabla f(\bar{\mathbf{x}}_k)\|}\right) - m_k(\bar{\mathbf{0}}) \\
&= -\Delta_k \|\nabla f(\bar{\mathbf{x}}_k)\| + \frac{1}{2} \frac{\Delta_k^2}{\|\nabla f(\bar{\mathbf{x}}_k)\|^2} \underbrace{\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)}_{\le 0} \\
&\le -\Delta_k \|\nabla f(\bar{\mathbf{x}}_k)\| \\
&\le -\|\nabla f(\bar{\mathbf{x}}_k)\| \min\left(\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right)
\end{aligned}
$$

Hence,

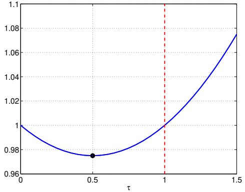
$$
m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k^c) \ge \|\nabla f(\bar{\mathbf{x}}_k)\| \min\left(\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right) \ge \frac{1}{2} \|\nabla f(\bar{\mathbf{x}}_k)\| \min\left(\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right)
$$

Proof of Lemma — Case#2

Case#2 $\left(\nabla f(\bar{\mathbf{x}}_k) B_k \nabla f(\bar{\mathbf{x}}) > 0, \text{ and } \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)} \leq 1\right)$:

In this scenario the Cauchy point is in the interior of the trust region, and $m_k(\bar{\mathbf{p}}_k^c) - m_k(\bar{\mathbf{0}}) =$

$$
\begin{aligned}
&= -\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^4}{\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)} + \frac{1}{2}\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^4}{(\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k))^2} \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) \\
&= -\frac{1}{2}\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^4}{\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)} \\
&\leq -\frac{1}{2}\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^4}{\|B_k\|\,\|\nabla f(\bar{\mathbf{x}}_k)\|^2} = -\frac{1}{2}\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^2}{\|B_k\|} \\
&\leq -\frac{1}{2}\|\nabla f(\bar{\mathbf{x}}_k)\| \min\left(\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right)
\end{aligned}
$$



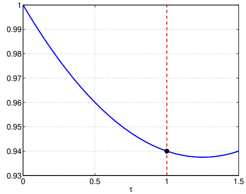
Use the minus sign to flip the inequality, and we're there!

Proof of Lemma — Case#3

Case#3 $\left(\nabla f(\bar{\mathbf{x}}_k) B_k \nabla f(\bar{\mathbf{x}}) > 0, \text{ and } \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)} > 1\right)$:

We note that in this scenario $\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) < \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k}$, and $m_k(\bar{\mathbf{p}}_k^c) - m_k(\bar{\mathbf{0}}) =$

$$
\begin{aligned}
&= -\frac{\Delta_k}{\|\nabla f(\bar{\mathbf{x}}_k)\|}\|\nabla f(\bar{\mathbf{x}}_k)\|^2 + \frac{1}{2}\frac{\Delta_k^2}{\|\nabla f(\bar{\mathbf{x}}_k)\|^2}\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) \\
&\leq -\Delta_k \|\nabla f(\bar{\mathbf{x}}_k)\| + \frac{1}{2}\frac{\Delta_k^2}{\|\nabla f(\bar{\mathbf{x}}_k)\|^2}\frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k} \\
&= -\frac{1}{2}\Delta_k \|\nabla f(\bar{\mathbf{x}}_k)\| \\
&\leq -\frac{1}{2}\|\nabla f(\bar{\mathbf{x}}_k)\| \min\left(\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right)
\end{aligned}
$$



Use the minus sign to flip the inequality, and we're there!

Global Convergence: Tool #2 — A Theorem

Theorem

Let $\bar{\mathbf{p}}_k$ *be any vector,* $\|\bar{\mathbf{p}}_k\| \leq \Delta_k$*, such that*

$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k) \geq c_2(m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k^c))$$

then

$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k) \geq \frac{c_2}{2}\|\nabla f(\bar{\mathbf{x}}_k)\| \min\left[\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right].$$

Both the dogleg, and 2-D subspace minimization algorithms (as well as Steihaug's algorithm) fall into this category, with $c_2 = 1$, since they all produce $\bar{\mathbf{p}}_k$ which give at least as much descent as the Cauchy point, *i.e.* $m_k(\bar{\mathbf{p}}_k) \leq m_k(\bar{\mathbf{p}}_k^c)$.

We are going to use this result to show convergence for the trust region algorithm (see next slide).

The Trust Region Algorithm

Algorithm: Trust Region

Set $k = 1$, $\widehat{\Delta} > 0$, $\Delta_0 \in (0, \widehat{\Delta})$, and $\eta \in [0, \frac{1}{4}]$
While optimality condition not satisfied
  Get $\bar{\mathbf{p}}_k$ (approximate solution)
  Evaluate ρ_k
  if $\rho_k < \frac{1}{4}$
    $\Delta_{k+1} = \frac{1}{4}\Delta_k$
  else
    if $\rho_k > \frac{3}{4}$ and $\|\bar{\mathbf{p}}_k\| = \Delta_k$
      $\Delta_{k+1} = \min(2\Delta_k, \widehat{\Delta})$
    else
      $\Delta_{k+1} = \Delta_k$
    endif
  endif
  if $\rho_k > \eta$
    $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k + \bar{\mathbf{p}}_k$
  else
    $\bar{\mathbf{x}}_{k+1} = \bar{\mathbf{x}}_k$
  endif
  $k = k + 1$
End-While

Convergence to Stationary Points

Case $\eta = 0$

accept any step which produces descent in f — we can show that the sequence of gradients $\{\nabla f(\bar{\mathbf{x}}_k)\}$ has a **limit point** at zero.

Case $\eta > 0$

accept a step only if the decrease in f is at least some fixed fraction of the predicted decrease — we can show the stronger result $\{\nabla f(\bar{\mathbf{x}}_k)\} \to \bar{\mathbf{0}}$.

In order for the proof(s) to work, we must assume that the model Hessians B_k are uniformly bounded, *i.e.* $\|B_k\| \leq \beta$, and that f is bounded below on the levelset $\{\bar{\mathbf{x}} \in \mathbb{R}^n : f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$.

The trust-region bound can be relaxed so that the results hold as long as the solution to the subproblems satisfy

$$\|\bar{\mathbf{p}}_k\| \leq \gamma \Delta_k, \quad \text{for some constant } \gamma \geq 1.$$

Convergence to Stationary Points: $\eta = 0$

Theorem

Let $\eta = 0$ in the trust region algorithm. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is continuously differentiable and bounded below on the bounded set $\{\bar{\mathbf{x}} \in \mathbb{R}^n \,:\, f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$, and that all approximate solutions to the trust-region subproblem satisfy the inequalities

$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k) \geq c_1 \|\nabla f(\bar{\mathbf{x}}_k)\| \, \min\left[\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right],$$

and

$$\|\bar{\mathbf{p}}_k\| \leq \gamma \Delta_k,$$

for some positive constants c_1 and γ. **Then** *we have*

$$\liminf_{k \to \infty} \|\nabla f(\bar{\mathbf{x}}_k)\| = 0.$$

Convergence to Stationary Points: $\eta > 0$

Theorem

Let $\eta \in \left(0, \frac{1}{4}\right)$ in the trust region algorithm. Suppose that $\|B_k\| \leq \beta$ for some constant β, that f is Lipschitz continuously differentiable and bounded below on the bounded set $\{\bar{\mathbf{x}} \in \mathbb{R}^n \,:\, f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$, and that all approximate solutions to the trust-region subproblem satisfy the inequalities

$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k) \geq c_1 \|\nabla f(\bar{\mathbf{x}}_k)\| \min\left[\Delta_k, \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|}{\|B_k\|}\right].$$

and

$$\|\bar{\mathbf{p}}_k\| \leq \gamma \Delta_k$$

for some positive constants c_1 and γ. **Then** *we have*

$$\lim_{k \to \infty} \nabla f(\bar{\mathbf{x}}_k) = \bar{\mathbf{0}}.$$

Proofs: Convergence to Stationary Points

The complete proofs are in $\texttt{NW}^{1st}$ pp.90–91, and pp.92–93; or $\texttt{NW}^{2nd}$ pp.80–82, and pp.82–83.

The proofs are based on manipulation of ρ — the ratio of actual (objective) reduction and predicted (model) reduction; Taylor's theorem; then deriving a contradiction from the supposition $\|\nabla f(\bar{\mathbf{x}}_k)\| \geq \epsilon$ using careful selection of scalings and bounds for Δ_k.

Definition (lim sup and lim inf)

Let $\{s_n\}$ be a sequence of real numbers. Let E be the set of values x so that $s_{n_k} \to x$ for some subsequence $\{s_{n_k}\}$. This set E contains all sub-sequential limits, plus possibly $\pm\infty$; let

$$s^* = \sup E, \quad s_* = \inf E$$

The values s^* and s_* are the upper and lower limits of $\{s_n\}$, and we use the notation

$$\limsup_{n\to\infty} s_n = s^*, \quad \liminf_{n\to\infty} s_n = s_*$$

Convergence: Iterative "Nearly Exact" Solutions $\bar{\mathbf{p}}_k^*$, for Trust-Region Newton

Theorem (NW[2nd] p.92, proof in Moré & Sorensen (1983))

Let $\eta \in \left(0, \frac{1}{4}\right)$ *in the algorithm on slide 11, let* $B_k = \nabla^2 f(\bar{\mathbf{x}}_k)$, *and suppose that* $\bar{\mathbf{p}}_k$ *at each iteration satisfy*

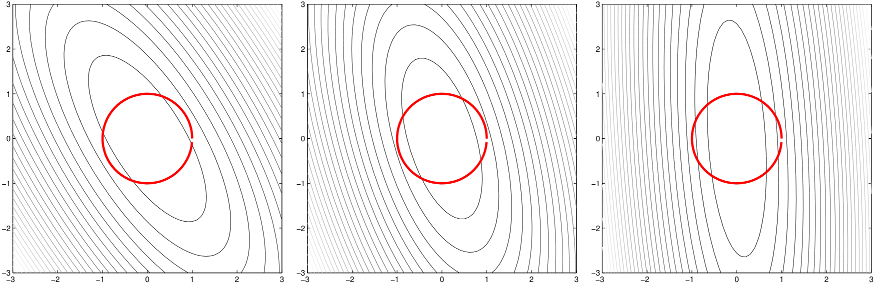
$$m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k) \geq c_1(m_k(\bar{\mathbf{0}}) - m_k(\bar{\mathbf{p}}_k^*)),$$

and $\|\bar{\mathbf{p}}_k\| \leq \gamma \Delta_k$, *for some positive constant* γ, *and* $c_1 \in (0, 1]$. **Then**

$$\lim_{k \to \infty} \|\nabla \mathbf{f}(\bar{\mathbf{x}}_k)\| = \mathbf{0}.$$

If, in addition, the set $\{\bar{\mathbf{x}} \in \mathbb{R}^n : f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$ *is compact, then* **either** *the algorithm terminates at a point* $\bar{\mathbf{x}}_k$ *at which the second order necessary conditions for a local minimum hold,* **or** $\{\bar{\mathbf{x}}_k\}$ *has a limit point* $\bar{\mathbf{x}}^* \in \{\bar{\mathbf{x}} \in \mathbb{R}^n : f(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}_0)\}$ *at which the conditions hold.*

Enhancement: Scaling — The Problem



As we have seen before (in the context of steepest descent / line-search), **scaling** (ill-conditioning) can cause problems. — If the objective is more sensitive to changes in one variable than other, the contour lines stretch out to be narrow ellipses (in 2D).

Clearly, a circular trust-region may be quite limiting in this scenario. — The radius is limited by the sensitive variable.

Enhancement: Scaling — The Solution

The solution to the problem of poor scaling is to use **elliptical** trust regions. We define a diagonal scaling matrix

$$D = \mathrm{diag}(d_1, d_2, \dots, d_n), \quad d_i > 0.$$

Then, the constraint $\|D\bar{\mathbf{p}}\| \leq \Delta$ defines an elliptical trust region, and we get the following scaled trust-region subproblem:

$$\min_{\bar{\mathbf{p}} \in \mathbb{R}^n \,:\, \|D\bar{\mathbf{p}}\| \leq \Delta_k} f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2} \bar{\mathbf{p}}^T B_k \bar{\mathbf{p}}.$$

The scaling matrix can be built using information about the gradient $\nabla f(\bar{\mathbf{x}}_k)$ and the Hessian $\nabla^2 f(\bar{\mathbf{x}}_k)$ along the solution path. — We can allow $D = D_k$ to change from iteration to iteration.

All our analysis/algorithms still work with scaling added — but we get factors of D^{-2}, D^{-1}, D, and D^2 in our expressions.

Feature: Non-Euclidean Trust Regions

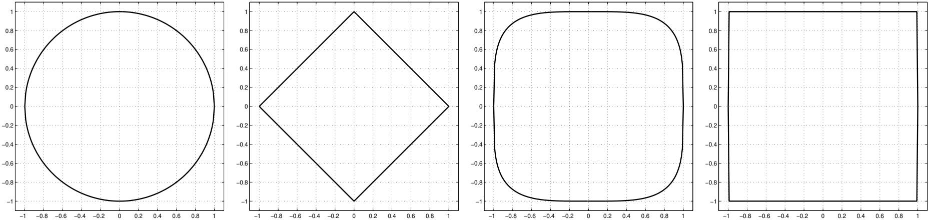


Figure: Illustration of (unscaled) trust region boundaries for, from left-to-right: $\|\bar{\mathbf{p}}\|_2 \leq \Delta_k$, $\|\bar{\mathbf{p}}\|_1 \leq \Delta_k$, $\|\bar{\mathbf{p}}\|_4 \leq \Delta_k$, and $\|\bar{\mathbf{p}}\|_\infty \leq \Delta_k$.

Most of the time using trust regions based on norms with $q \neq 2$:

$$\|\bar{\mathbf{p}}\|_q \leq \Delta_k \text{ (unscaled)}, \quad \|D\bar{\mathbf{p}}\|_q \leq \Delta_k \text{ (scaled)}$$

cause us a giant head-ache. There are however some situations when such regions come in handy...

Feature: Non-Euclidean Trust Regions

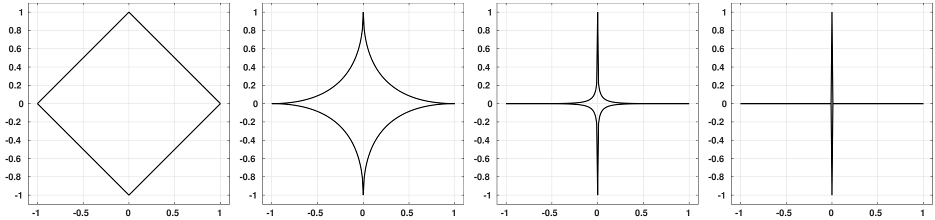


Figure: Illustration of (unscaled) trust region boundaries for, from left-to-right: $\|\bar{\mathbf{p}}\|_1 \leq \Delta_k$, $\|\bar{\mathbf{p}}\|_{\frac{1}{2}} \leq \Delta_k$, $\|\bar{\mathbf{p}}\|_{\frac{1}{4}} \leq \Delta_k$, and $\|\bar{\mathbf{p}}\|_{\frac{1}{8}} \leq \Delta_k$.

Using $q < 1$ leads to non-convex trust regions, which may be a bit of a pain?!?

This may, however, be useful/necessary for non-convex optimization problems.

Feature: Non-Euclidean Trust Regions

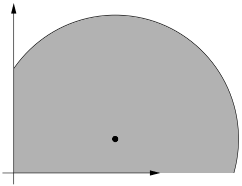
For **constrained** problems, *e.g.*

$$\min_{\bar{\mathbf{x}} \in \mathbb{R}^n} f(\bar{\mathbf{x}}), \quad \text{subject to} \quad x_i \geq 0, \;\; i = 1, 2, \ldots, n$$

the trust-region subproblem may be

$$\min_{\bar{\mathbf{p}} \in \mathbb{R}^n} m_k(\bar{\mathbf{p}}), \quad \text{subject to} \quad \bar{\mathbf{x}}_k + \bar{\mathbf{p}} \geq 0, \; \text{(component-wise)}, \; \|\bar{\mathbf{p}}\| \leq \Delta_k$$

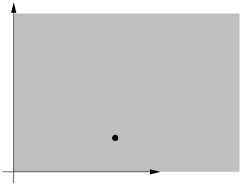
This trust region is the intersection of the disk centered at $\bar{\mathbf{x}}_k$ and the first quadrant. It could look like this:



Feature: Non-Euclidean Trust Regions

Such a region is hard to describe, and hard to work with.

If, instead, we work with the $\|\cdot\|_\infty$-norm, the trust region is the intersection of the square with sides Δ_k centered at $\bar{\mathbf{x}}_k$ and the first quadrant:



Much easier to work with...

Index

Reference(s):

MS-1983 J.J. Moré and D.C. Sorensen, *Computing a Trust Region Step*, SIAM Journal on Scientific and Statistical Computing, 4 (1983), pp. 553 – 572.