Numerical Optimization

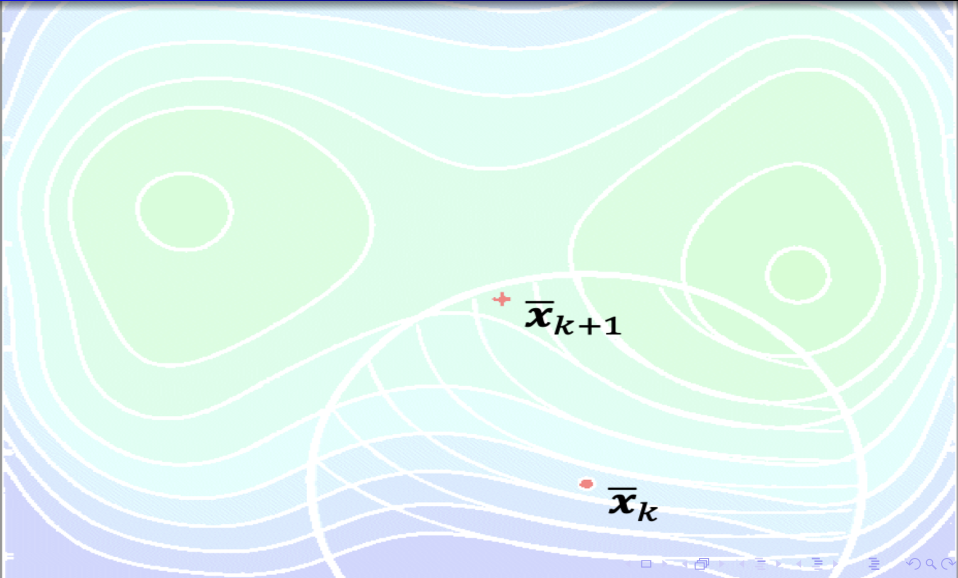
Lecture Notes #7

Trust-Region Methods: Introduction / Cauchy Point

Fall 2024

Outline

Introduction – Trust Region Methods



Key ideas: Trust Region Methods

The Idea:

- Build a, usually quadratic, model around the current point $\bar{\mathbf{x}}_k$.
- Next, define a region in which we trust the model to be a good representation of the objective f.
- Let the next iterate $\bar{\mathbf{x}}^*_{k+1}$ be the (approximate) optimizer of the **model** in the **"trust region."** simultaneously.
- If the new point $\bar{\mathbf{x}}^*_{k+1}$ is not acceptable, we reduce the size of the trust region, and repeat.

Trust Region Methods — The Quadratic Model

The **"model"** is based on the Taylor expansion of the objective f at the current point $\bar{\mathbf{x}}_k$ —

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

where B_k is a symmetric matrix.

If $B_k \neq \nabla^2 f(\bar{\mathbf{x}}_k)$ the **error** in the model **is quadratic** in $\bar{\mathbf{p}}$, *i.e.*

$$\|m_k(\bar{\mathbf{p}}) - f(\bar{\mathbf{x}}_k + \bar{\mathbf{p}})\| \sim \mathcal{O}\left(\|\bar{\mathbf{p}}\|^2\right),$$

If $B_k = \nabla^2 f(\bar{\mathbf{x}}_k)$ the model agrees with the first three terms of the expansion and the **error** in the model is **cubic**

$$\|m_k(\bar{\mathbf{p}}) - f(\bar{\mathbf{x}}_k + \bar{\mathbf{p}})\| \sim \mathcal{O}\left(\|\bar{\mathbf{p}}\|^3\right).$$

Trust Region Methods — The Quadratic Model

When the first three terms of the quadratic model agrees with the Taylor expansion, *i.e.* $B_k = \nabla^2 f(\bar{\mathbf{x}}_k)$, the algorithm is called **the trust-region Newton Method**.

The locally constrained **trust region problem** is

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

where T_k is the trust region.

Note: If B_k is positive definite, and $\bar{\mathbf{p}}_k^B = -B_k^{-1} \nabla f(\bar{\mathbf{x}}_k) \in T_k$, then the **full step** is allowed.

Illustration: The Quadratic Model

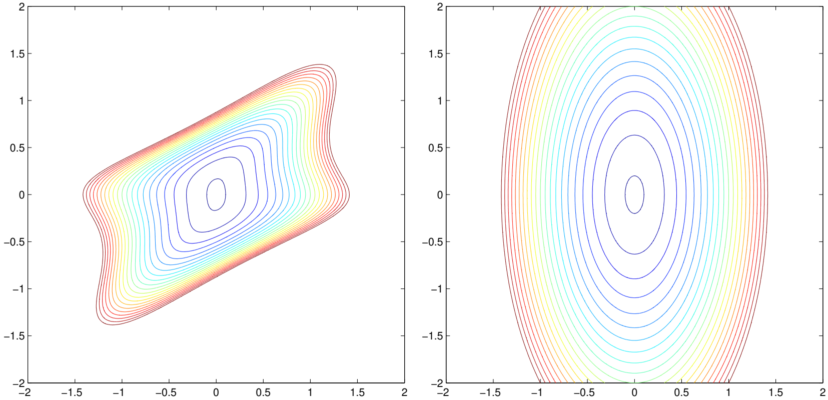


Figure: The picture to the left shows the contour lines of the objective $f(\bar{\mathbf{x}}) = x_1^2 + x_2^2/4 + 4(x_1 - x_2)^2 \cdot \sin^2(x_2)$ and the picture to the right shows the same contour lines for the model $m_k(\bar{\mathbf{p}})$ whose first three terms agree with the Taylor expansion of the objective.

Illustration: The Quadratic Model

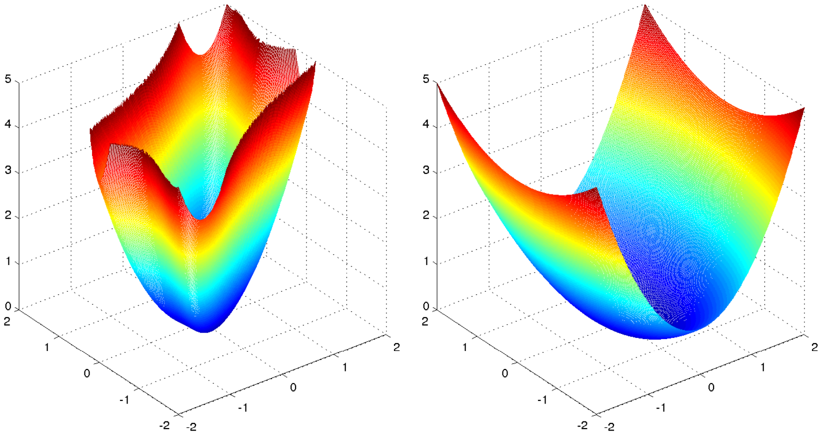


Figure: The picture to the left mesh plot of the objective $f(\bar{\mathbf{x}}) = x_1^2 + x_2^2/4 + 4(x_1 - x_2)^2 \cdot \sin^2(x_2)$ and the picture to the right shows the mesh plot for the model $m_k(\bar{\mathbf{p}})$ whose first three terms agree with the Taylor expansion of the objective.

Trust Region Methods — The Trust Region

Usually, the **Trust Region** T_k is defined by its radius Δ_k:

$$T_k = \{\bar{\mathbf{x}} \in \mathbb{R}^n \,:\, \|\bar{\mathbf{x}}\| \leq \Delta_k\}.$$

Note: If B_k is positive definite, and $\bar{\mathbf{p}}_k^B = -B_k^{-1}\nabla f(\bar{\mathbf{x}}_k) \in T_k$, (i.e. $\|B_k^{-1}\nabla f(\bar{\mathbf{x}}_k)\| \leq \Delta_k$) then the **full step** is allowed.

Note: If $\|B_k^{-1}\nabla f(\bar{\mathbf{x}}_k)\| > \Delta_k$ then the full step is not allowed, and we must find the optimal (approximate solution) to the (locally) **constrained problem**

$$\min_{\|\bar{\mathbf{p}}\| \leq \Delta_k} \left[f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2}\bar{\mathbf{p}}^T B_k \bar{\mathbf{p}} \right].$$

The Base-Line Trust Region Algorithm – How to choose Δ_k

First, we define a ratio measuring the success of a step —

Definition

Given a step $\bar{\mathbf{p}}_k$ we define the ratio

$$\rho_k = \frac{\text{actual reduction}}{\text{predicted reduction}} = \frac{f(\bar{\mathbf{x}}_k) - f(\bar{\mathbf{x}}_k + \bar{\mathbf{p}}_k)}{m_k(0) - m_k(\bar{\mathbf{p}}_k)}$$

The predicted reduction is always non-negative (the step $\bar{\mathbf{p}}_k = 0$ is part of the trust region). Thus if $\rho_k < 0$ the step must be rejected (since $f(\bar{\mathbf{x}}_k + \bar{\mathbf{p}}_k) > f(\bar{\mathbf{x}}_k)$).

The Base-Line Trust Region Algorithm

If $\rho_k < 0$	We shrink the size of the trust region.
If $\rho_k \approx 0$	Then we shrink the size of the trust region.
If $\rho_k \approx 1$	Then the model m_k is in good agreement with the objective f; in this case it is (probably) safe to expand the trust region for the next iteration.

Otherwise we keep the size of the trust region.

The Base-Line Trust Region Algorithm (Algorithm for choosing Δ_k) 3 of 3

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

Trust Region Algorithm: Missing Parts - (Approximating $\bar{\mathbf{p}}_k$)

Clearly, in order to make use of this "algorithm" we must turn our attention to the solution of

$$\min_{\|\bar{\mathbf{p}}\|\leq\Delta_k} \left[f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2}\bar{\mathbf{p}}^T B_k \bar{\mathbf{p}} \right]. \qquad (\textbf{Get } \bar{\mathbf{p}}_k)$$

We look at the easiest approximation:

— the **Cauchy point**, the minimizer of $m_k(\bar{\mathbf{p}})$ in the steepest descent direction.

Then we study three improvements to the Cauchy point:

— **Dogleg method**; used when B_k is positive definite.

— **2-D Subspace Minimization**; can be used when B_k is indefinite.

— **Steihaug's Method**; appropriate when $B_k = \nabla^2 f(\bar{\mathbf{x}}_k)$ and this matrix is large and sparse (most entries are zeros.)

The Cauchy Point

For global convergence we can be quite sloppy in the minimization of the model $m_k(\bar{\mathbf{p}})$ — all we must require is **sufficient reduction** in the model. This is quantified in terms of the Cauchy point $\bar{\mathbf{p}}_k^c$ —

Algorithm: Cauchy Point Calculation

Find the minimizer for the linear model $l_k(\bar{\mathbf{p}}) = f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k)$

$$\bar{\mathbf{p}}_k^s = \underset{\|\bar{\mathbf{p}}\| \leq \Delta_k}{\arg\min} \left[f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) \right].$$

Let $\tau_k > 0$ be the scalar that minimizes $m_k(\tau \bar{\mathbf{p}}_k^s)$ subject to satisfying the trust-region constraint, *i.e.*

$$\tau_k = \underset{\tau > 0}{\arg\min}\, m_k(\tau \bar{\mathbf{p}}_k^s), \quad \text{such that }, \; \|\tau \bar{\mathbf{p}}_k^s\| \leq \Delta_k.$$

Let $\bar{\mathbf{p}}_k^c = \tau_k \bar{\mathbf{p}}_k^s$. This is the Cauchy point.

The Cauchy Point — Explicit Expressions

We can write down some of the quantities explicitly, *e.g.*

$$\bar{\mathbf{p}}_k^s = -\Delta_k \frac{\nabla f(\bar{\mathbf{x}}_k)}{\|\nabla f(\bar{\mathbf{x}}_k)\|},$$

is the full step to the trust-region boundary.

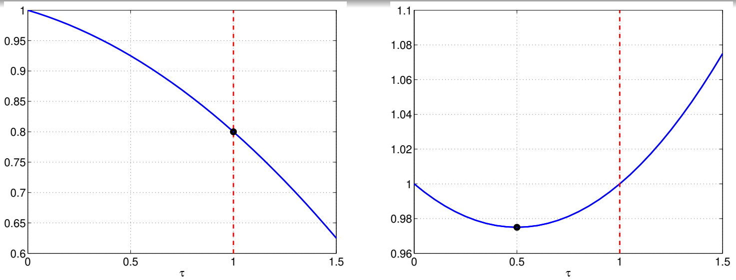
Case: $\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) \leq 0$

$m_k(\tau \bar{\mathbf{p}}_k^s)$ decreases monotonically with τ, whenever $\nabla f(\bar{\mathbf{x}}_k) \neq 0$. Hence, τ_k is the largest τ which keeps satisfing the trust-region condition; by construction of $\bar{\mathbf{p}}_k^s$, this means $\tau_k = 1$.

Case: $\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k) > 0$

$m_k(\tau \bar{\mathbf{p}}_k^s)$ is a convex quadratic in τ; hence τ_k is the smaller of the minimizer of the quadratic, or 1.

The Cauchy Point — Explicit Expressions



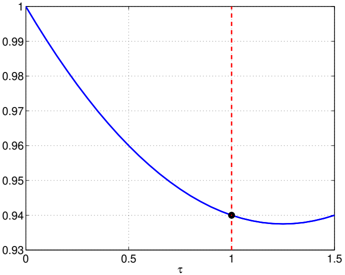


Figure: The three possible scenarios for selection of τ. $y-$axis denotes $m_k(\tau\bar{\mathbf{p}}_k^s)$.

The Cauchy Point — Explicit Expressions

The unconstrained minimizer of the quadratic is

$$\tau_k^* = \frac{\|\nabla f(\bar{\mathbf{x}}_k)\|^3}{\Delta_k \nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)}.$$

Hence we have, for the **Cauchy point**

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

The Cauchy point is cheap to calculate — no matrix inversions, or factorizations are required.

A trust-region method will be globally convergent if its steps $\bar{\mathbf{p}}_k$ give reductions in the models $m_k(\bar{\mathbf{p}})$ that is at least some fixed multiple of the decrease attained by the Cauchy point in each iteration.

The Cauchy Point — Are We Done?

The Cauchy point $\bar{\mathbf{p}}_k^c$ gives us sufficient reduction for global convergence and it is cheap-and-easy to compute. Is there any reason to look for other (approximate) solutions of

$$\underset{\|\bar{\mathbf{p}}\|\leq\Delta_k}{\arg\min}\left[f(\bar{\mathbf{x}}_k)+\bar{\mathbf{p}}^T\nabla f(\bar{\mathbf{x}}_k)+\frac{1}{2}\bar{\mathbf{p}}^T B_k\bar{\mathbf{p}}\right] \quad ???$$

The Cauchy Point — Are We Done?

The Cauchy point $\bar{\mathbf{p}}_k^c$ gives us sufficient reduction for global convergence and it is cheap-and-easy to compute. Is there any reason to look for other (approximate) solutions of

$$\underset{\|\bar{\mathbf{p}}\| \leq \Delta_k}{\arg\min} \left[f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2} \bar{\mathbf{p}}^T B_k \bar{\mathbf{p}} \right] \quad ???$$

Well, yes. Using the Cauchy point as our step means that we have implemented the **Steepest Descent** method, with a particular step length. From previous discussion (and HW#1) we know that steepest descent converges slowly (linearly) even when the step length is chosen optimally.

$\therefore$ there is room for improvement

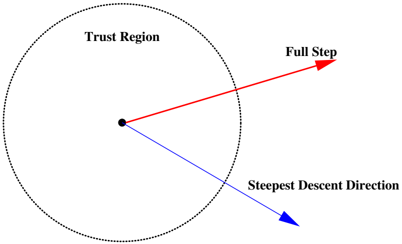
The Dogleg Method

Strategy: Dogleg

Method: Dogleg (for Trust-region).

Use When: The model Hessian B_k is positive definite.

At a point $\bar{\mathbf{x}}_k$ we have already looked at two steps — a step in the steepest descent direction, and the full step.



The Dogleg Method

The **full step** is given by the unconstrained minimum of the quadratic model

$$\bar{\mathbf{p}}_k^{\mathrm{FS}} = -B_k^{-1}\nabla f(\bar{\mathbf{x}}_k).$$

The step in the **steepest descent direction** is given by the unconstrained minimum of the quadratic model along the steepest descent direction

$$\bar{\mathbf{p}}_k^{U} = -\frac{\nabla f(\bar{\mathbf{x}}_k)^T \nabla f(\bar{\mathbf{x}}_k)}{\nabla f(\bar{\mathbf{x}}_k)^T B_k \nabla f(\bar{\mathbf{x}}_k)} \nabla f(\bar{\mathbf{x}}_k).$$

When the trust region is small, the quadratic term $\frac{1}{2}\bar{\mathbf{p}}^T B_k \bar{\mathbf{p}}$ is small, so the minimum of

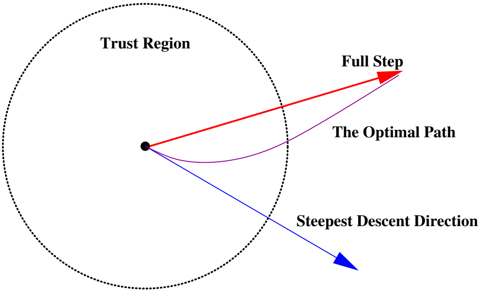
$$\operatorname*{arg\,min}_{\|\bar{\mathbf{p}}\| \leq \Delta_k} \left[f(\bar{\mathbf{x}}_k) + \bar{\mathbf{p}}^T \nabla f(\bar{\mathbf{x}}_k) + \frac{1}{2}\bar{\mathbf{p}}^T B_k \bar{\mathbf{p}} \right],$$

is achieved very close to the steepest descent direction.

The Dogleg Method

On the other hand, as the trust region gets larger ($\Delta_k \to \infty$) the optimum will move to the full step.

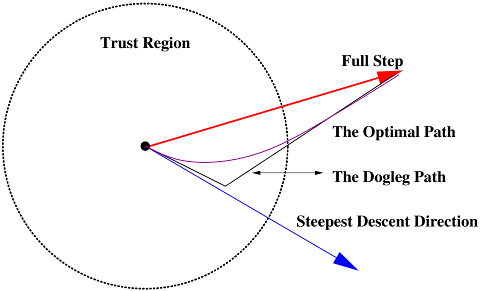
If we plot the optimum as a function of the size of the trust region, we get a smooth path:



The Dogleg Method

The idea of the dogleg method is to **(i)** approximate this path, since the analytical expression for it is quite expensive; and **(ii)** to optimize the model $m_k(\bar{\mathbf{p}})$ along the approximate path subject to the trust region constraint.

The approximate path is a line segment running from $\bar{\mathbf{0}}$ to $\bar{\mathbf{p}}_k^U$, connected to a second line segment running from $\bar{\mathbf{p}}_k^U$ to $\bar{\mathbf{p}}_k^{FS}$, as shown in the figure below



The Dogleg Method

Formally, the dogleg path can be described by one parameter τ

$$\tilde{\bar{p}}(\tau) = \begin{cases} \tau \bar{\mathbf{p}}_k^U & 0 \leq \tau \leq 1 \\ \bar{\mathbf{p}}_k^U + (\tau - 1)(\bar{\mathbf{p}}_k^{\mathrm{FS}} - \bar{\mathbf{p}}_k^U) & 1 \leq \tau \leq 2 \end{cases}$$

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

Lemma

Let B_k be positive definite, then

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

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

This means that the optimum along the dogleg path is achieved at the point where the path exits the trust-region (if it does), otherwise the full step is allowed and optimal.

The Dogleg Method

If the full step is not allowed, then the exit point for the dogleg path is given by the scalar quadratic equation

$$\left\| \bar{\mathbf{p}}_k^U + (\tau - 1)(\bar{\mathbf{p}}_k^{\mathrm{FS}} - \bar{\mathbf{p}}_k^U) \right\|^2 = \Delta_k^2, \quad \tau \in [1, 2]$$

assuming that $\bar{\mathbf{p}}_k^U$ is allowable, otherwise the exit point is along the steepest descent path

$$\left\| \tau \bar{\mathbf{p}}_k^U \right\|^2 = \Delta_k^2, \quad \tau \in [0, 1].$$

The Dogleg Method (Algorithm)

Algorithm: The Dogleg Step

```
If( ‖p̄_k^U‖ ≥ Δ_k ), then
```

$$\bar{\mathbf{p}}_k^{\mathrm{DL}} = \Delta_k \cdot \bar{\mathbf{p}}_k^{\mathrm{U}} \,/\, \|\bar{\mathbf{p}}_k^{\mathrm{U}}\|,$$

```
elseif( ‖p̄_k^FS‖ ≤ Δ_k ), then
```

$$\bar{\mathbf{p}}_k^{\mathrm{DL}} = \bar{\mathbf{p}}_k^{\mathrm{FS}},$$

```
else
```

$$\bar{\mathbf{p}}_k^{\mathrm{DL}} = \bar{\mathbf{p}}_k^{\mathrm{U}} + (\tau^* - 1)(\bar{\mathbf{p}}_k^{\mathrm{FS}} - \bar{\mathbf{p}}_k^{\mathrm{U}})$$

where $\tau^* \in [1, 2]$ so that $\|\bar{\mathbf{p}}_k^{\mathrm{U}} + (\tau^* - 1)(\bar{\mathbf{p}}_k^{\mathrm{FS}} - \bar{\mathbf{p}}_k^{\mathrm{U}})\|^2 = \Delta_k^2$

```
end
```

Next time we will look at dealing with indefinite model Hessians B_k...

Index