Trust-Region Methods

Oswin Krause, NO, 2022

KØBENHAVNS UNIVERSITET

Last-Week: Line-Search based Gradient Descent

1. Set $m(p) = f(x) + p^T \nabla f(x)$
2. Pick $p = -\nabla f(x)$
3. Find α such, that $f(x + \alpha p)$ fulfills Wolfe conditions
4. Set $x \to x + \alpha p$ and go to 1.

This week: Trust-Region Newton (Idea)

1. Set $m(p) = f(x) + p^T \nabla f(x) + \frac{1}{2} p^T \nabla^2 f(x) p$ (second order Taylor)
2. $p = \min_{p'} m(p')$ such, that $\|p\| \leq \Delta$
3. Adjust Δ based on how well $m(p)$ approximates $f(x+p)$
4. If $f(x+p)$ sufficiently better than $f(x)$
   4.1 Set $x \to x + p$
5. Go to 1.

Adapting Δ

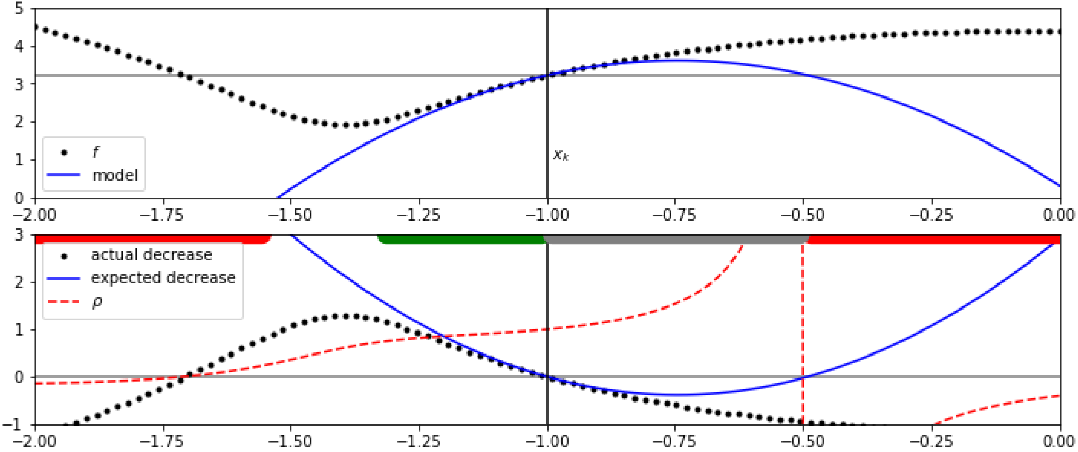
- In Trust-Region algorithms, the trust-region replaces the line-search.
- Δ represents the radius in which we trust our model to approximate the function sufficiently well.
- We need to adapt Δ if our model under-performed in the past

$\rho(p)$ a measure for model quality

$$\rho = \frac{\overbrace{f(x) - f(x+p)}^{\text{Actual decrease}}}{\underbrace{m(0) - m(p)}_{\text{Expected decrease}}}$$

- Expected decrease should always be positive: we solve for the minimum.
- $\rho < 0$: Model predicts decrease where function increases.
- $\rho \approx 1$: Model approximates function well
- But: $\rho = 1$ is not a good target for adapting Δ
  - Too small steps.
  - Goal: adapt Δ such that it prevents bad steps ($\rho \lesssim 0$) but does not shorten good steps.

$\rho(p)$ a measure for model quality



Adapting $\rho(p)$

function $\textsc{AdjustTrustRegion}(\Delta, \rho, \|p\|)$
    if $\rho < 1/4$ **then** $\triangleright$ model-mismatch too large
        $\Delta \leftarrow 1/4\Delta$ $\triangleright$ shrink region
    else if $\rho > 3/4$ and $\|p\| = \Delta$ **then** $\triangleright$ if a good step wants to leave the region
        $\Delta \leftarrow \min(2\Delta, \Delta_{\mathsf{max}})$ $\triangleright$ increase region
    end if
    return Δ $\triangleright$ otherwise do nothing
end function

The Trust-Region Problem

We need to solve the trust-region problems

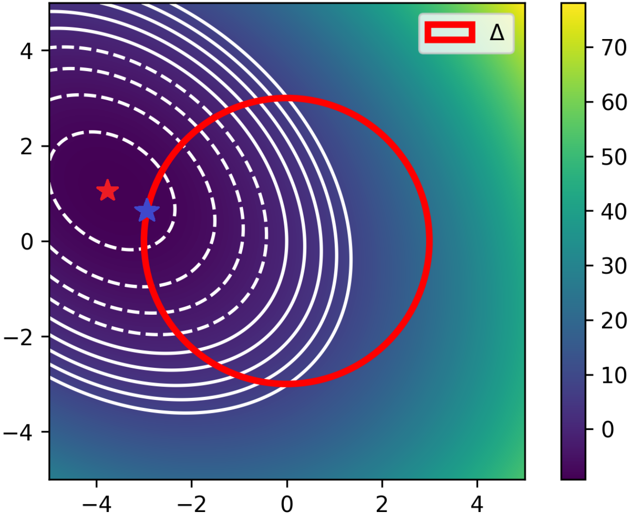
$$\min_{p\in\mathbb{R}^n} f + g^T p + \frac{1}{2} p^T B p$$
$$\text{s.t.} \|p\| \leq \Delta$$

Here, f, g, B are parameters of the local model approximation, for example

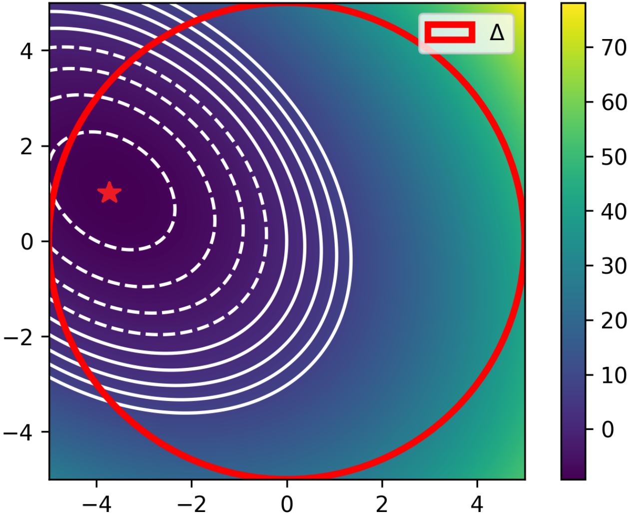
- $f = f(x)$ function-value
- $g = \nabla f(x)$ gradient
- $B = \nabla^2 f(x)$ Hessian

How do these problems look like?

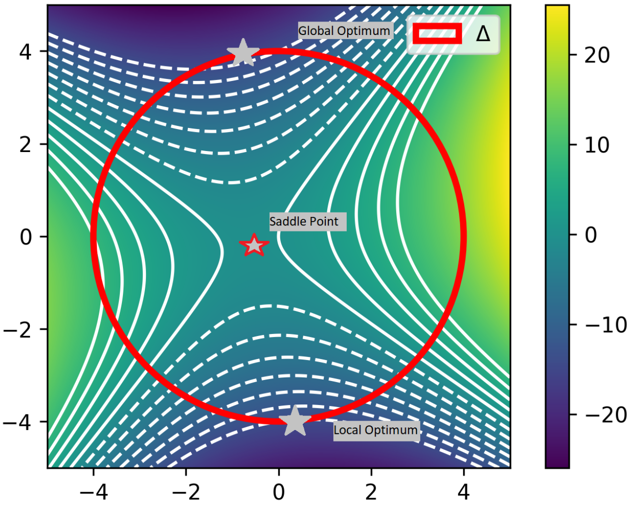
Trust-Region Problem: Positive Definite Hessian



Trust-Region Problem: Positive Definite Hessian, Large Radius



Trust-Region Problem: Indefinite Hessian



Approximately Solving the Trust-Region Problem

We usually only need some "better" point. No perfect solution required

- Approach 1: Cauchy Point
  - Find the optimum of m in direction $p^C = -\alpha g$
  - Good: Cheap, Simple, line-search.
  - Bad: We could as well just do a line-search on f instead.

Approximately Solving the Trust-Region Problem

We usually only need some "better" point. No perfect solution required

- Approach 1: Cauchy Point
  - Find the optimum of m in direction $p^C = -\alpha g$
  - Good: Cheap, Simple, line-search.
  - Bad: We could as well just do a line-search on f instead.
- Approach 2: Dog-Leg
  - Define a path that first goes through p^C and then towards optimum $p^N = -B^{-1}g$
  - Good: At least as much progress as Cauchy, and might get as good as Newton step.
  - Bad: Newton step only defined for positive definite Hessian.

Approximately Solving the Trust-Region Problem

We usually only need some "better" point. No perfect solution required

- Approach 1: Cauchy Point
  - Find the optimum of m in direction $p^C = -\alpha g$
  - Good: Cheap, Simple, line-search.
  - Bad: We could as well just do a line-search on f instead.
- Approach 2: Dog-Leg
  - Define a path that first goes through p^C and then towards optimum $p^N = -B^{-1}g$
  - Good: At least as much progress as Cauchy, and might get as good as Newton step.
  - Bad: Newton step only defined for positive definite Hessian.
- Approach 3: Two-Dimensional Subspace minimization
  - $\min_{\alpha,\beta} m(\alpha p^C + \beta p^N)$, such, that $\|\alpha p^C + \beta p^N\| \leq \Delta$
  - Good: At least as good as Dog-Leg and still easy to compute
  - Bad: Also requires PD Hessian

Approximately Solving the Trust-Region Problem

We usually only need some "better" point. No perfect solution required

- Approach 1: Cauchy Point
  - Find the optimum of m in direction $p^C = -\alpha g$
  - Good: Cheap, Simple, line-search.
  - Bad: We could as well just do a line-search on f instead.
- Approach 2: Dog-Leg
  - Define a path that first goes through p^C and then towards optimum $p^N = -B^{-1}g$
  - Good: At least as much progress as Cauchy, and might get as good as Newton step.
  - Bad: Newton step only defined for positive definite Hessian.
- Approach 3: Two-Dimensional Subspace minimization
  - $\min_{\alpha,\beta} m(\alpha p^C + \beta p^N)$, such, that $\|\alpha p^C + \beta p^N\| \leq \Delta$
  - Good: At least as good as Dog-Leg and still easy to compute
  - Bad: Also requires PD Hessian
- Can we find a solution that works for indefinite Hessians?

Intuition: Solving the Trust-Region Problem

- An infeasible solution p has $\|p\| > \Delta$

Intuition: Solving the Trust-Region Problem

- An infeasible solution p has $\|p\| > \Delta$
- Idea: Add penalisation term based on $\|p\|^2 = p^T p$

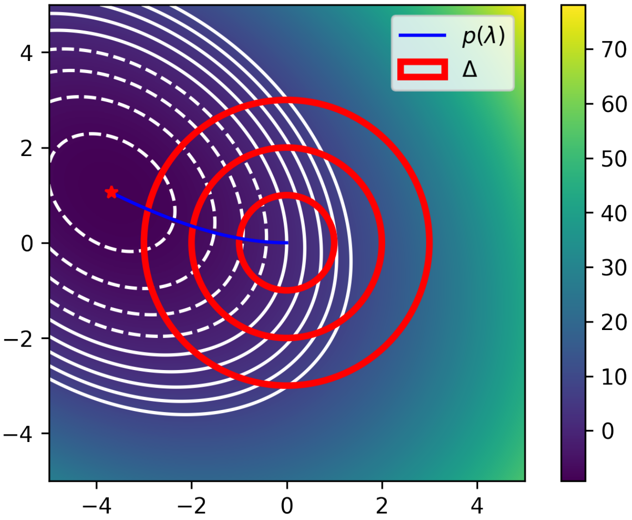
Intuition: Solving the Trust-Region Problem

- An infeasible solution p has $\|p\| > \Delta$
- Idea: Add penalisation term based on $\|p\|^2 = p^T p$
- Adapt model:

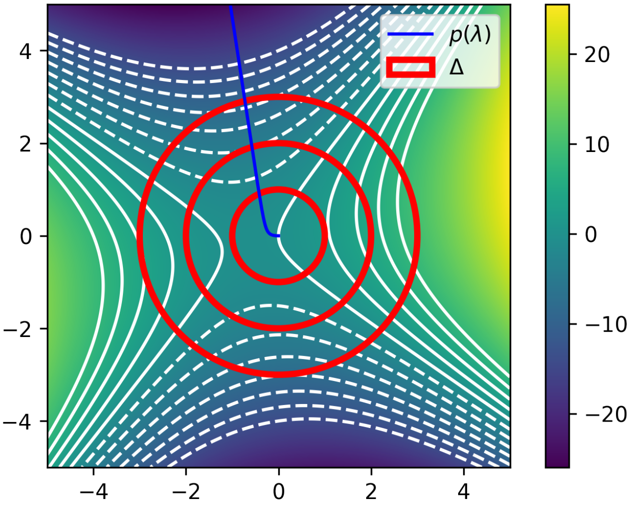
$$\hat{m}(p) = m(p) + \frac{\lambda}{2} p^T p$$

- Does this idea work?
  - Clearly: Steps must become shorter as λ increases
  - if m has indefinite hessian, a large λ can give positive curvature.
  - Which $\lambda > 0$ leads to the right solution?
  - Is this the global optimum?

Penalized solution paths: Indefinite Hessian



Penalized solution paths: Indefinite Hessian



After Visualisation: Might this be correct?

- In both examples, our set of solutions seemed to have passed through the optimum
- We will show, the global optimum lies on this set.

The core Theorem of this algorithm

Theorem (4.1)

Let

$$\min_{p\in\mathbb{R}^d} f + g^T p + \frac{1}{2}p^T Bp$$
$$s.t. \|p\| \leq \Delta$$

The vector p is a global solution of the optimization problem if and only if p is feasible and there is a scalar $\lambda \geq 0$ such, that the following conditions are satisfied:

$$(B + \lambda I)p^* = -g$$
$$\lambda \cdot (\|p^*\| - \Delta) = 0$$
$$(B + \lambda I) \text{ is positive semi-definite}$$

Complementary condition

- We call

$$\lambda \cdot (\|p^*\| - \Delta) = 0$$

  A complementary Condition

Complementary condition

- We call
$$\lambda \cdot (\|p^*\| - \Delta) = 0$$
A complementary Condition
- This can only be fulfilled, if
  - Either, $\lambda = 0$
  - Or $\|p^*\| = \Delta$
  - Both might hold simultaneously under rare conditions.
- The book uses a theorem from chapter 12 to proof Theorem 4.1. We will provide an elementary proof for a slightly weaker version.

Theorem (4.1, (weak))

Let

$$\min_{p\in\mathbb{R}^n} f + g^T p + \frac{1}{2} p^T B p$$
$$s.t. \|p\| \leq \Delta$$

such that for the eigenvector q_n ***of the smallest eigenvalue*** λ_n ***of*** B***, either*** $\lambda_n > 0$ ***or*** $g^T q_n \neq 0$***.***
The vector p *is a global solution of the optimization problem if and only if* p *is feasible and there is a scalar* $\lambda \geq 0$ *such, that the following conditions are satisfied:*

$$(B + \lambda I)p^* = -g$$
$$\lambda \cdot (\|p^*\| - \Delta) = 0$$
$$(B + \lambda I) \text{ is positive } \textbf{definite}$$

The pair p^*, λ ***is the unique global optimum.***

Proof:

Step 1: Show that if a feasible pair (λ, p^*) exists that fulfills the three conditions, then p^* is a solution of the optimization problem.

Proof:

Step 1: Show that if a feasible pair (λ, p^*) exists that fulfills the three conditions, then p^* is a solution of the optimization problem.

Step 1.1: Show that p^* is the minimum of the penalized model:

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

Proof:

Step 1: Show that if a feasible pair (λ, p^*) exists that fulfills the three conditions, then p^* is a solution of the optimization problem.

Step 1.1: Show that p^* is the minimum of the penalized model:

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

The gradient is given by $\nabla \hat{m}(p) = g + (B + \lambda I) p$

Proof:

Step 1: Show that if a feasible pair (λ, p^*) exists that fulfills the three conditions, then p^* is a solution of the optimization problem.

Step 1.1: Show that p^* is the minimum of the penalized model:

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

The gradient is given by $\nabla \hat{m}(p) = g + (B + \lambda I) p$

Inserting p^* fulfilling condition $(B + \lambda I) p^* = -g$ leads to

$$\nabla \hat{m}(p^*) = g + (B + \lambda I) p^* = g - g = 0$$

Proof:

Step 1: Show that if a feasible pair (λ, p^*) exists that fulfills the three conditions, then p^* is a solution of the optimization problem.

Step 1.1: Show that p^* is the minimum of the penalized model:

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

The gradient is given by $\nabla \hat{m}(p) = g + (B + \lambda I) p$

Inserting p^* fulfilling condition $(B + \lambda I) p^* = -g$ leads to

$$\nabla \hat{m}(p^*) = g + (B + \lambda I) p^* = g - g = 0$$

Since $(B + \lambda I)$ is positive definite, p^* is the unique minimizer of $\hat{m}$.

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

$$\hat{m}(p) - \hat{m}(p^*) > 0$$

$$m(p) - m(p^*) + \frac{\lambda}{2}(p^T p - (p^*)^T p^*) > 0$$

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

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

$$\hat{m}(p) - \hat{m}(p^*) > 0$$

$$m(p) - m(p^*) + \frac{\lambda}{2}(p^T p - (p^*)^T p^*) > 0$$

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

By complementary condition (2), $\lambda \cdot (\|p^*\| - \Delta) = 0$.

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

$$\hat{m}(p) - \hat{m}(p^*) > 0$$

$$m(p) - m(p^*) + \frac{\lambda}{2}(p^T p - (p^*)^T p^*) > 0$$

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

By complementary condition (2), $\lambda \cdot (\|p^*\| - \Delta) = 0$. We have one of

- $\lambda = 0 \Rightarrow m(p) > m(p^*)$

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

$$\hat{m}(p) - \hat{m}(p^*) > 0$$

$$m(p) - m(p^*) + \frac{\lambda}{2}(p^T p - (p^*)^T p^*) > 0$$

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

By complementary condition (2), $\lambda \cdot (\|p^*\| - \Delta) = 0$. We have one of

- $\lambda = 0 \Rightarrow m(p) > m(p^*)$
- $\|p^*\| = \Delta \Rightarrow m(p) > m(p^*) + \underbrace{\frac{\lambda}{2}}_{\geq 0} \underbrace{(\Delta^2 - p^T p)}_{\geq 0}$

Proof:

Step 1.2: Show that p^* is global optimum of the original problem.

Let $p \neq p^*$ with $\|p\| \leq \Delta$. Since p^* is optimum of $\hat{m}$ it holds

$$\hat{m}(p) - \hat{m}(p^*) > 0$$

$$m(p) - m(p^*) + \frac{\lambda}{2}(p^T p - (p^*)^T p^*) > 0$$

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

By complementary condition (2), $\lambda \cdot (\|p^*\| - \Delta) = 0$. We have one of

- $\lambda = 0 \Rightarrow m(p) > m(p^*)$
- $\|p^*\| = \Delta \Rightarrow m(p) > m(p^*) + \underbrace{\frac{\lambda}{2}}_{\geq 0} \underbrace{(\Delta^2 - p^T p)}_{\geq 0}$

This shows Step 1 as $m(p) > m(p^*)$.

Proof: Intermezzo

Where are we?

- We have shown that if a pair p^*, λ exists, p^* is a solution of our penalized model.
- Further, p^* is the global optimum of the original problem
- We still need to show, that
  - Each problem can be solved by our penalization approach.
  - The solution is unique.

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Lets have a look at the conditions

- Third condition, $B + \lambda I$ is PD

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Lets have a look at the conditions

- Third condition, $B + \lambda I$ is PD
  - Fulfilled for $\lambda > -\lambda_n$, where λ_n smallest eigenvalue of B

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Lets have a look at the conditions

- Third condition, $B + \lambda I$ is PD
  - Fulfilled for $\lambda > -\lambda_n$, where λ_n smallest eigenvalue of B
- First condition $(B + \lambda I)p^* = -g$

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Lets have a look at the conditions

- Third condition, $B + \lambda I$ is PD
  - Fulfilled for $\lambda > -\lambda_n$, where λ_n smallest eigenvalue of B
- First condition $(B + \lambda I)p^* = -g$
  - Can always be found from λ that fulfills third condition.

Proof:

Step 2: Show that for all $\Delta > 0$ a unique pair λ, p^* exists, $\lambda > 0$, p^* feasible, that fulfills all three conditions.

Lets have a look at the conditions

- Third condition, $B + \lambda I$ is PD
  - Fulfilled for $\lambda > -\lambda_n$, where λ_n smallest eigenvalue of B
- First condition $(B + \lambda I)p^* = -g$
  - Can always be found from λ that fulfills third condition.
- Second condition $\lambda \cdot (\|p\| - \Delta) = 0$, $\lambda \geq 0$
  - This and feasibility of p^* requires some work.

Proof:

Step 2.1: Define λ-Path

p^* is a function depending on λ:

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

Proof:

Step 2.1: Define λ-Path

p^* is a function depending on λ:

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

Let λ_i eigenvalues of B with eigenvectors q_i. Then

$$\|p^*(\lambda)\|^2 = \sum_{i=1}^{n} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2}$$

Proof:

Step 2.2: Need to show existence of solution:

- Case 1: B is PD
  - Either $\|p^*(0)\| \leq \Delta \Rightarrow$ unconstrained optimum is feasible
  - Or $\|p^*(\lambda)\| = \Delta,$ for some $\lambda > 0 \Rightarrow$ we can find feasible p^*

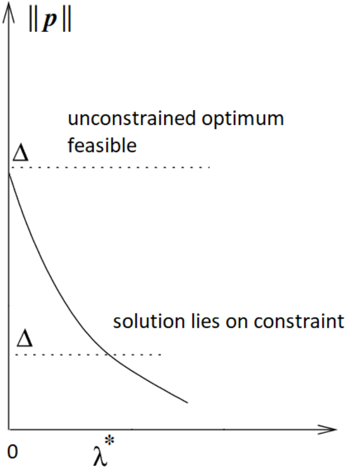
Proof:

Step 2.2: Need to show existence of solution:

- Case 1: B is PD
  - Either $\|p^*(0)\| \leq \Delta \Rightarrow$ unconstrained optimum is feasible
  - Or $\|p^*(\lambda)\| = \Delta, \quad$ for some $\lambda > 0 \Rightarrow$ we can find feasible p^*
- Case 2: B is not PD
  - Unconstrained optimum does not exist (due to our weaker condition $q_n^T g \neq 0$)
  - $\|p^*(\lambda)\| = \Delta, \quad$ for some $\lambda > -\lambda_n$

Proof:

Step 2.2, Case 1: B PD.



Proof:

Step 2.2, Case 1: B PD.

- $p^*(0) = -(B + \lambda I)^{-1} g = -B^{-1} g$ exists and is minimizer of m
- If $\|p^*(0)\| > \Delta$

Proof:

Step 2.2, Case 1: B PD.

- $p^*(0) = -(B + \lambda I)^{-1} g = -B^{-1} g$ exists and is minimizer of m
- If $\|p^*(0)\| > \Delta$

  Limit of $p^*(\lambda)$ as $\lambda \to \infty$:

$$\|p^*(\lambda)\|^2 = \sum_{i=1}^{n} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2} \xrightarrow{\lambda \to \infty} 0$$

Proof:

Step 2.2, Case 1: B PD.

- $p^*(0) = -(B + \lambda I)^{-1} g = -B^{-1} g$ exists and is minimizer of m
- If $\|p^*(0)\| > \Delta$

  Limit of $p^*(\lambda)$ as $\lambda \to \infty$:

$$\|p^*(\lambda)\|^2 = \sum_{i=1}^{n} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2} \xrightarrow{\lambda \to \infty} 0$$

  Easy to show: $\|p^*(\lambda)\|^2$ continuous and strictly monotonous decreasing for $\lambda > 0$

Proof:

Step 2.2, Case 1: B PD.

- $p^*(0) = -(B + \lambda I)^{-1} g = -B^{-1} g$ exists and is minimizer of m
- If $\|p^*(0)\| > \Delta$

  Limit of $p^*(\lambda)$ as $\lambda \to \infty$:

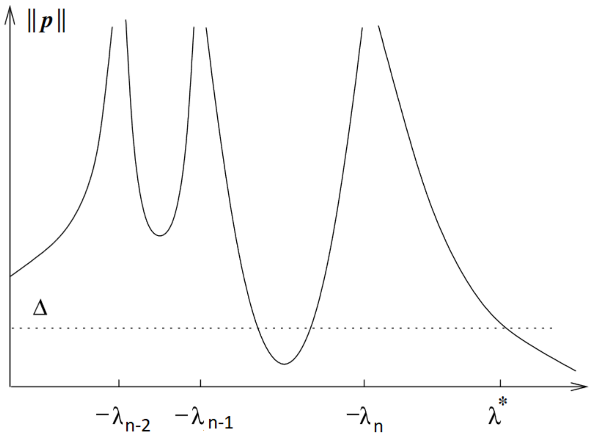
  $$\|p^*(\lambda)\|^2 = \sum_{i=1}^{n} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2} \xrightarrow{\lambda \to \infty} 0$$

  Easy to show: $\|p^*(\lambda)\|^2$ continuous and strictly monotonous decreasing for $\lambda > 0$

  $\to$ there exists a unique λ with $\|p^*(\lambda)\| = \Delta$
- $\|p^*(0)\| \leq \Delta$
  - Unconstrained optimum is feasible.
  - Since $\|p^*(\lambda)\|^2$ is monotonous decreasing, this solution is unique.

Proof:

Step 2.2, Case 2: B not PD.



Proof:

Step 2.2: Case 2:B not PD.

- We have $\lambda_i + \lambda > 0$ for $\lambda > -\lambda_n$ and $q_n^T g \neq 0$ by assumption.
- Limit $\lambda \to \infty$

$$\|p^*(\lambda)\|^2 \xrightarrow{\lambda \to \infty} 0$$

Proof:

Step 2.2: Case 2:B not PD.

- We have $\lambda_i + \lambda > 0$ for $\lambda > -\lambda_n$ and $q_n^T g \neq 0$ by assumption.
- Limit $\lambda \to \infty$

$$\|p^*(\lambda)\|^2 \xrightarrow{\lambda \to \infty} 0$$

- Limit $\lambda \to -\lambda_n$

$$\|p^*(\lambda)\|^2 = \underbrace{\sum_{i=1}^{n-1} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2}}_{\geq 0} + \underbrace{(q_n^T g)^2}_{>0} \frac{1}{\underbrace{(\lambda_n + \lambda)}_{\to 0}{}^2} \xrightarrow{\lambda \to -\lambda_n} \infty$$

Proof:

Step 2.2: Case 2:B not PD.

- We have $\lambda_i + \lambda > 0$ for $\lambda > -\lambda_n$ and $q_n^T g \neq 0$ by assumption.
- Limit $\lambda \to \infty$

$$\|p^*(\lambda)\|^2 \xrightarrow{\lambda \to \infty} 0$$

- Limit $\lambda \to -\lambda_n$

$$\|p^*(\lambda)\|^2 = \underbrace{\sum_{i=1}^{n-1} (q_i^T g)^2 \frac{1}{(\lambda_i + \lambda)^2}}_{\geq 0} + \underbrace{(q_n^T g)^2}_{>0} \frac{1}{\underbrace{(\lambda_n + \lambda)}_{\to 0}{}^2} \xrightarrow{\lambda \to -\lambda_n} \infty$$

- $\|p^*(\lambda)\|^2$ continuous and monotonous decreasing for $\lambda > -\lambda_n$ leads to the result.

What is missing to the full Theorem?

Are there cases, which are not covered?

- There can be problems where the optimal solution is not unique due to $q_n^T g = 0$.
- The book calls this "the hard case"
- There is an assignment about this.

How to find λ?

- We have shown a suitable λ exists under very broad conditions!
- How can we find it?
- Two approaches:
  - Bisection algorithm (see theoretical assignment)
  - Book gives another approach to quickly find λ

How to check correctness of solution?

- Once we found p^*, λ how do we know our solution is correct?
- Check, whether Theorem 4.1 holds!