Part 1: A gentle introduction to nonlinear optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\; f(x) \;\text{ subject to }\; c_{\mathcal{E}}(x) = 0 \;\text{ and }\; c_{\mathcal{I}}(x) \geqslant 0$$

Course on continuous optimization, STFC-RAL, February 2021

WHAT IS NONLINEAR PROGRAMMING?

Nonlinear optimization $\equiv$ nonlinear programming

$$\underset{x}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c_{\mathcal{E}}(x) = 0 \;\; \text{and} \;\; c_{\mathcal{I}}(x) \geqslant 0$$

where

objective function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

constraints $c_{\mathcal{E}} : \mathbb{R}^n \longrightarrow \mathbb{R}^{m_e}$ $(m_e \leqslant n)$ and
$c_{\mathcal{I}} : \mathbb{R}^n \longrightarrow \mathbb{R}^{m_i}$

- there may also be integrality restrictions
- concentrate on minimization since

$$\max_{x \in \mathcal{F}} f(x) = - \min_{x \in \mathcal{F}} (-f(x))$$

AN EXAMPLE

Optimization of
a high-pressure
gas network

Transco
National
Transmission
System

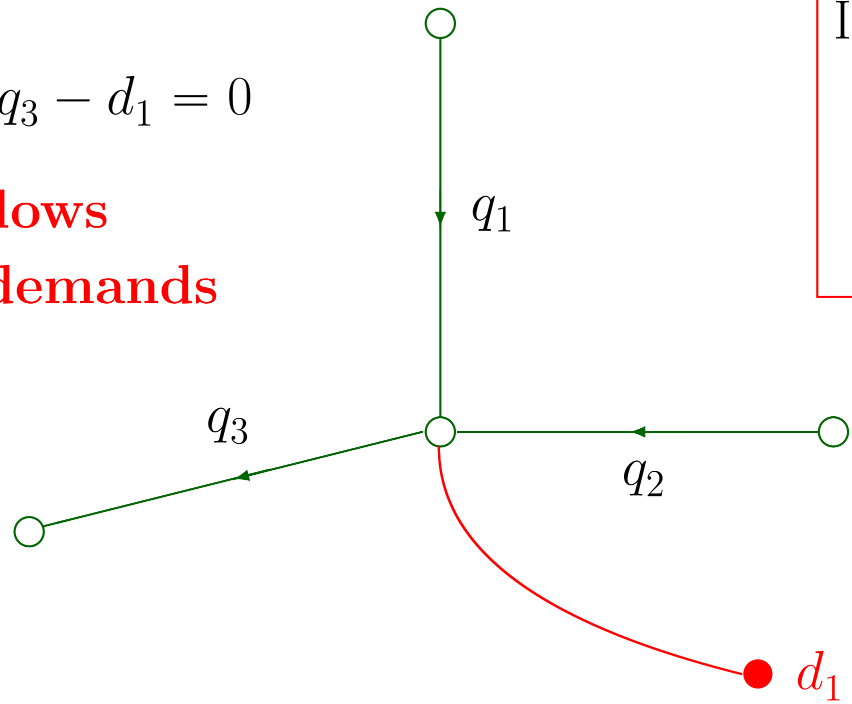
British Gas (Transco)
Oxford University
RAL

NODE EQUATIONS

$$q_1 + q_2 - q_3 - d_1 = 0$$

where q_i **flows**

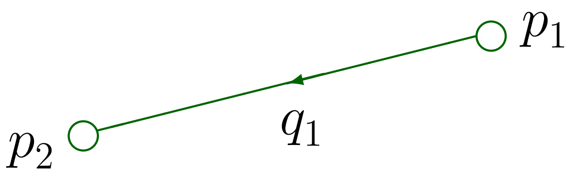
d_i **demands**



In general: $Aq - d = 0$

- linear
- sparse
- structured

PIPE EQUATIONS



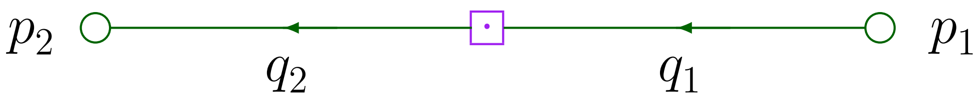
$$p_2^2 - p_1^2 + k_1 q_1^{2.8359} = 0$$

where p_i **pressures**
q_i **flows**
k_i **constants**

In general: $A^T p^2 + K q^{2.8359} = 0$

- non-linear
- sparse
- structured

COMPRESSOR CONSTRAINTS



$$q_1 - q_2 + z_1 \cdot c_1(p_1, q_1, p_2, q_2) = 0$$

where p_i **pressures**

q_i **flows**

z_i **0–1 variables**

$= 1$ if machine is on

c_i **nonlinear functions**

In general: $A_2^T q + z \cdot c(p, q) = 0$

- non-linear
- sparse
- structured
- 0–1 variables

OTHER CONSTRAINTS

Bounds on pressures and flows

$$\begin{aligned} p_{\min} \;\leqslant\; p \;\leqslant\; p_{\max} \\ q_{\min} \;\leqslant\; q \;\leqslant\; q_{\max} \end{aligned}$$

- simple bounds on variables

OBJECTIVES

Many possible objectives

- maximize / minimize sum of pressures
- minimize compressor fuel costs
- minimize supply

+ combinations of these

STATISTICS

British Gas National Transmission System

- 199 nodes
- 196 pipes
- 21 machines

Steady state problem

$\sim$400 variables

24-hour variable demand problem with 10 minute discretization

$\sim$58,000 variables

Challenge: Solve this in real time

TYPICAL PROBLEM

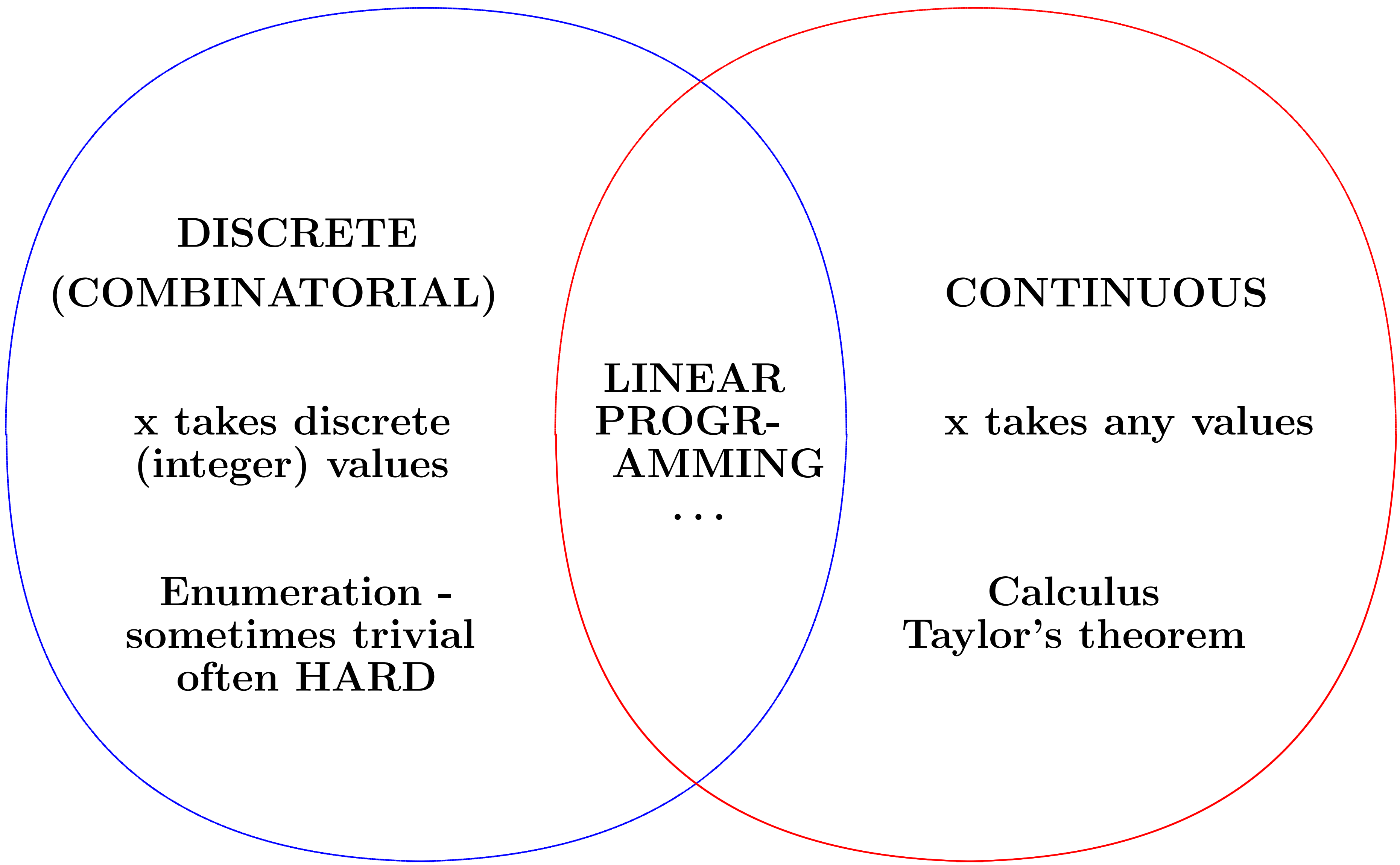
This problem is typical of real-world, large-scale applications

- simple bounds
- linear constraints
- nonlinear constraints
- structure
- global solution “required”
- integer variables
- discretization

(SOME) OTHER APPLICATION AREAS

- minimum energy problems
- gas production models
- hydro-electric power scheduling
- structural design problems
- portfolio selection
- parameter determination in financial markets
- production scheduling problems
- computer tomography (image reconstruction)
- efficient models of alternative energy sources
- traffic equilibrium models
- **machine learning/neural nets**

CLASSIFICATION OF OPTIMIZATION PROBLEMS



OPTIMIZATION PROBLEMS

Unconstrained minimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x)$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

Equality constrained minimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

where the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(m \leqslant n)$

Inequality constrained minimization:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) \geqslant 0$$

where $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(m$ may be larger than $n)$

OPTIMALITY CONDITIONS

Optimality is **hidden**; it needs further thought and work to verify

Optimality conditions are useful because:

- they provide a means of guaranteeing that a candidate solution is indeed optimal (**sufficient conditions**), and
- they indicate when a point is not optimal (**necessary conditions**)

Furthermore they

- guide in the design of algorithms, since lack of optimality $\Longleftrightarrow$ indication of improvement

THE GRADIENT

Let $x \in \mathrm{IR}^n$

Suppose that $f(x)$ is continuously differentiable ($f \in C^1$).

Then its **gradient** $g(x)$ is the vector whose i-th component

$$g_i(x) = \frac{\partial f(x)}{\partial x_i}$$

for $1 \leqslant i \leqslant n$

E.g, if

$$f(x) = x_1^2 + x_1 x_2$$

then

$$g(x) = \begin{pmatrix} 2x_1 + x_2 \\ x_1 \end{pmatrix}$$

THE HESSIAN MATRIX

Suppose that $f(x)$ is twice-continuously differentiable ($f \in C^2$). Then its **Hessian** (Otto Hesse, 1811–1874) $H(x)$ is the matrix whose i,j-th component

$$H_{i,j}(x) = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$$

for $1 \leqslant i,j \leqslant n$

E.g, if

$$f(x) = x_1^2 + x_1 x_2$$

then

$$H(x) = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$$

Notice that the Hessian is always **symmetric**

THE JACOBIAN MATRIX

Suppose that $c(x)$ is vector-valued and continuously differentiable ($c : \mathbb{R}^n \to \mathbb{R}^m, c \in C^1$). Then its **Jacobian** (Carl Jacobi, 1804-1851) $J(x)$ is the matrix whose i, j-th component

$$J_{i,j}(x) = \frac{\partial c_i(x)}{\partial x_j}$$

for $1 \leqslant i \leqslant m$ and $1 \leqslant j \leqslant n$

E.g, if

$$c(x) = \begin{pmatrix} x_1^2 \\ x_1 + x_2^3 \end{pmatrix}$$

then

$$J(x) = \begin{pmatrix} 2x_1 & 0 \\ 1 & 3x_2^2 \end{pmatrix}$$

Notice that the i-th row of the Jacobian is the transpose of the gradient of $c_i(x)$. Also that if $c(x) = g(x)$, then $J(x) = H(x)$

INNER PRODUCTS AND NORMS

Suppose that $x, y \in \mathrm{IR}^n$. Then the **inner product** $\langle x, y \rangle$ between x and y is the component-wise sum

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

This defines the (Euclidean) **norm**

$$\|x\|_2 = \sqrt{\langle x, x \rangle} \equiv \sqrt{\sum_{i=1}^{n} x_i^2}$$

Notice that $\|x\|_2$ is always **non-negative** and only zero when $x = 0$

- If S is a symmetric matrix, $\|S\| = \max_{\|x\|=1} \|Sx\|$
- There are other norms, e.g., $\|x\|_1 = \sum_{i=1}^{n} |x_i|$ and $\|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$
- if we don't say otherwise $\| \cdot \| = \| \cdot \|_2$

EIGENPAIRS & POSITIVE-DEFINITE MATRICES

Let S be a real, symmetric $n \times n$ matrix.

S is said to have an **eigenpair** (λ, v) if

$$Sv = \lambda v,$$

where the **eigenvalue** λ is real and its **eigenvector** v has $\|v\| = 1$.

- S has n eigenvalues λ_i, and associated eigenvectors v_i, $1 \leqslant i \leqslant n$
- the eigenvectors are mutually orthogonal i.e., $\langle v_i, v_j \rangle = 0$ if $i \neq j$.
- $V = (v_1, \ldots, v_n)$, S has a **spectral decomposition**

$$S = V^T \Lambda V, \text{ where } \Lambda = \mathrm{diag}(\lambda_i)$$

S is **positive (semi) definite** if (equivalently)

- $\lambda_i > 0$ $(\geqslant 0)$ for $1 \leqslant i \leqslant n$
- $\langle u, Su \rangle > 0$ $(\geqslant 0)$ for all nonzero vectors u

LIPSCHITZ CONTINUITY (don't panic!!)

- $\mathcal{X}$ and $\mathcal{Y}$ sets
- $F : \mathcal{X} \to \mathcal{Y}$
- $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{Y}}$ are norms

Then

- F is **Lipschitz** (Rudolf Lipschitz, 1832–1903) **continuous at** $x \in \mathcal{X}$ if $\exists\ \gamma(x)$ such that

  $$\|F(z) - F(x)\|_{\mathcal{Y}} \leqslant \gamma(x)\|z - x\|_{\mathcal{X}}$$

  for all $z \in \mathcal{X}$.

- F is **Lipschitz continuous throughout/in** $\mathcal{X}$ if $\exists\ \gamma$ such that

  $$\|F(z) - F(x)\|_{\mathcal{Y}} \leqslant \gamma\|z - x\|_{\mathcal{X}}$$

  for all x and $z \in \mathcal{X}$.

Essentially controls how far $F(z)$ is from $F(x)$ as z approaches x

TAYLOR-SERIES APPROXIMATIONS

A fundamental question is:

if we have a function f and know its value and derivatives at x, can we say anything about f at a nearby point $x + s$?

This question was addressed by Brook **Taylor** (1685–1731), who showed that in many cases a series approximation

$$f(x+s) \approx T_p(s) := f(x) + \sum_{i=1}^{p} \frac{f^{(i)}(x)[s]^i}{i!},$$

where $f^{(i)}(x)$ is the i-th derivative of f at x, is increasingly accurate as $p \to \infty$ (NB …there is a lot hidden here in the notation!)

Computationally useful for $p = 1$ and 2:

$$\begin{aligned} m^L(x+s) &= T_1(s) = f(x) + \langle g(x), s \rangle \\ m^Q(x+s) &= T_2(s) = f(x) + \langle g(x), s \rangle + \tfrac{1}{2}\langle s, H(x)s \rangle \end{aligned}$$

A USEFUL TAYLOR APPROXIMATION

Theorem 1.1. Let $\mathcal{S}$ be an open subset of $\mathbb{R}^n$, and suppose $f : \mathcal{S} \to \mathbb{R}$ is continuously differentiable throughout $\mathcal{S}$. Suppose further that $g(x)$ is Lipschitz continuous at x, with Lipschitz constant $\gamma^L(x)$ in some appropriate vector norm. Then, if the segment $x + \theta s \in \mathcal{S}$ for all $\theta \in [0, 1]$,

$$\boxed{\begin{array}{c} |f(x+s) - m^L(x+s)| \leqslant \frac{1}{2}\gamma^L(x)\|s\|^2, \quad \text{where} \\ m^L(x+s) = f(x) + \langle g(x), s\rangle. \end{array}}$$

If f is twice continuously differentiable throughout $\mathcal{S}$ and $H(x)$ is Lipschitz continuous at x, with Lipschitz constant $\gamma^Q(x)$,

$$\boxed{\begin{array}{c} |f(x+s) - m^Q(x+s)| \leqslant \frac{1}{6}\gamma^Q(x)\|s\|^3, \quad \text{where} \\ m^Q(x+s) = f(x) + \langle g(x), s\rangle + \frac{1}{2}\langle s, H(x)s\rangle. \end{array}}$$

ANOTHER USEFUL TAYLOR APPROXIMATION

Theorem 1.2. Let $\mathcal{S}$ be an open subset of IR^n, and suppose $F : \mathcal{S} \to \mathrm{IR}^m$ is continuously differentiable throughout $\mathcal{S}$. Suppose further that $\nabla_x F(x)$ is Lipschitz continuous at x, with Lipschitz constant $\gamma^L(x)$ in some appropriate vector norm and its induced matrix norm. Then, if the segment $x + \theta s \in \mathcal{S}$ for all $\theta \in [0, 1]$,

$$\boxed{\begin{array}{c}\|F(x+s) - M^L(x+s)\| \leqslant \tfrac{1}{2}\gamma^L(x)\|s\|^2, \quad \text{where} \\ M^L(x+s) = F(x) + \nabla_x F(x) s.\end{array}}$$

COROLLARY — NEWTON'S METHOD

Given a Lipschitz C^1 function $F : \mathrm{I\!R}^n \rightarrow \mathrm{I\!R}^n$, Taylor $\Longrightarrow$

$$\|F(x+s) - M^L(x+s)\| \leqslant \tfrac{1}{2}\gamma^L(x)\|s\|^2, \quad \text{where}$$
$$M^L(x+s) = F(x) + \nabla_x F(x) s$$

From given x with small $F(x)$, pick s so that

$$M^L(x+s) = F(x) + \nabla_x F(x) s = 0$$

$\Longrightarrow$

$$\|F(x+s)\| \leqslant \tfrac{1}{2}\gamma^L(x)\|s\|^2 \leqslant \gamma^L(x)\|(\nabla_x F(x))^{-1}\|^2\|F(x)\|^2$$

$\Longrightarrow$ usually quadratic rate of decrease

Choosing $s : \nabla_x F(x) s = -F(x)$ is **Newton's method**
for finding a root of the nonlinear system $F(x) = 0$

BLOCK NEWTON

Given Lipschitz C^1 function $F : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^{n+m}$ such that

$$F(x, y) = \begin{pmatrix} b(x, y) \\ c(x, y) \end{pmatrix}$$

with $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $b : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^n$ and $c : \mathbb{R}^{n+m} \longrightarrow \mathbb{R}^m$

Newton equations are

$$\begin{pmatrix} \nabla_x b(x, y) & \nabla_y b(x, y) \\ \nabla_x c(x, y) & \nabla_y c(x, y) \end{pmatrix} \begin{pmatrix} s_x \\ s_y \end{pmatrix} = - \begin{pmatrix} b(x, y) \\ c(x, y) \end{pmatrix}$$

to get an improvement $x + s_x$ and $y + s_y$

Part 2: Unconstrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Course on continuous optimization, STFC-RAL, February 2021

UNCONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x)$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

- assume that $f \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

CONTENT

We shall discuss:

- optimality conditions
- quadratic minimization
- linesearch methods
- trust-region methods
- (regularization methods)

OPTIMALITY CONDITIONS FOR UNCONSTRAINED MINIMIZATION

First-order necessary optimality:

Theorem 2.1. Suppose that $f \in C^1$, and that x_* is a local minimizer of $f(x)$. Then

$$g(x_*) = 0.$$

Second-order necessary optimality:

Theorem 2.2. Suppose that $f \in C^2$, and that x_* is a local minimizer of $f(x)$. Then $g(x_*) = 0$ and $H(x_*)$ is positive semi-definite, that is

$$\langle s, H(x_*)s \rangle \geqslant 0 \;\text{ for all }\; s \in \mathrm{IR}^n.$$

OPTIMALITY CONDITIONS (cont.)

Second-order sufficient optimality:

Theorem 2.3. Suppose that $f \in C^2$, that x_* satisfies the condition $g(x_*) = 0$, and that additionally $H(x_*)$ is positive definite, that is

$$\langle s, H(x_*)s \rangle > 0 \;\text{ for all }\; s \neq 0 \in \mathrm{I\!R}^n.$$

Then x_* is an isolated local minimizer of f.

MINIMIZING A CONVEX QUADRATIC FUNCTION

Generic convex quadratic problem: (B sym. positive definite)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; q(x) = \langle g, x \rangle + \tfrac{1}{2}\langle x, Bx \rangle$$

If x_* is a minimizer, necessarily

$$\nabla q(x_*) = g + Bx_* = 0 \implies Bx_* = -g$$

Since B is positive definite, x_* is the unique (global) minimizer

How do we find x_*?

- by **factorization**
  - dense/spares Cholesky factorization of $B = LL^T$, L triangular
  - Forward and back solution $Lz = -g$ then $L^T x_* = z$
- approximately by **iteration**

ITERATIVE QUADRATIC MINIMIZATION

Many possible methods, the most effective is the method of **conjugate gradients**:

Given:

- a sequence of linearly-independent vectors $\{p_j\}$, $0 \leqslant j \leqslant n-1$
- a sequence of expanding matrices $P_j = (p_0, \ldots, p_{j-1})$
- a sequence of expanding subspaces

$$\mathcal{P}_j = \{x : x = P_j v \text{ for some } v \in \mathrm{IR}^j\}$$

Generate a sequence of successively improving estimates

$$x_j = \arg\min_{x \in \mathcal{P}_j} q(x)$$

$\Longrightarrow x_n = x_*$

CONJUGATE GRADIENTS — THE CLEVER PARTS

Let $g_j = \nabla q(x_j) = Bx_j + g$

- **(easy)** **if** we can select p_j so that $\{p_i\}$ are ***B*-conjugate**, i.e.,

$$\langle p_j, Bp_i \rangle = 0 \;\; \text{for} \;\; i \leqslant j$$

$\Longrightarrow$

$$x_{j+1} = x_j + \alpha_j p_j, \;\; \text{where} \;\; \alpha_j = -\frac{\langle p_j, g_j \rangle}{\langle p_j, Bp_j \rangle}$$

- **(trivial)**

$$g_{j+1} = g_j + \alpha_j B p_j$$

- **(messy)** we **can** select p_j so that $\{p_i\}$ are B-conjugate via

$$p_{j+1} = -g_{j+1} + \beta_j p_j, \;\; \text{where} \;\; \beta_j = \frac{\|g_{j+1}\|}{\|g_j\|}$$

CONJUGATE-GRADIENT (CG) METHOD

Set $x_0 = 0$, $g_0 = g$, $p_0 = -g$ and $i = 0$.

Until g_i "small", iterate

$$\alpha_i = -\langle g_i, p_i \rangle / \langle p_i, Bp_i \rangle \equiv \arg\min_{\alpha} q(x_i + \alpha p_i)$$

$$x_{i+1} = x_i + \alpha_i p_i$$

$$g_{i+1} = g_i + \alpha_i B p_i \equiv \nabla q(x_{i+1})$$

$$\beta_i = \|g_{i+1}\|_2^2 / \|g_i\|_2^2$$

$$p_{i+1} = -g_{i+1} + \beta_i p_i$$

and increase i by 1

Important features:

- $q(x_j) \leqslant q(x_{j-1})$
- $x_n = x_*$ (in exact arithmetic)
- may stop earlier if B is structured, e.g. clustered eigenvalues
- can accelerate by **preconditioning**

ITERATIVE METHODS FOR GENERAL $f(x)$

- in practice very rare to be able to provide explicit minimizer of f
- iterative method: given starting "guess" x_0, generate sequence

$$\{x_k\}, \quad k = 1, 2, \ldots$$

- **AIM:** ensure that (a subsequence) has some favourable limiting properties:
  - satisfies first-order necessary conditions
  - satisfies second-order necessary conditions

Notation: $f_k = f(x_k)$, $g_k = g(x_k)$, $H_k = H(x_k)$.

Part 2a: Linesearch methods for unconstrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x)$$

Course on continuous optimization, STFC-RAL, February 2021

LINESEARCH METHODS

- calculate a **search direction** d_k from x_k
- ensure that this direction is a **descent direction**, i.e.,

$$\langle g_k, d_k \rangle < 0 \ \text{ if } \ g_k \neq 0$$

(the **slope** $\langle d_k, g_k \rangle$ is negative) so that, for small steps along d_k, the objective function **will** be reduced (Taylor's theorem)

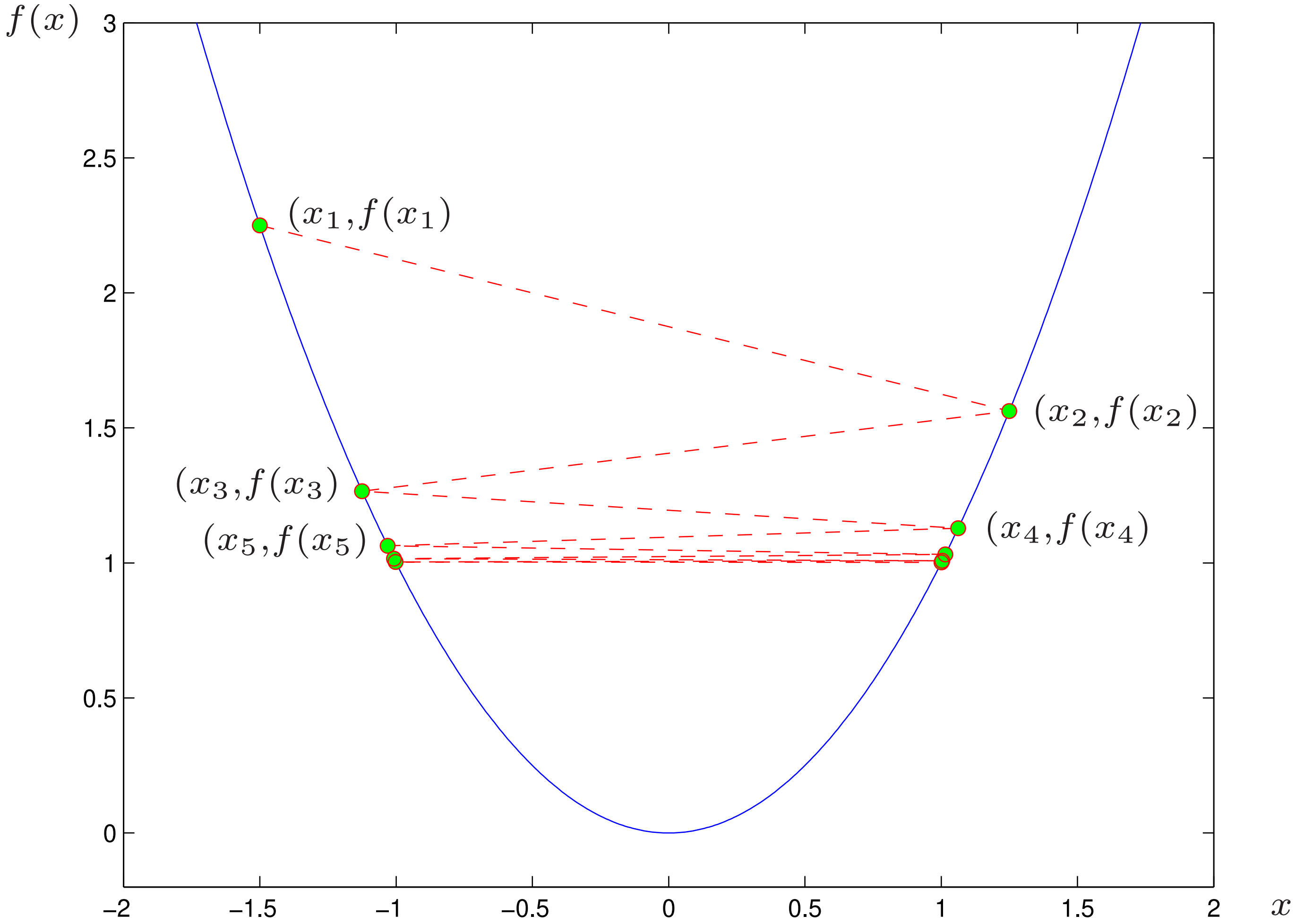
- calculate a suitable **steplength** $\alpha_k > 0$ so that

$$f(x_k + \alpha_k d_k) < f_k$$

- computation of α_k is the **linesearch**—may itself be an iteration
- generic linesearch method:

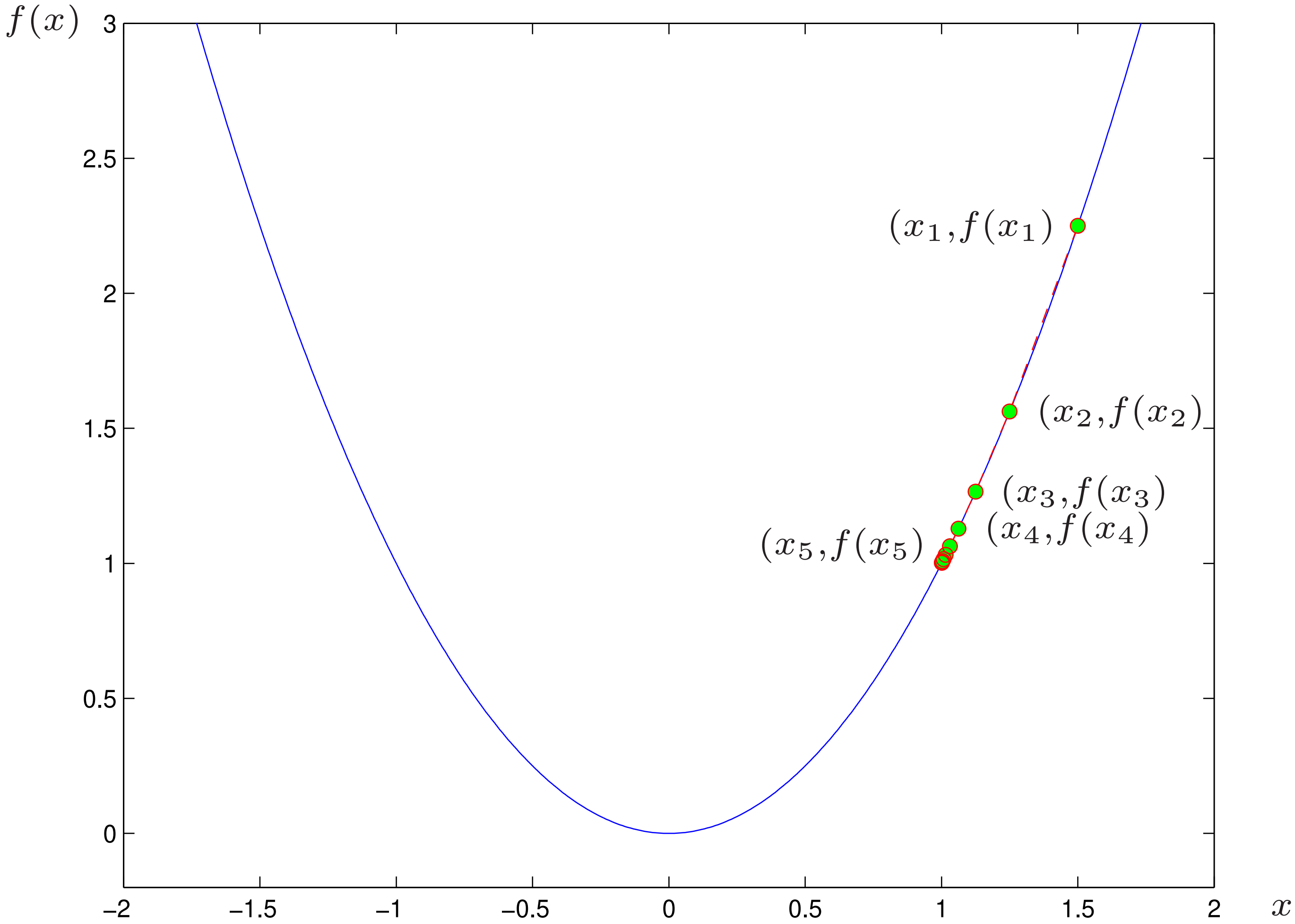
$$x_{k+1} = x_k + \alpha_k d_k$$

STEPS MIGHT BE TOO LONG



The objective function $f(x) = x^2$ and the iterates $x_{k+1} = x_k + \alpha_k d_k$ generated by the descent directions $d_k = (-1)^{k+1}$ and steps $\alpha_k = 2 + 3/2^{k+1}$ from $x_0 = 2$

STEPS MIGHT BE TOO SHORT



The objective function $f(x) = x^2$ and the iterates $x_{k+1} = x_k + \alpha_k d_k$ generated by the descent directions $d_k = -1$ and steps $\alpha_k = 1/2^{k+1}$ from $x_0 = 2$

PRACTICAL LINESEARCH METHODS

- in early days, pick α_k to minimize

$$f(x_k + \alpha d_k)$$

  - **exact** linesearch—univariate minimization
  - rather expensive and certainly not cost effective
- modern methods: **inexact** linesearch
  - ensure steps are neither too long nor too short
  - try to pick "useful" initial stepsize for fast convergence
  - best methods are either
    - "backtracking- Armijo" or
    - "Armijo-Goldstein"

    based

BACKTRACKING LINESEARCH

Procedure to find the stepsize α_k:

Given $\alpha_{\text{init}} > 0$ (e.g., $\alpha_{\text{init}} = 1$)
let $\alpha^{(0)} = \alpha_{\text{init}}$ and $l = 0$
Until $f(x_k + \alpha^{(l)} d_k)$ "<" f_k
  set $\alpha^{(l+1)} = \tau\alpha^{(l)}$, where $\tau \in (0, 1)$ (e.g., $\tau = \frac{1}{2}$)
  and increase l by 1
Set $\alpha_k = \alpha^{(l)}$

- this prevents the step from getting too small ... but does not prevent too large steps relative to decrease in f
- need to tighten requirement

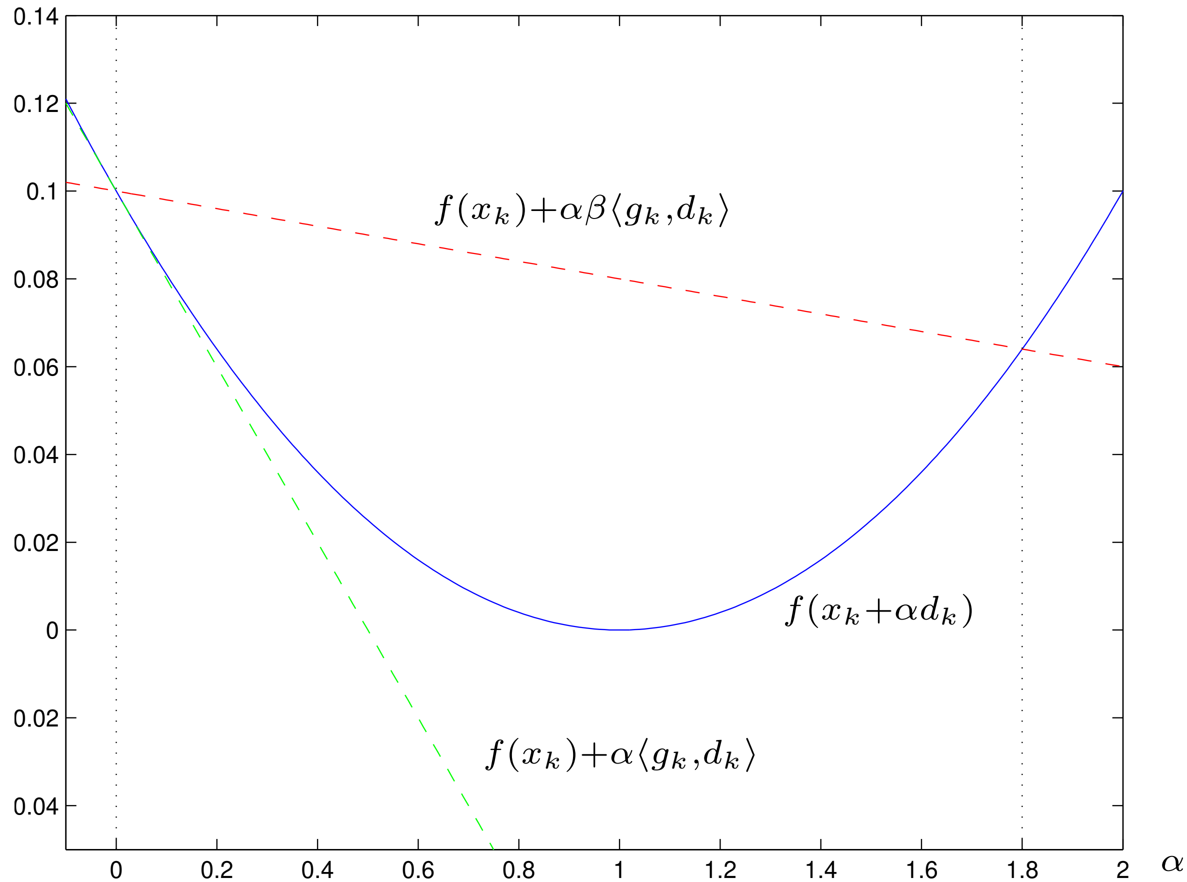
$$f(x_k + \alpha^{(l)} d_k) \text{“}<\text{”} f_k$$

ARMIJO CONDITION

In order to prevent large steps relative to decrease in f, instead require

$$f(x_k + \alpha_k d_k) \leqslant f(x_k) + \beta \alpha_k \langle g_k, d_k \rangle$$

for some $\beta \in (0, 1)$ (e.g., $\beta = 0.1$ or even $\beta = 0.0001$)



BACKTRACKING-ARMIJO LINESEARCH

Procedure to find the stepsize α_k:

Given $\alpha_{\text{init}} > 0$ (e.g., $\alpha_{\text{init}} = 1$)
let $\alpha^{(0)} = \alpha_{\text{init}}$ and $l = 0$
Until $f(x_k + \alpha^{(l)} d_k) \leqslant f(x_k) + \beta \alpha^{(l)} \langle g_k, d_k \rangle$
    set $\alpha^{(l+1)} = \tau \alpha^{(l)}$, where $\tau \in (0,1)$ (e.g., $\tau = \frac{1}{2}$)
    and increase l by 1
Set $\alpha_k = \alpha^{(l)}$

SATISFYING THE ARMIJO CONDITION

Theorem 2.4. Suppose that $f \in C^1$, that $g(x)$ is Lipschitz continuous with Lipschitz constant $\gamma(x)$, that $\beta \in (0,1)$ and that d is a descent direction at x. Then the Armijo condition

$$f(x + \alpha d) \leqslant f(x) + \alpha\beta\langle g(x), d\rangle$$

is satisfied for all $\alpha \in [0, \alpha_{\max(x)}]$, where

$$\alpha_{\max} = \frac{2(\beta - 1)\langle g(x), d\rangle}{\gamma(x)\|d\|_2^2}$$

THE ARMIJO LINESEARCH TERMINATES

Corollary 2.5. Suppose that $f \in C^1$, that $g(x)$ is Lipschitz continuous with Lipschitz constant γ_k at x_k, that $\beta \in (0,1)$ and that d_k is a descent direction at x_k. Then the stepsize generated by the backtracking-Armijo linesearch terminates with

$$\alpha_k \geqslant \min\left(\alpha_{\text{init}}, \frac{2\tau(\beta-1)\langle g_k, d_k\rangle}{\gamma_k \|d_k\|_2^2}\right)$$

GENERIC LINESEARCH METHOD

Given an initial guess x_0, let $k = 0$

Until convergence:

Find a descent direction d_k at x_k

Compute a stepsize α_k using a backtracking-Armijo linesearch along d_k

Set $x_{k+1} = x_k + \alpha_k d_k$, and increase k by 1

GLOBAL CONVERGENCE THEOREM

Theorem 2.6. Suppose that $f \in C^1$ and that g is Lipschitz continuous on $\mathbb{R}^n$. Then, for the iterates generated by the Generic Linesearch Method,

either

$$g_l = 0 \;\text{ for some }\; l \geqslant 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

or

$$\lim_{k \to \infty} \min\left(|\langle d_k, g_k \rangle|, \frac{|\langle d_k, g_k \rangle|}{\|d_k\|_2} \right) = 0.$$

METHOD OF STEEPEST DESCENT

The search direction

$$d_k = -g_k$$

gives the so-called **steepest-descent** direction.

- d_k is a descent direction
- d_k solves the problem

$$\begin{array}{ll} \underset{d \in \mathbb{R}^n}{\text{minimize}} & m_k^L(x_k + d) := f_k + \langle g_k, d \rangle \\ \text{subject to} & \|d\|_2 = \|g_k\|_2 \end{array}$$

Any method that uses the steepest-descent direction is a **method of steepest descent**.

GLOBAL CONVERGENCE FOR STEEPEST DESCENT

Theorem 2.7. Suppose that $f \in C^1$ and that g is Lipschitz continuous on IR^n. Then, for the iterates generated by the Generic Linesearch Method using the steepest-descent direction,

either

$$g_l = 0 \ \text{ for some } \ l \geqslant 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

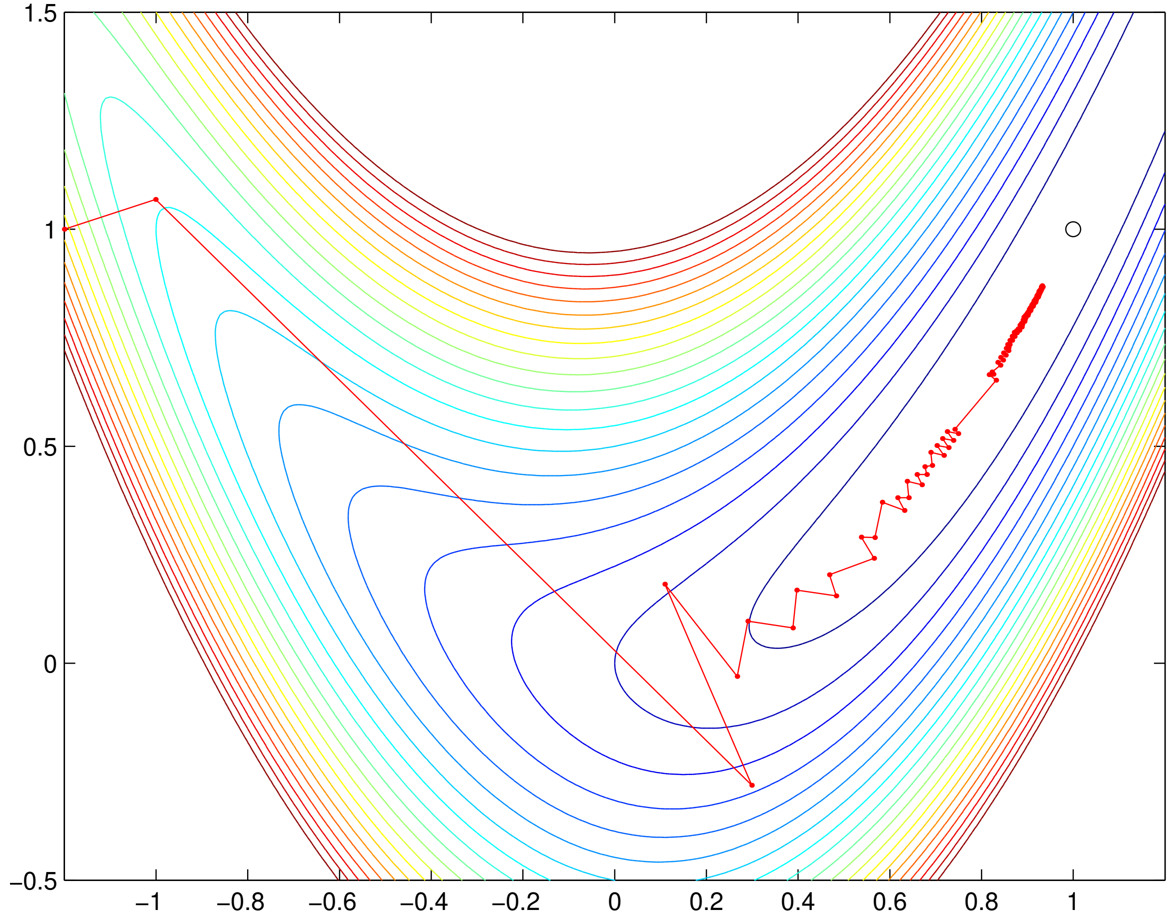
or

$$\lim_{k \to \infty} g_k = 0.$$

METHOD OF STEEPEST DESCENT (cont.)

- archetypical globally convergent method
- many other methods resort to steepest descent in bad cases
- not scale invariant
- convergence is usually very (very!) slow (linear)
- numerically often not convergent at all

STEEPEST DESCENT EXAMPLE



Contours for the objective function $f(x, y) = 10(y - x^2)^2 + (x - 1)^2$, and the iterates generated by the Generic Linesearch steepest-descent method

MORE GENERAL DESCENT METHODS

Let B_k be a **symmetric, positive definite** matrix, and define the search direction d_k so that

$$B_k d_k = -g_k$$

Then

- d_k is a descent direction as $\langle g_k, d_k \rangle = -\langle d_k, B_k d_k \rangle < 0$
- d_k solves the problem

$$\underset{d \in \mathbb{R}^n}{\text{minimize}} \;\; m_k^Q(x_k + d) := f_k + \langle g_k, d \rangle + \tfrac{1}{2} \langle d, B_k d \rangle$$

- if the Hessian H_k is positive definite, and $B_k = H_k$, this is **Newton's method**

MORE GENERAL GLOBAL CONVERGENCE

Theorem 2.8. Suppose that $f \in C^1$ and that g is Lipschitz continuous on $\mathbb{R}^n$. Then, for the iterates generated by the Generic Linesearch Method using the more general descent direction, either

$$g_l = 0 \text{ for some } l \geqslant 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

or

$$\lim_{k \to \infty} g_k = 0$$

provided that the eigenvalues of B_k are uniformly bounded and bounded away from zero.

MORE GENERAL DESCENT METHODS (cont.)

- may be viewed as “scaled” steepest descent
- convergence is often faster than steepest descent
- can be made scale invariant for suitable B_k

CONVERGENCE OF NEWTON'S METHOD

Theorem 2.9. Suppose that $f \in C^2$ and that H is Lipschitz continuous on IR^n. Then suppose that the iterates generated by the Generic Linesearch Method with $\alpha_{\text{init}} = 1$ and $\beta < \frac{1}{2}$, in which the search direction is chosen to be the Newton direction $d_k = -H_k^{-1} g_k$ whenever possible, has a limit point x_* for which $H(x_*)$ is positive definite. Then

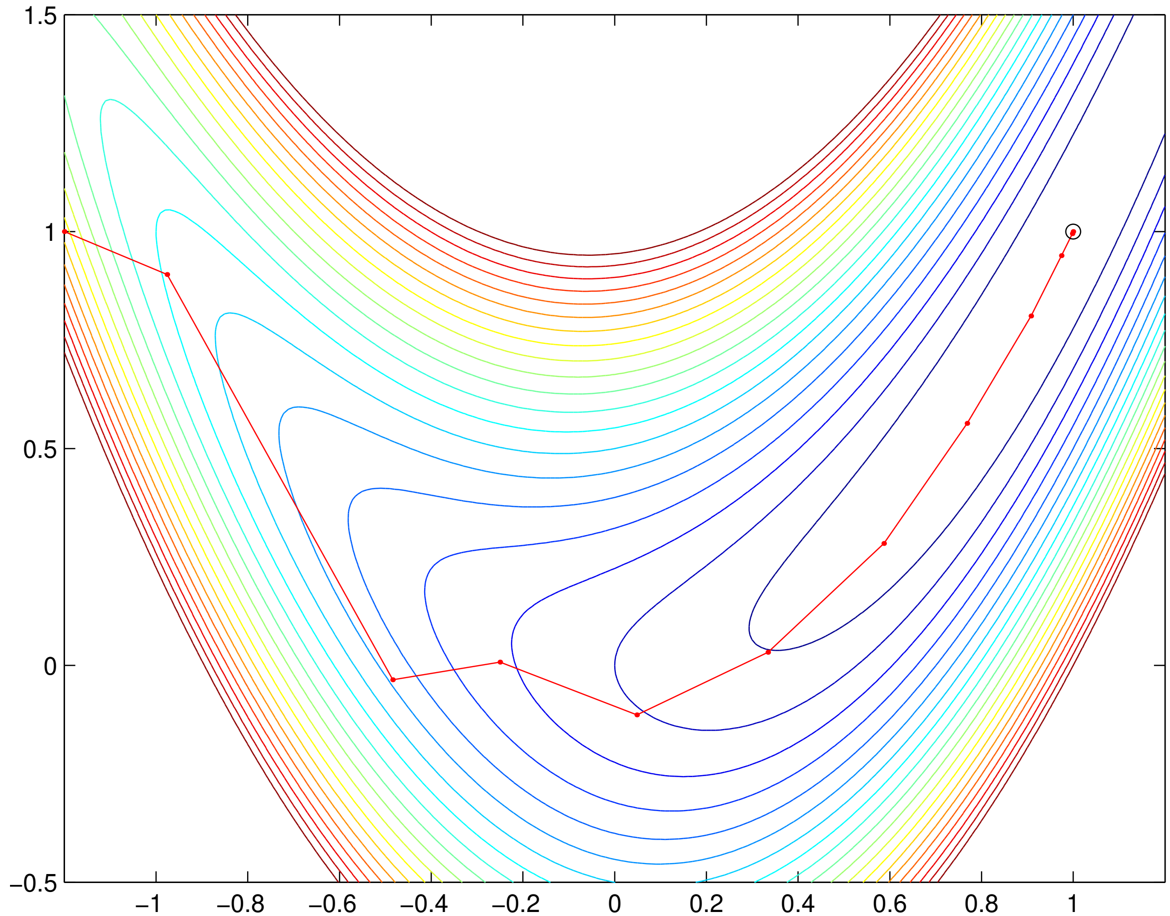
(i) $\alpha_k = 1$ for all sufficiently large k,

(ii) the entire sequence $\{x_k\}$ converges to x_*, and

(iii) the rate is Q-quadratic, i.e, there is a constant $\kappa \geqslant 0$.

$$\lim_{k \to \infty} \frac{\|x_{k+1} - x_*\|_2}{\|x_k - x_*\|_2^2} \leqslant \kappa.$$

NEWTON METHOD EXAMPLE



Contours for the objective function $f(x,y) = 10(y - x^2)^2 + (x-1)^2$, and the iterates generated by the Generic Linesearch Newton method

MODIFIED NEWTON METHODS

If H_k is indefinite, it is usual to solve instead

$$(H_k + M_k)d_k \equiv B_k d_k = -g_k$$

where

- M_k chosen so that $B_k = H_k + M_k$ is "sufficiently" positive definite
- $M_k = 0$ when H_k is itself "sufficiently" positive definite

Possibilities:

- If H_k has the spectral decomposition $H_k = V_k^T \Lambda_k V_k$ then

$$B_k \equiv H_k + M_k = V_k^T \max(\epsilon, |\Lambda_k|) V_k$$

- $M_k = \max(0, \epsilon - \lambda_{\min}(H_k))I$
- **Modified Cholesky**: $B_k \equiv H_k + M_k = L_k L_k^T$

QUASI-NEWTON METHODS

Various attempts to approximate H_k:

1. **Finite-difference** approximations:

$$(H_k)e_i \approx \frac{g(x_k + he_i) - g_k}{h} = (B_k)e_i$$

for some "small" scalar $h > 0$

- needs n evaluations of g to get H, fewer if sparse
- may need to symmetrize $H_k = \frac{1}{2}(H_k + H_k^T)$
- obviously parallel

QUASI-NEWTON METHODS (continued)

2. **Secant** approximations: try to ensure the **secant condition**

$$B_{k+1}s_k = y_k, \quad \text{where} \quad s_k = x_{k+1} - x_k \quad \text{and} \quad y_k = g_{k+1} - g_k$$

Why? Because $H_k s_k = y_k$ when f is quadratic

Examples:

- **Symmetric Rank-1 method** (but may be indefinite or even fail):

$$B_{k+1} = B_k + \frac{(y_k - B_k s_k)(y_k - B_k s_k)^T}{\langle y_k - B_k s_k, s_k \rangle}$$

- **BFGS method**: (symmetric and positive definite if $\langle y_k, s_k \rangle > 0$):

$$B_{k+1} = B_k + \frac{y_k y_k^T}{\langle y_k, s_k \rangle} - \frac{B_k s_k s_k^T B_k}{\langle s_k, B_k s_k \rangle}$$

Generally a low-rank (rank-one or -two) update of the existing B_k

LIMITED-MEMORY METHODS

Quasi-Newton methods pick

$$B_{k+1} = B_k + \text{low-rank matrix combination}(y_k, s_k, B_k) \;\; \text{where}$$
$$s_k = x_{k+1} - x_k \;\; \text{and} \;\; y_k = g_{k+1} - g_k$$

$\Longrightarrow$

$$B_{k+1} = B_0 + \text{matrix combination}(y_1, \ldots, y_k, s_1, \ldots, s_k, B_0)$$

Limited-memory methods pick

$$B_{k+1} = B_j + \text{matrix combination}(y_{j+1}, \ldots, y_k, s_{j+1}, \ldots, s_k, B_{j+1})$$

for some j close to k

- re-initialize using simple B_j (e.g $B_j = I \Longrightarrow B_{k+1}$ is a low-rank modification of B_j using data $\{y_{j+1}, \ldots, y_k, s_{j+1}, \ldots, s_k\}$
- efficient formulae to compute $d_{k+1} = -B_{k+1}^{-1} g_{k+1}$
- **L-BFGS** using BFGS formula

USE CG TO MINIMIZE CONVEX QUADRATIC MODEL

For convex models (B_k positive definite)

$$d_k = \text{(approximate)} \arg\min_{d \in \mathbb{R}^n} m_k^Q(x_k + d) f_k + \langle g_k, d \rangle + \tfrac{1}{2}\langle d, B_k d \rangle$$

Can apply conjugate-gradients method to minimize

$$q(d) = m_k^Q(x_k + d)$$

Stop CG when

$$\|\nabla q(d_k)\| \leqslant \min(\|g_k\|^\omega, \eta)\|g_k\| \quad (0 < \eta, \omega < 1)$$

$\Longrightarrow$ fast convergence

NONLINEAR CONJUGATE-GRADIENT METHODS

method for minimizing quadratic $f(x)$

Given x_0 and $g(x_0)$, set $p_0 = -g(x_0)$ and $i = 0$.
Until $g(x_k)$ "small" iterate

$$\alpha_i = \arg\min_{\alpha} f(x_i + \alpha p_i)$$
$$x_{i+1} = x_i + \alpha_i p_i$$
$$\beta_i = \|g(x_{i+1})\|_2^2 / \|g(x_i)\|_2^2$$
$$p_{i+1} = -g(x_{i+1}) + \beta_i p_i$$

and increase i by 1

may also be used for nonlinear $f(x)$ (Fletcher & Reeves)

- replace calculation of α_i by suitable linesearch
- other methods pick different β_i to ensure descent (Polyak–Ribière, Hestenes–Stiefel, Hager–Zhang ...)

Part 2b: Trust-region methods for unconstrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x)$$

Course on continuous optimization, STFC-RAL, February 2021

UNCONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x)$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$

- assume that $f \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

LINESEARCH VS TRUST-REGION METHODS

- **Linesearch methods**
  - pick descent direction d_k
  - pick stepsize α_k to "reduce" $f(x_k + \alpha d_k)$
  - $x_{k+1} = x_k + \alpha_k d_k$
- **Trust-region methods**
  - pick step s_k to reduce "model" of $f(x_k + s)$
  - accept $x_{k+1} = x_k + s_k$ if decrease in model inherited by $f(x_k + s_k)$
  - otherwise set $x_{k+1} = x_k$, "refine" model

TRUST-REGION MODEL PROBLEM

Model $f(x_k + s)$ by:

- linear model

$$m_k^L(s) = f_k + \langle s, g_k \rangle$$

- quadratic model — symmetric B_k

$$m_k^Q(s) = f_k + \langle g_k, s \rangle + \tfrac{1}{2}\langle s, B_k s \rangle$$

Major difficulties:

- models may not resemble $f(x_k + s)$ if s is large
- models may be unbounded from below
  - linear model - always unless $g_k = 0$
  - quadratic model - always if B_k is indefinite, possibly if B_k is only positive semi-definite

THE TRUST REGION

Prevent model $m_k(s)$ from unboundedness by imposing a **trust-region** constraint

$$\|s\| \leqslant \Delta_k$$

for some "suitable" scalar **radius** $\Delta_k > 0$

$\Longrightarrow$ **trust-region subproblem**

$$\text{approx} \quad \underset{s \in \mathbb{R}^n}{\text{minimize}} \; m_k(s) \; \text{ subject to } \; \|s\| \leqslant \Delta_k$$

- in theory does not depend on norm $\| \cdot \|$
- in practice it might!

OUR MODEL

For simplicity, concentrate on the second-order (Newton-like) model

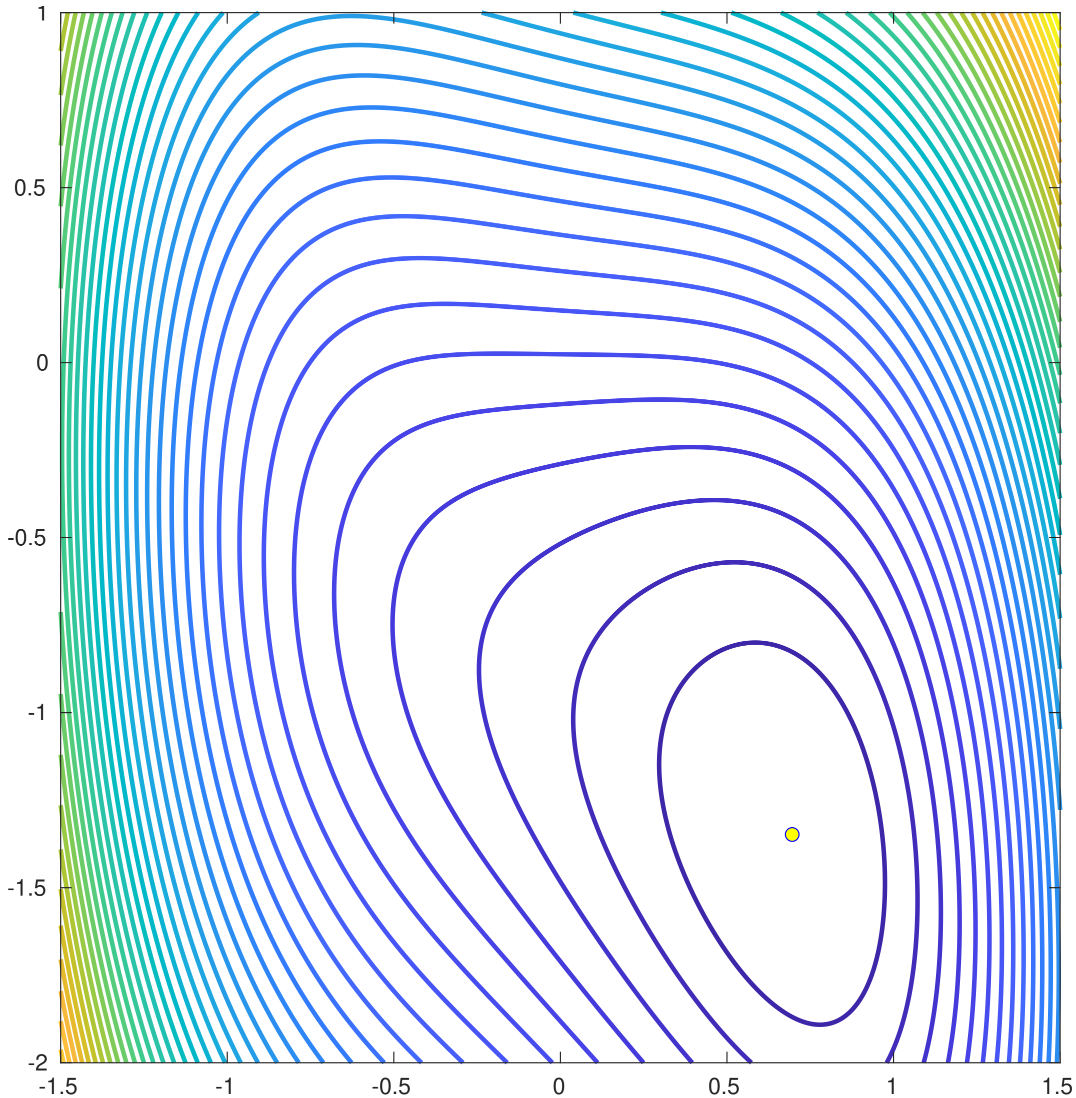
$$m_k(s) = m_k^Q(s) = f_k + \langle s, g_k \rangle + \tfrac{1}{2}\langle s, B_k s \rangle$$

and the ℓ_2-trust region norm $\| \cdot \| = \| \cdot \|_2$

Note:

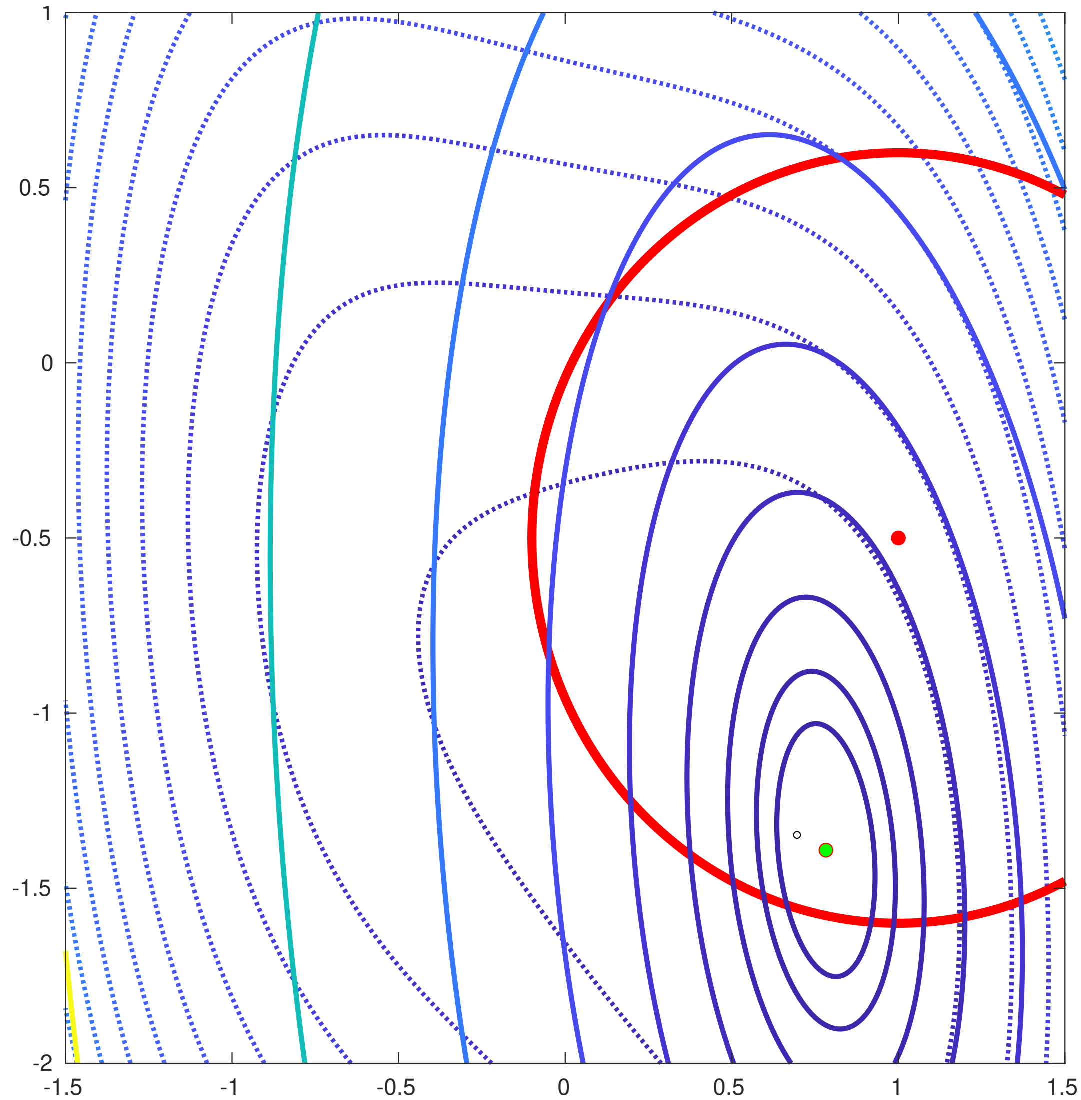
- $B_k = H_k$ is allowed
- analysis for other trust-region norms simply adds extra constants in following results

TRUST-REGION EXAMPLES



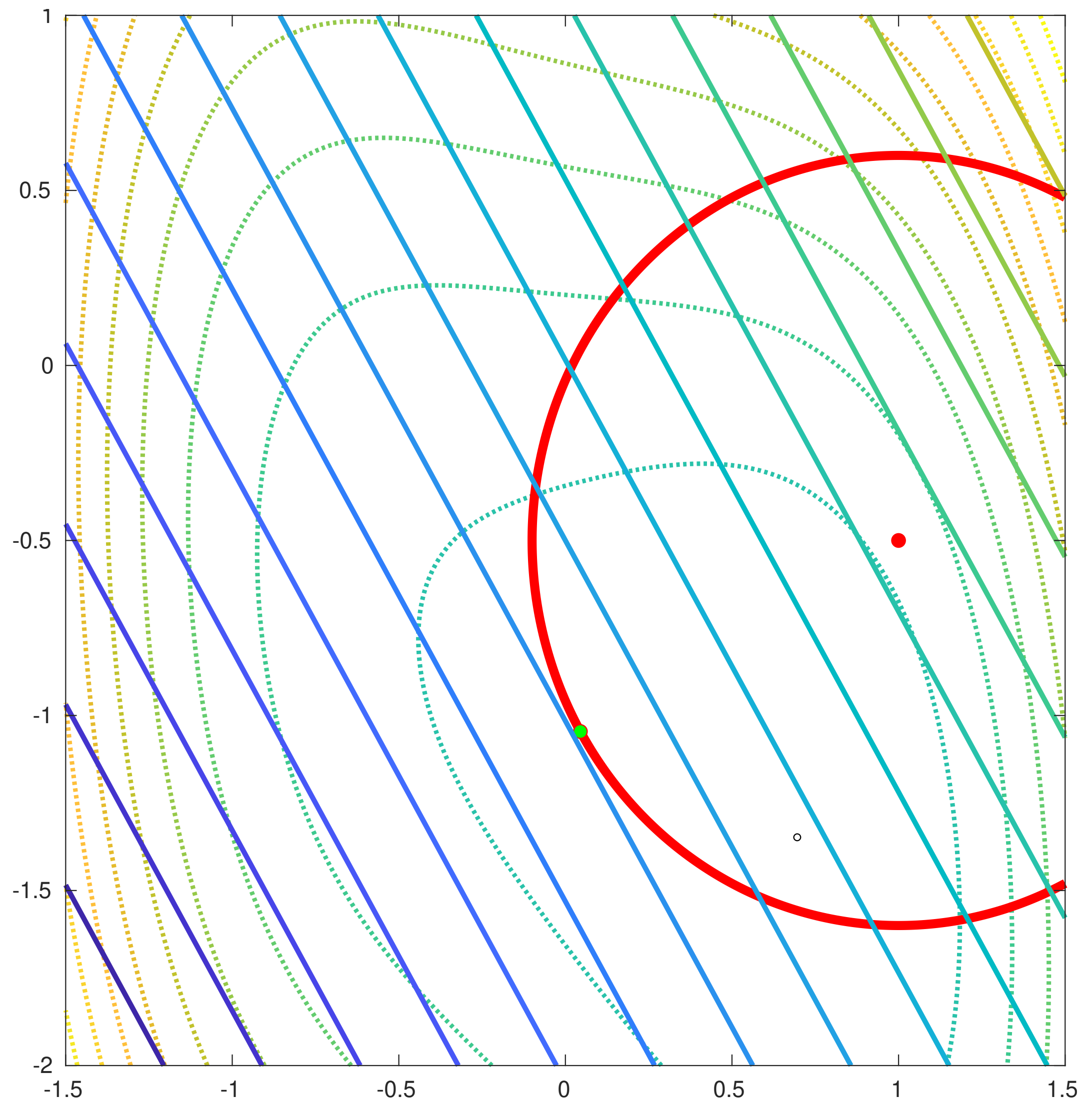
Contours for the objective function $f(x, y) = x^4 + xy + (y + 1)^2$

TRUST-REGION EXAMPLES (cont)



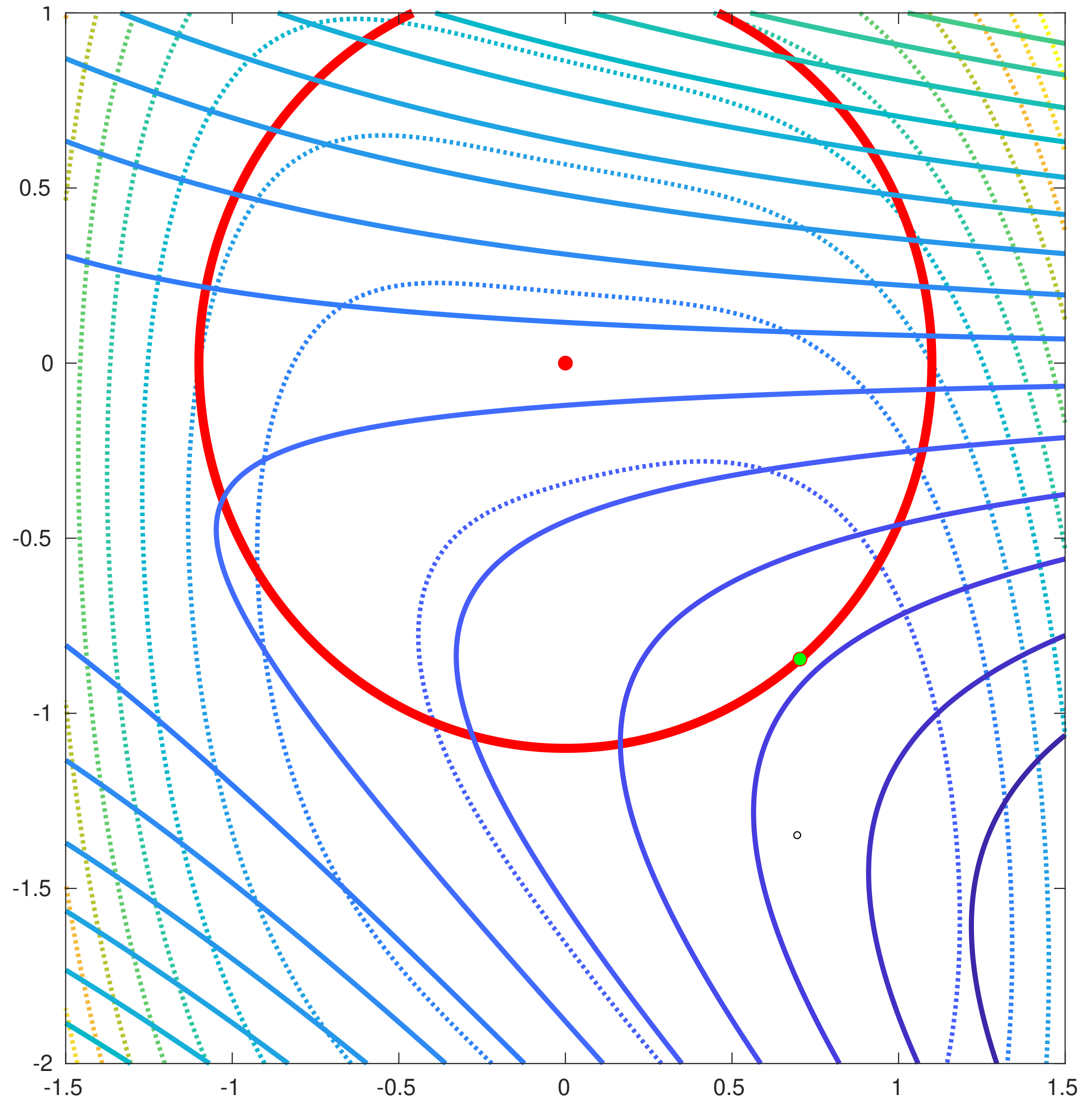
Contours of quadratic model $m_k(s)$ at $(1, -0.5)$ with radius $\Delta = 1.1$

TRUST-REGION EXAMPLES (cont)



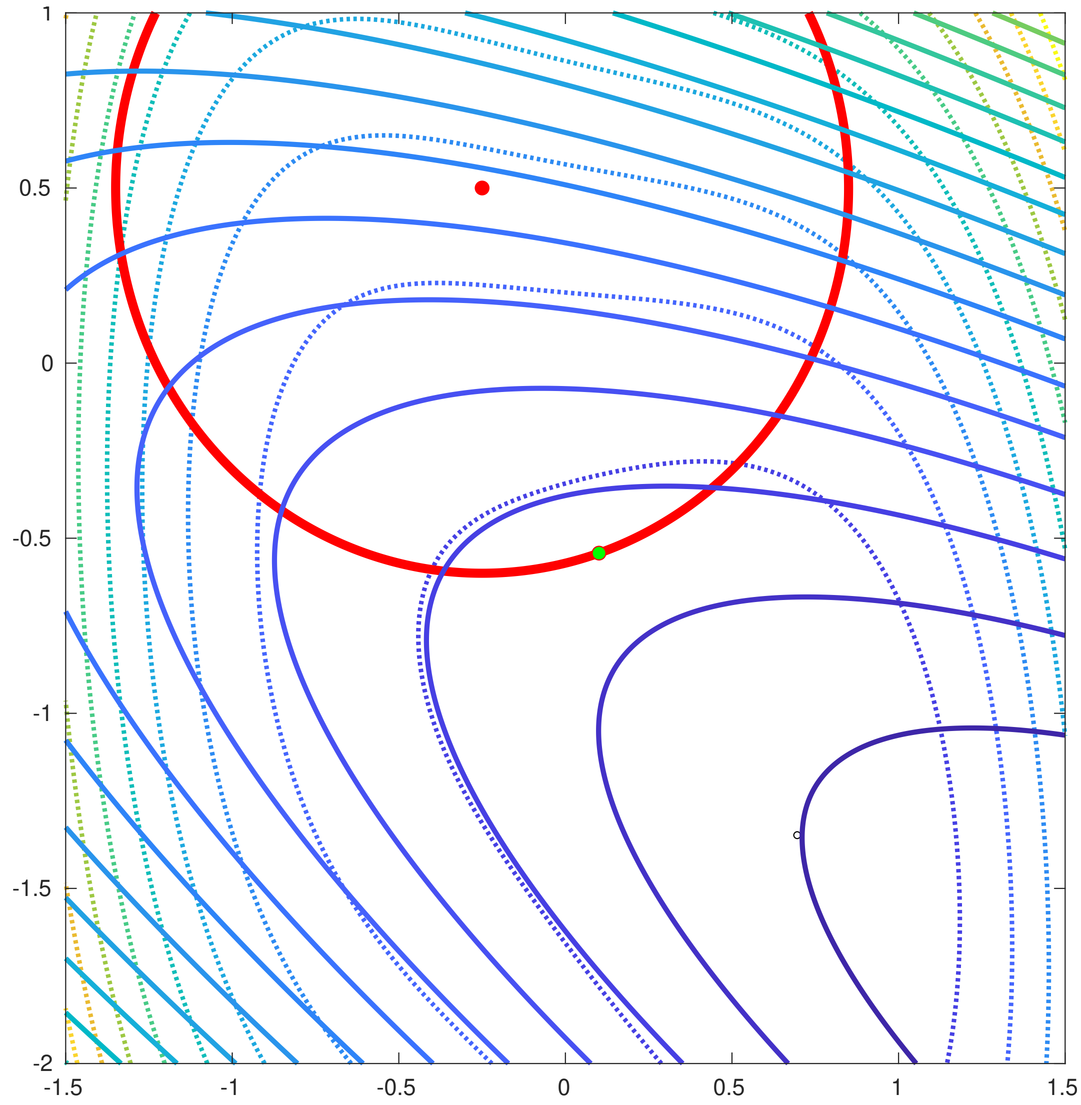
Contours of linear model $m_k(s)$ at $(1, -0.5)$ with radius $\Delta = 1.1$

TRUST-REGION EXAMPLES (cont)



Contours of quadratic model $m_k(s)$ at $(0,0)$ with radius $\Delta = 1.1$

TRUST-REGION EXAMPLES (cont)



Contours of quadratic model $m_k(s)$ at $(-0.25, 0.5)$ with radius $\Delta = 1.1$

BASIC TRUST-REGION METHOD

Given $k = 0$, $\Delta_0 > 0$ and x_0, until "convergence" do:

Build the second-order model $m_k(s)$ of $f(x_k + s)$.

"Solve" the trust-region subproblem to find s_k

for which $m_k(s_k)$ "<" f_k and $\|s_k\| \leqslant \Delta_k$, and define

$$\rho_k = \frac{f_k - f(x_k + s_k)}{f_k - m_k(s_k)}.$$

If $\rho_k \geqslant \eta_v$ [**very successful**] $\boxed{0 < \eta_v < 1}$

set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \gamma_i \Delta_k$ $\boxed{\gamma_i \geqslant 1}$

Otherwise if $\rho_k \geqslant \eta_s$ then [**successful**] $\boxed{0 < \eta_s \leqslant \eta_v < 1}$

set $x_{k+1} = x_k + s_k$ and $\Delta_{k+1} = \Delta_k$

Otherwise [**unsuccessful**]

set $x_{k+1} = x_k$ and $\Delta_{k+1} = \gamma_d \Delta_k$ $\boxed{0 < \gamma_d < 1}$

Increase k by 1

“SOLVE” THE TRUST REGION SUBPROBLEM?

At the very least

- aim to achieve as much reduction in the model as would an iteration of steepest descent
- **Cauchy point**: $s_k^{\mathrm{C}} = -\alpha_k^{\mathrm{C}} g_k$ where

$$\begin{aligned} \alpha_k^{\mathrm{C}} &= \arg\min_{\alpha > 0} m_k(-\alpha g_k) \;\text{ subject to }\; \alpha\|g_k\| \leqslant \Delta_k \\ &= \arg\min_{0 < \alpha \leqslant \Delta_k/\|g_k\|} m_k(-\alpha g_k) \end{aligned}$$

  - minimize 1-D quadratic on line segment $\Longrightarrow$ very easy!
- require that

$$m_k(s_k) \leqslant m_k(s_k^{\mathrm{C}}) \;\text{ and }\; \|s_k\| \leqslant \Delta_k$$

- in practice, hope to do far better than this

ACHIEVABLE MODEL DECREASE

Theorem 2.10. If $m_k(s)$ is the second-order model and s_k^{C} is its Cauchy point within the trust-region $\|s\| \leqslant \Delta_k$,

$$f_k - m_k(s_k^{\mathrm{C}}) \geqslant \tfrac{1}{2}\|g_k\| \min\left[\frac{\|g_k\|}{1+\|B_k\|}, \Delta_k\right].$$

Corollary 2.11. If $m_k(s)$ is the second-order model, and s_k is an improvement on the Cauchy point within the trust-region $\|s\| \leqslant \Delta_k$,

$$f_k - m_k(s_k) \geqslant \tfrac{1}{2}\|g_k\| \min\left[\frac{\|g_k\|}{1+\|B_k\|}, \Delta_k\right].$$

DIFFERENCE BETWEEN MODEL AND FUNCTION

Lemma 2.12. Suppose that $f \in C^2$, and that the true and model Hessians satisfy the bounds $\|H(x)\| \leqslant \kappa_h$ for all x and $\|B_k\| \leqslant \kappa_b$ for all k and some $\kappa_h \geqslant 1$ and $\kappa_b \geqslant 0$. Then

$$|f(x_k + s_k) - m_k(s_k)| \leqslant \kappa_d \Delta_k^2,$$

where $\kappa_d = \frac{1}{2}(\kappa_h + \kappa_b)$, for all k.

ULTIMATE PROGRESS AT NON-OPTIMAL POINTS

Lemma 2.13. Suppose that $f \in C^2$, that the true and model Hessians satisfy the bounds $\|H_k\| \leqslant \kappa_h$ and $\|B_k\| \leqslant \kappa_b$ for all k and some $\kappa_h \geqslant 1$ and $\kappa_b \geqslant 0$. Suppose furthermore that $g_k \neq 0$ and that

$$\Delta_k \leqslant \left(\frac{1 - \eta_v}{\kappa_h + \kappa_b} \right) \|g_k\|.$$

Then iteration k is very successful and

$$\Delta_{k+1} \geqslant \Delta_k.$$

RADIUS WON'T SHRINK TO ZERO AT NON-OPTIMAL POINTS

Lemma 2.14. Suppose that $f \in C^2$, that the true and model Hessians satisfy the bounds $\|H_k\| \leqslant \kappa_h$ and $\|B_k\| \leqslant \kappa_b$ for all k and some $\kappa_h \geqslant 1$ and $\kappa_b \geqslant 0$. Suppose furthermore that there is a constant $\epsilon > 0$ such that

$$\|g_k\| \geqslant \epsilon \;\text{ for all } k.$$

Then

$$\Delta_k \geqslant \kappa_\epsilon \;\text{ where }\; \kappa_\epsilon := \epsilon \gamma_d \left(\frac{1 - \eta_v}{\kappa_h + \kappa_b} \right)$$

for all k.

POSSIBLE FINITE TERMINATION

Lemma 2.15. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Suppose furthermore that there are only finitely many successful iterations. Then $x_k = x_*$ for all sufficiently large k and $g(x_*) = 0$.

GLOBAL CONVERGENCE OF ONE SEQUENCE

Theorem 2.16. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Then either

$$g_l = 0 \quad \text{for some} \quad l \geqslant 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

or

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

GLOBAL CONVERGENCE

Theorem 2.17. Suppose that $f \in C^2$, and that both the true and model Hessians remain bounded for all k. Then either

$$g_l = 0 \ \text{ for some } \ l \geqslant 0$$

or

$$\lim_{k \to \infty} f_k = -\infty$$

or

$$\lim_{k \to \infty} g_k = 0.$$

II: SOLVING THE TRUST-REGION SUBPROBLEM

$$\text{(approximately) } \underset{s \in \mathbb{R}^n}{\text{minimize}} \;\; q(s) \equiv \langle g, s \rangle + \tfrac{1}{2}\langle s, Bs \rangle \;\; \text{subject to} \;\; \|s\| \leqslant \Delta$$

AIM: find s_* so that

$$q(s_*) \leqslant q(s^{\mathrm{C}}) \;\; \text{and} \;\; \|s_*\| \leqslant \Delta$$

Might solve

- exactly $\Longrightarrow$ Newton-like method
- approximately $\Longrightarrow$ steepest descent/conjugate gradients

THE ℓ_2-NORM TRUST-REGION SUBPROBLEM

$$\underset{s \in \mathbb{R}^n}{\text{minimize}} \;\; q(s) \equiv \langle s, g \rangle + \tfrac{1}{2} \langle s, Bs \rangle \;\; \text{subject to} \;\; \|s\|_2 \leqslant \Delta$$

Solution characterisation result:

Theorem 2.18. Any global minimizer s_* of $q(s)$ subject to $\|s\|_2 \leqslant \Delta$ satisfies the equation

$$(B + \lambda_* I) s_* = -g,$$

where $B + \lambda_* I$ is positive semi-definite,

$$\lambda_* \geqslant 0 \;\; \text{and} \;\; \lambda_* (\|s_*\|_2 - \Delta) = 0.$$

If $B + \lambda_* I$ is positive definite, s_* is unique.

ALGORITHMS FOR THE ℓ_2-NORM SUBPROBLEM

Two cases:

- B positive-semi definite and $Bs = -g$ satisfies $\|s\|_2 \leqslant \Delta$

  $\Longrightarrow \; s_* = s$

- B indefinite or $Bs = -g$ satisfies $\|s\|_2 > \Delta$

  $\Longrightarrow$

  - $(B + \lambda_* I)s_* = -g$ and $\langle s_*, s_* \rangle = \Delta^2$
  - nonlinear (quadratic) system in s and λ
  - concentrate on this

EQUALITY CONSTRAINED ℓ_2-NORM SUBPROBLEM

Suppose B has spectral decomposition

$$B = V^T \Lambda V$$

- V orthogonal matrix of eigenvectors
- Λ diagonal matrix of eigenvalues: $\lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n$

Require $B + \lambda I = V^T(\Lambda + \lambda I)V$ positive semi-definite $\Longrightarrow \lambda \geqslant -\lambda_1$

Define

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

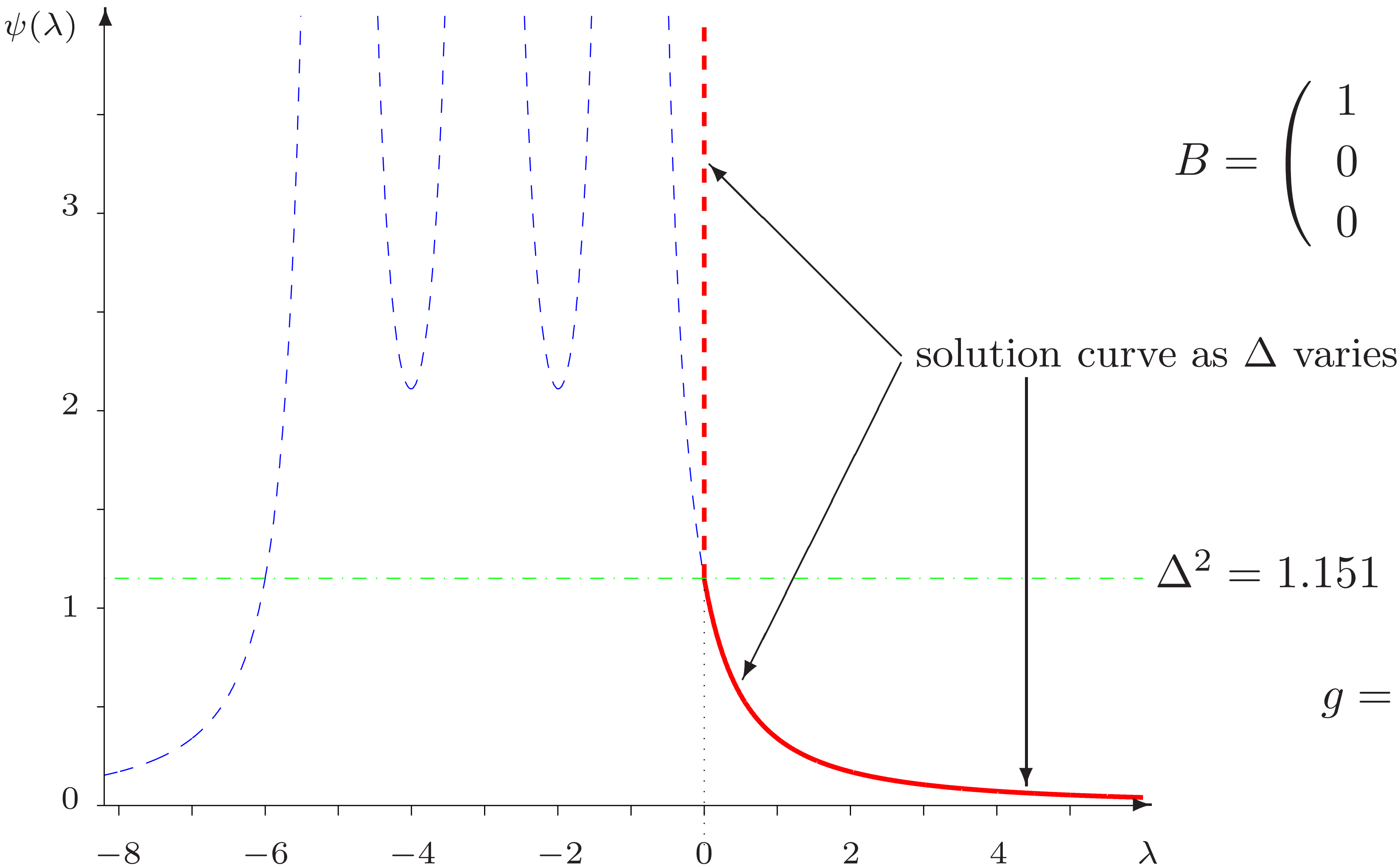
Require the **secular function**

$$\psi(\lambda) := \|s(\lambda)\|_2^2 = \Delta^2$$

Note $\qquad (\gamma_i = \langle e_i, V g \rangle)$

$$\psi(\lambda) = \|V^T(\Lambda + \lambda I)^{-1} V g\|_2^2 = \sum_{i=1}^{n} \frac{\gamma_i^2}{(\lambda_i + \lambda)^2}$$

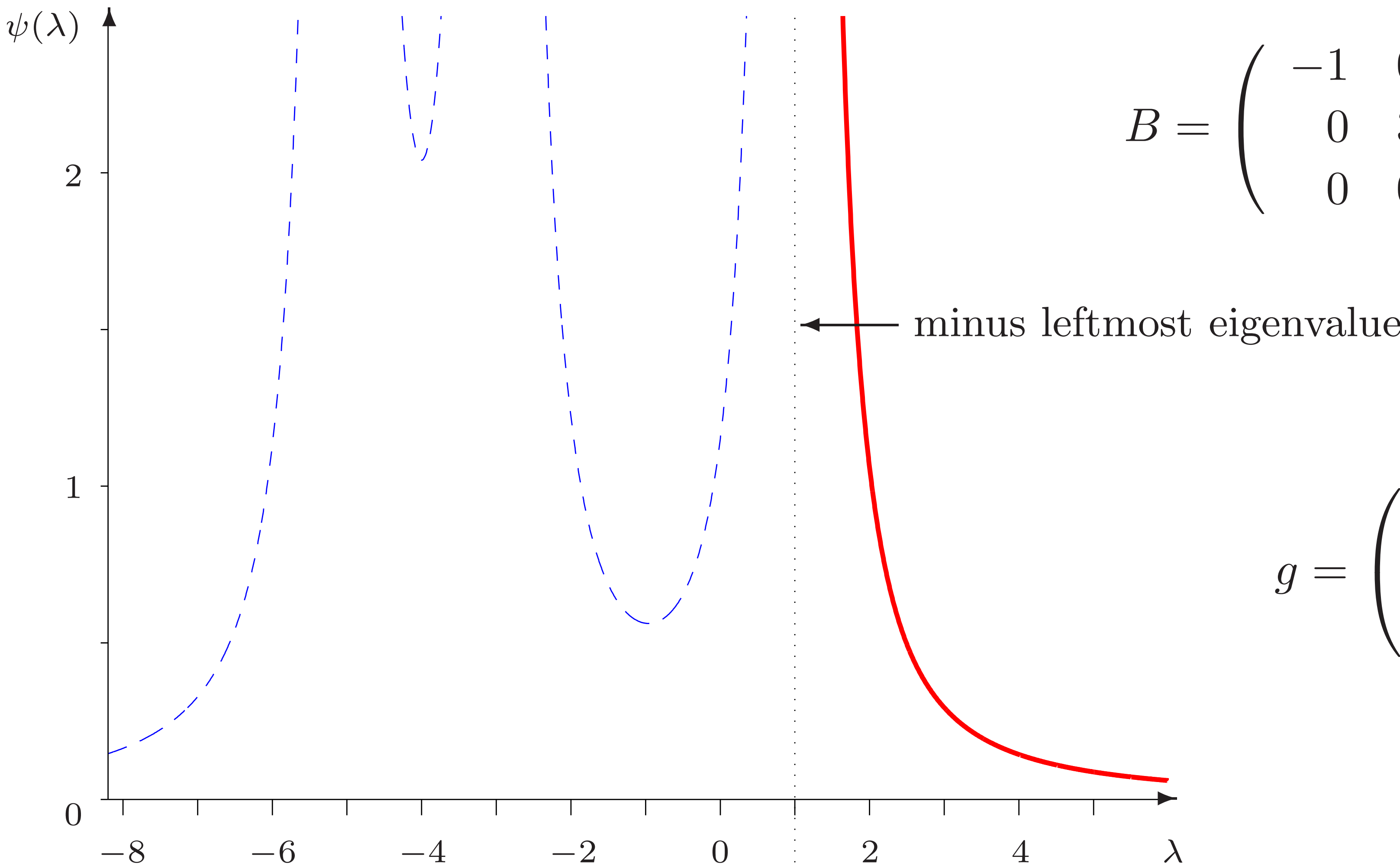
CONVEX EXAMPLE



$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

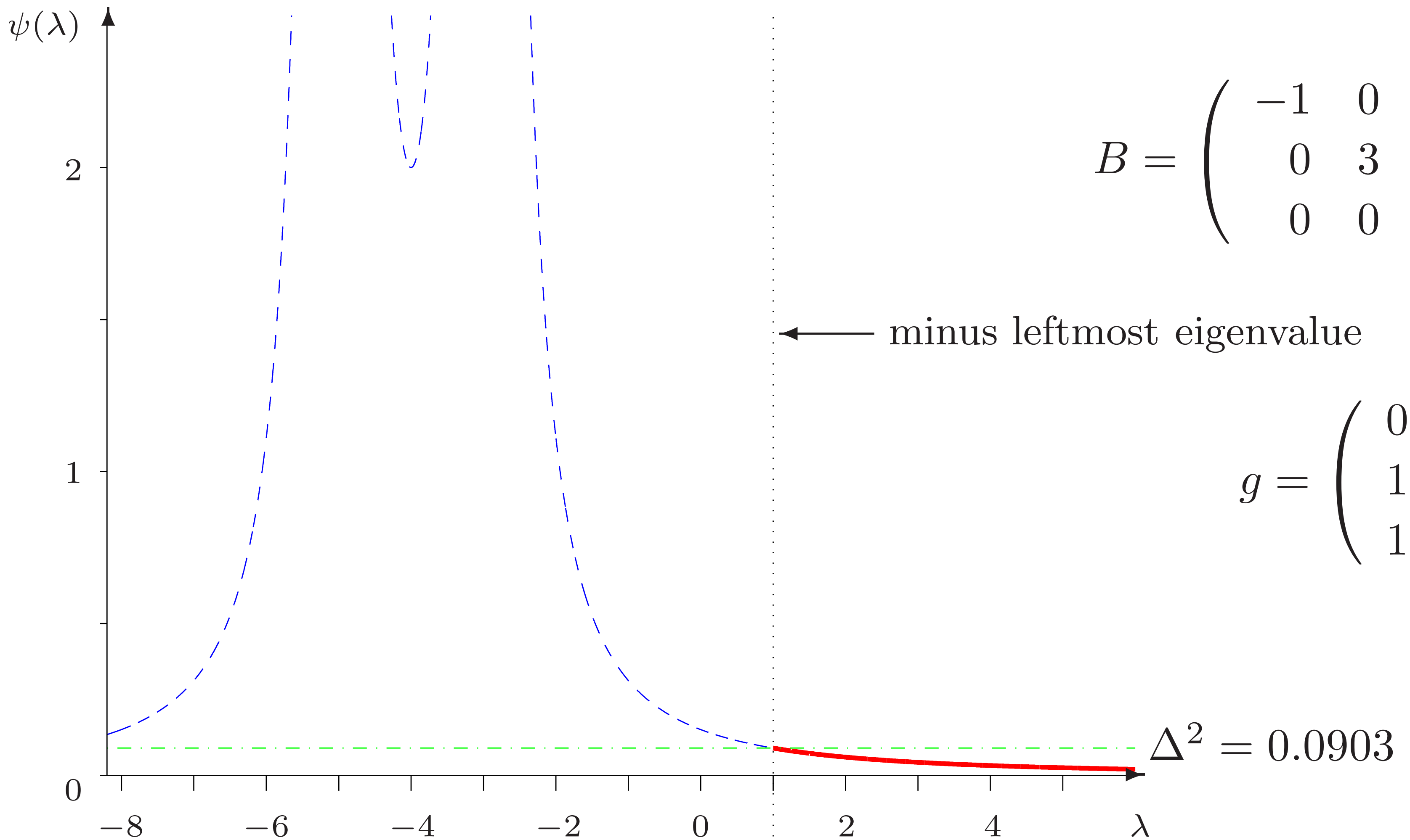
NONCONVEX EXAMPLE



$$B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

THE “HARD” CASE



$$B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$$

$$g = \begin{pmatrix} 0 \\ 1 \\ 1 \end{pmatrix}$$

SUMMARY

For indefinite B:

Hard case occurs when g orthogonal to eigenvector v_1 for most negative eigenvalue λ_1 and Δ "too large"

- OK if radius Δ is small enough
- No "obvious" solution to equations ... but solution is actually of the form

  $$s_{\text{lim}} + \sigma v_1$$

  where

  - $s_{\text{lim}} = \lim_{\lambda \xrightarrow{+} -\lambda_1} s(\lambda)$
  - $\|s_{\text{lim}} + \sigma v_1\|_2 = \Delta$
- very rare in practice ("probability 0" event)

HOW TO SOLVE $\|\mathbf{s}(\lambda)\|_2 = \Delta$

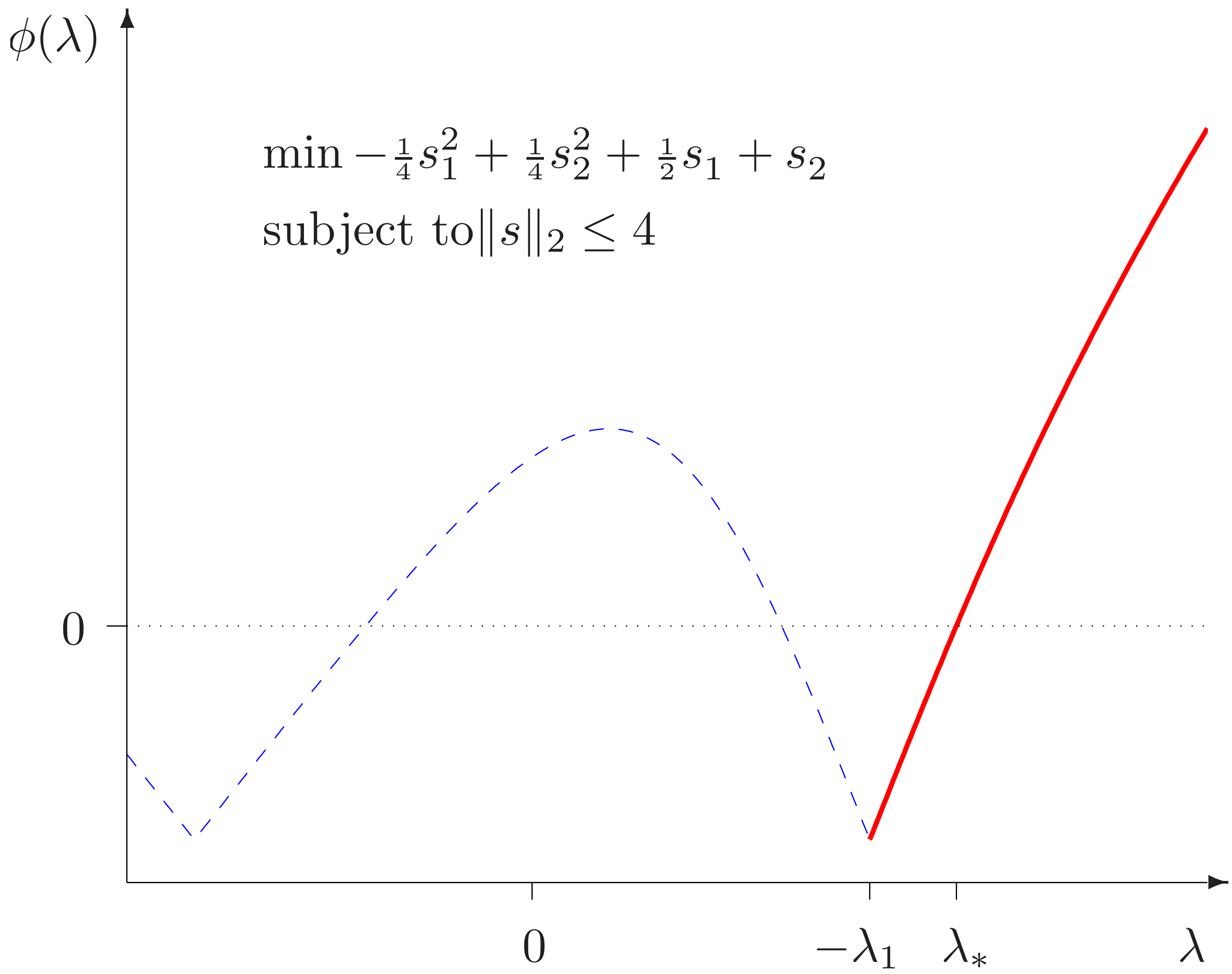
DON'T!!

Solve instead the **secular equation**

$$\phi(\lambda) := \frac{1}{\|s(\lambda)\|_2} - \frac{1}{\Delta} = 0$$

- no poles
- smallest at eigenvalues (except in hard case!)
- analytic function $\Longrightarrow$ ideal for Newton
- global convergent (ultimately quadratic rate except in hard case)
- need to safeguard to protect Newton from the hard & interior solution cases

THE SECULAR EQUATION



NEWTON'S METHOD & THE SECULAR EQUATION

Let $\lambda > -\lambda_1$ and $\Delta > 0$ be given

Until "convergence" do:

Factorize $B + \lambda I = LL^T$

Solve $LL^T s = -g$

Solve $Lw = s$

Replace λ by

$$\lambda + \left(\frac{\|s\|_2 - \Delta}{\Delta} \right) \left(\frac{\|s\|_2^2}{\|w\|_2^2} \right)$$

SOLVING THE LARGE-SCALE PROBLEM

- when n is large, factorization may be impossible
- may instead try to use an iterative method to approximate
  - steepest descent leads to the Cauchy point
  - obvious generalization: conjugate gradients . . . but
    - what about the trust region?
    - what about negative curvature $\langle s, Bs \rangle \leqslant 0$?

CONJUGATE GRADIENTS TO "MINIMIZE" q(s)

Set $s_0 = 0$, $g_0 = g$, $p_0 = -g$ and $i = 0$

Until g_i "small" or breakdown, iterate

$$\alpha_i = \|g_i\|_2^2 / \langle p_i, Bp_i \rangle$$
$$s_{i+1} = s_i + \alpha_i p_i$$
$$g_{i+1} = g_i + \alpha_i B p_i$$
$$\beta_i = \|g_{i+1}\|_2^2 / \|g_i\|_2^2$$
$$p_{i+1} = -g_{i+1} + \beta_i p_i$$

and increase i by 1

Important features

- $g_j = Bs_j + g$ for all $j = 0, \ldots, i$
- $\langle d_j, g_{i+1} \rangle = 0$ for all $j = 0, \ldots, i$
- $\langle g_j, g_{i+1} \rangle = 0$ for all $j = 0, \ldots, i$

CRUCIAL PROPERTY OF CONJUGATE GRADIENTS

Theorem 2.19. Suppose that the conjugate gradient method is applied to minimize $q(s)$ starting from $s_0 = 0$, and that

$$\langle p_i, Bp_i \rangle > 0 \;\text{ for } 0 \leqslant i \leqslant k.$$

Then the iterates s_j satisfy the inequalities

$$\|s_j\|_2 < \|s_{j+1}\|_2$$

for $0 \leqslant j \leqslant k - 1$.

TRUNCATED CONJUGATE GRADIENTS

Apply the conjugate gradient method, but terminate at iteration i if

1. $\langle d_i, B d_i \rangle \leqslant 0 \Longrightarrow$ problem unbounded along d_i
2. $\|s_i + \alpha_i d_i\|_2 > \Delta \Longrightarrow$ solution on trust-region boundary

In both cases, stop with $s_* = s_i + \alpha^{\text{B}} d_i$, where α^{B} chosen as positive root of

$$\|s_i + \alpha^{\text{B}} d_i\|_2 = \Delta$$

Crucially

$$q(s_*) \leqslant q(s^{\text{C}}) \text{ and } \|s_*\|_2 \leqslant \Delta$$

$\Longrightarrow$ TR algorithm converges to a first-order critical point

HOW GOOD IS TRUNCATED C.G.?

In the convex case ... very good

Theorem 2.20. Suppose that the truncated conjugate gradient method is applied to minimize $q(s)$ and that B is positive definite. Then the truncated and actual solutions to the problem, s_* and s_*^{M}, satisfy the bound

$$q(s_*) \leqslant \tfrac{1}{2} q(s_*^{\mathrm{M}})$$

In the non-convex case ... maybe poor

- e.g., if $g = 0$ and B is indefinite $\Longrightarrow$ $q(s_*) = 0$
- instead continue using equivalent Lanczos method to solve trust-region subproblem in subspace (GLTR method, see notes)

Part 2c: Miscellaneous methods for unconstrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ \tfrac{1}{2}\|c(x)\|_2^2$$

Course on continuous optimization, STFC-RAL, February 2021

AN ALTERNATIVE — CUBIC REGULARIZATION

Trust-region subproblem:

$$\text{(approx)} \underset{s \in \mathbb{R}^n}{\text{minimize}} \; f_k + \langle s, g_k \rangle + \tfrac{1}{2}\langle s, B_k s \rangle \;\text{ subject to }\; \|s\| \leqslant \Delta_k$$

for adjustable radius $\Delta_k > 0$

A modern alternative . . . the **cubic-regularization** subproblem:

$$\text{(approx)} \underset{s \in \mathbb{R}^n}{\text{minimize}} \; f_k + \langle s, g_k \rangle + \tfrac{1}{2}\langle s, B_k s \rangle + \tfrac{1}{3}\sigma_k \|s\|^3$$

for adjustable **weight** $\sigma_k > 0$

- can consider weight as "one over radius"
- solve regularization subproblem using related secular equation
- perform essentially the same in practice
- theoretical better worst-case behaviour

NONLINEAR LEAST-SQUARES

Given vector of **residuals** $c : \mathbb{R}^n \rightarrow \mathbb{R}^m$ find

$$\text{(approx) } \underset{x \in \mathbb{R}^n}{\text{minimize}} \; \|c(x)\|_2$$

Equivalent to the **smooth nonlinear least-squares** problem

$$\text{(approx) } \underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) = \tfrac{1}{2}\|c(x)\|_2^2$$

- the major use of unconstrained optimization
- model fitting to experimental data, e.g. $c_i(x) = r_i(x) - d_i$, where $r_i = r(x, p_i)$ and given parameters p_i
- $f(x)$ is bounded from below (by zero)

NOTATION

Use the following in what follows:

$$
\begin{array}{rcll}
a_i(x) & := & \nabla_x c_i(x) & \textbf{gradient} \text{ of } i\text{-th residual} \\
A(x) & := & [\nabla_x c^T(x)]^T \equiv \begin{pmatrix} a_1^T(x) \\ \cdots \\ a_m^T(x) \end{pmatrix} & \textbf{Jacobian matrix} \text{ of } c \\
H_i(x) & := & \nabla^2_{xx} c_i(x) & \textbf{Hessian} \text{ of } i\text{-th residual}
\end{array}
$$

DERIVATIVES OF THE LEAST-SQUARES FUNCTION

$$\text{(approx) } \underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) = \tfrac{1}{2}\|c(x)\|_2^2$$

- $g(x) = A^T(x)c(x)$
- $H(x) = A^T(x)A(x) + \sum_{i=1}^{m} c_i(x)H_i(x)$

Notice that

- if $c(x)$ is zero $\Longrightarrow H(x) = A^T(x)A(x)$
- if $c(x)$ is small $\Longrightarrow H(x) \approx A^T(x)A(x)$
- suggests using second-derivative models with $B_k = A_k^T A_k$

METHODS FOR NONLINEAR LEAST-SQUARES

$$\text{(approx) } \underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) = \tfrac{1}{2}\|c(x)\|_2^2$$

So long as c is twice-continuously differentiable, can use linesearch/trust-region/regularization method to minimize $f(x)$

Alternative: use **first-order Taylor model**

$$r_k(s) = c_k + A_k s$$

of the residual $c(x_k + s) \Longrightarrow$ **Gauss-Newton** model

$$\begin{aligned} m_k^{LS}(s) &= \tfrac{1}{2}\|r_k(s)\|_2^2 = \tfrac{1}{2}\|c_k + A_k s\|_2^2 \\ &= \tfrac{1}{2}\|c_k\|_2^2 + \langle s, A_k^T c_k \rangle + \tfrac{1}{2}\langle s, A_k^T A_k s \rangle \end{aligned}$$

of $f(x_k + s)$

METHODS FOR NONLINEAR LEAST-SQUARES (cont)

Gauss-Newton model:

$$\begin{aligned} m_k^{LS}(s) &= \tfrac{1}{2}\|r_k(s)\|_2^2 = \tfrac{1}{2}\|c_k + A_k s\|_2^2 \\ &= \tfrac{1}{2}\|c_k\|_2^2 + \langle s, A_k^T c_k \rangle + \tfrac{1}{2}\langle s, A_k^T A_k s \rangle \end{aligned}$$

- linesearch in direction d_k:

$$A_k^T A_k d_k = -A_k^T c_k$$

  - may fail if A_k is (or becomes) rank deficient

- trust-region imposes $\|s\| \leqslant \Delta_k$ implies implicitly

$$(A_k^T A_k + \lambda_k I) s_k = -A_k^T c_k$$

- $+$ quadratic regularization $\frac{1}{2}\sigma_k \|s\|_2^2$ implies explicitly

$$(A_k^T A_k + \sigma_k I) s_k = -A_k^T c_k$$

Last two are $\approx$ **Levenberg-Morrison-Marquardt** method

Part 3: Constrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \;\; \text{subject to} \;\; c(x) \left\{ \begin{array}{c} \geqslant \\ = \end{array} \right\} 0$$

Course on continuous optimization, STFC-RAL, February 2021

CONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) \left\{ \begin{array}{c} \geqslant \\ = \end{array} \right\} 0$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
and the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$

- assume that $f, \; c \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

CONTENT

We shall discuss:

- optimality conditions
- (gradient projection methods for bound constraints)
- penalty and augmented-Lagrangian methods
- barrier-function and interior-point methods
- (Sequential Quadratic Programming methods)

NOTATION

Use the following from now on:

$$
\begin{aligned}
a_i(x) &:= \nabla_x c_i(x) && \textbf{gradient} \text{ of } i\text{th constraint} \\
A(x) &:= [\nabla_x c^T(x)]^T \equiv \begin{pmatrix} a_1^T(x) \\ \cdots \\ a_m^T(x) \end{pmatrix} && \textbf{Jacobian matrix} \text{ of } c \\
H_i(x) &:= \nabla^2_{xx} c_i(x) && \textbf{Hessian} \text{ of } i\text{th constraint} \\
\ell(x,y) &:= f(x) - \langle y, c(x) \rangle && \textbf{Lagrangian} \text{ function, where} \\
& && y \text{ are } \textbf{Lagrange multipliers} \\
H(x,y) &:= \nabla^2_{xx} \ell(x,y) && \textbf{Hessian} \text{ of the Lagrangian} \\
&\equiv H(x) - \sum_{i=1}^{m} y_i H_i(x)
\end{aligned}
$$

EQUALITY CONSTRAINED MINIMIZATION

First-order necessary optimality:

Theorem 3.1. Suppose that $f,\ c \in C^1$, and that x_* is a local minimizer of $f(x)$ subject to $c(x) = 0$. Then, so long as a first-order constraint qualification holds, there exist a vector of Lagrange multipliers y_* such that

$$\begin{aligned} c(x_*) &= 0 \ \ (\textbf{primal feasibility}) \text{ and} \\ g(x_*) - A^T(x_*)y_* &= 0 \ \ (\textbf{dual feasibility}). \end{aligned}$$

EQUALITY CONSTRAINED MINIMIZATION (cont.)

Second-order necessary optimality:

Theorem 3.2. Suppose that $f,\ c \in C^2$, and that x_* is a local minimizer of $f(x)$ subject to $c(x) = 0$. Then, provided that first- and second-order constraint qualifications hold, there exist a vector of Lagrange multipliers y_* such that

$$\langle s, H(x_*, y_*)s \rangle \geqslant 0 \quad \text{for all} \quad s \in \mathcal{N}$$

where

$$\mathcal{N} = \{s \in \mathbb{R}^n \mid A(x_*)s = 0\}.$$

INEQUALITY CONSTRAINED MINIMIZATION

First-order necessary optimality:

Theorem 3.3. Suppose that $f,\ c \in C^1$, and that x_* is a local minimizer of $f(x)$ subject to $c(x) \geqslant 0$. Then, provided that a first-order constraint qualification holds, there exist a vector of Lagrange multipliers y_* such that

$$
\begin{aligned}
c(x_*) &\geqslant 0 \quad (\textbf{primal feasibility}), \\
g(x_*) - A^T(x_*)y_* &= 0 \\
\text{and} \ \ y_* &\geqslant 0 \quad (\textbf{dual feasibility}) \text{ and} \\
c_i(x_*)[y_*]_i &= 0 \quad (\textbf{complementary slackness}).
\end{aligned}
$$

Often known as the **Karush-Kuhn-Tucker (KKT)** conditions

- second-order conditions are more complicated!

SIMPLE-BOUND MINIMIZATION

First-order necessary optimality:

Theorem 3.4. Suppose that $f \in C^1$, and that x_* is a local minimizer of $f(x)$ subject to $x^{\mathrm{L}} \leqslant x \leqslant x^{\mathrm{U}}$. Then

$$x^{\mathrm{L}} \leqslant x_* \leqslant x^{\mathrm{U}} \quad \text{and} \quad P[x_* - \alpha g(x_*)] = x_*,$$

for all $\alpha \geqslant 0$, where the **projection** of x into the feasible region is

$$P_i[x] = \mathrm{mid}(x_i^{\mathrm{L}}, x_i, x_i^{\mathrm{U}}) = \begin{cases} x_i^{\mathrm{L}} & \text{if } x_i < x_i^{\mathrm{L}} \\ x_i^{\mathrm{U}} & \text{if } x_i > x_i^{\mathrm{U}} \\ x_i & \text{if } x_i^{\mathrm{L}} \leqslant x_i \leqslant x_i^{\mathrm{U}} \end{cases}$$

True more generally: if $\mathcal{F}$ is a closed, non-empty convex set, x_* is a local minimizer of $f(x) : x \in \mathcal{F}$, then $P_{\mathcal{F}}[x_* - \alpha g(x_*)] = x_*$ and $x_* \in \mathcal{F}$, where $P_{\mathcal{F}}(x) = \underset{y \in \mathcal{F}}{\arg\min} \, ||x - y||$ is the projection of x into $\mathcal{F}$

GRADIENT-PROJECTION METHODS

$$\underset{x\in\mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; x \in \text{(closed, convex)}\; \mathcal{F},$$

Generalise steepest-descent to cope with convex constraints, starting from $x_0 \in \mathcal{F}$

Linesearch variant:

$$d_k = P_{\mathcal{F}}[x_k - g(x_k)] - x_k$$

+ Armjio linesearch for $f(x_k + \alpha d_k)$ for $\alpha \in (0, 1]$

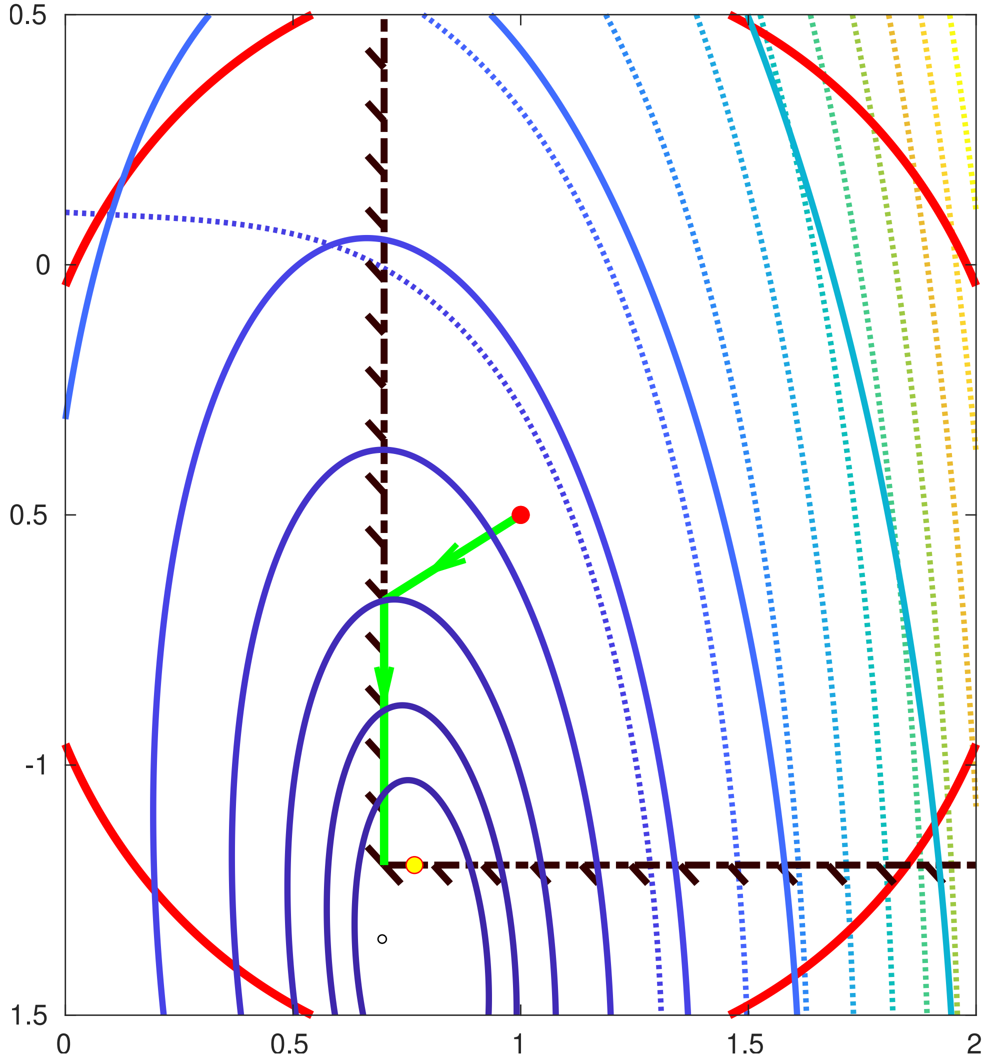
Trust-region variant: for model $m_k(s)$

$$s_k^{\rm c} = s_k(\alpha_k), \quad \text{where } \textbf{arc} \;\; s_k(\alpha) = P_{\mathcal{F}}[x_k - \alpha g(x_k)] - x_k$$

and

$$\alpha_k = \underset{\alpha > 0}{\arg\min}\, m_k(s_k(\alpha)) \;\; \text{subject to} \;\; \|s_k(\alpha)\| \leqslant \Delta_k$$

BOUND-CONSTRAINED TRUST-REGION EXAMPLE



Arc $s_k(\alpha)$ (green) from $(1, -0.5)$ with radius $\Delta = 1.1$ and $x \geqslant (0.7, -1.2)$

Part 3a: Penalty and augmented Lagrangian methods for equality constrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

Course on continuous optimization, STFC-RAL, February 2021

CONSTRAINTS AND MERIT FUNCTIONS

Two conflicting goals:

- minimize the objective function $f(x)$
- satisfy the constraints

Overcome this by minimizing a composite **merit function** $\Phi(x, p)$ for which

- p are parameters
- (some) minimizers of $\Phi(x, p)$ wrt x approach those of $f(x)$ subject to the constraints as p approaches some set $\mathcal{P}$
- only uses **unconstrained** minimization methods

AN EXAMPLE FOR EQUALITY CONSTRAINTS

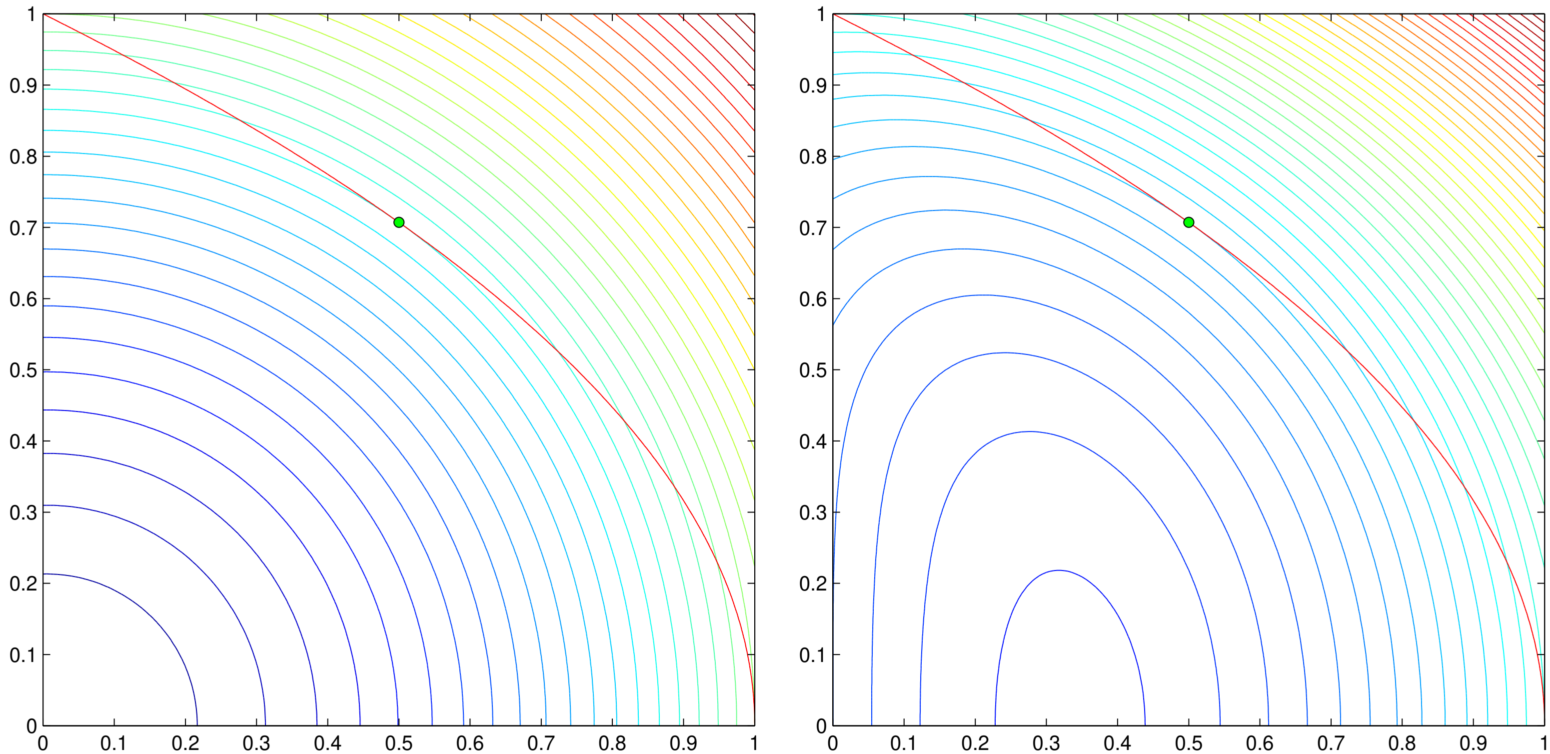
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

Merit function **(quadratic penalty function)**:

$$\Phi(x, \mu) = f(x) + \frac{1}{2\mu} \|c(x)\|_2^2$$

- required solution as μ approaches $\{0\}$ from above
- may have other useless stationary points

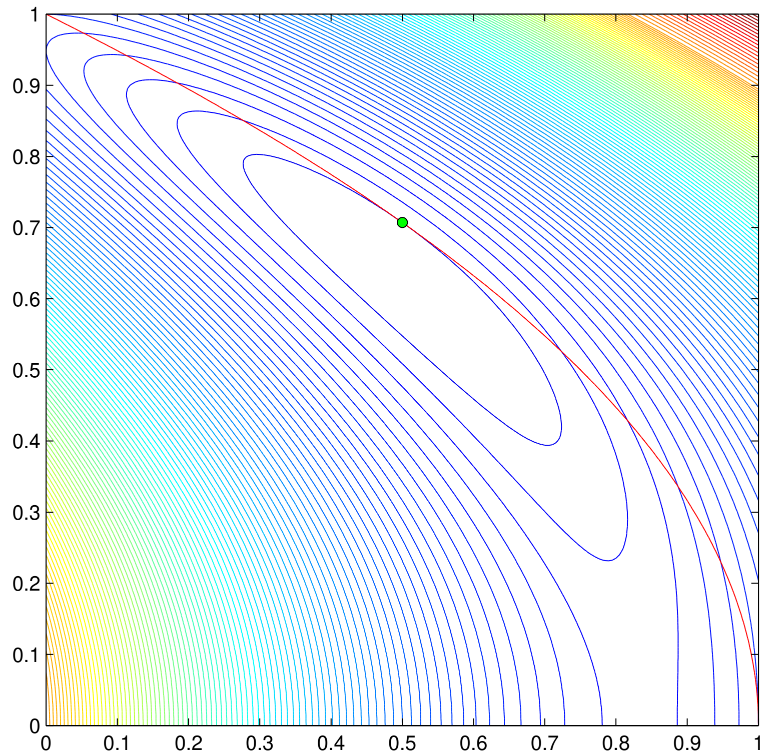
CONTOURS OF THE PENALTY FUNCTION



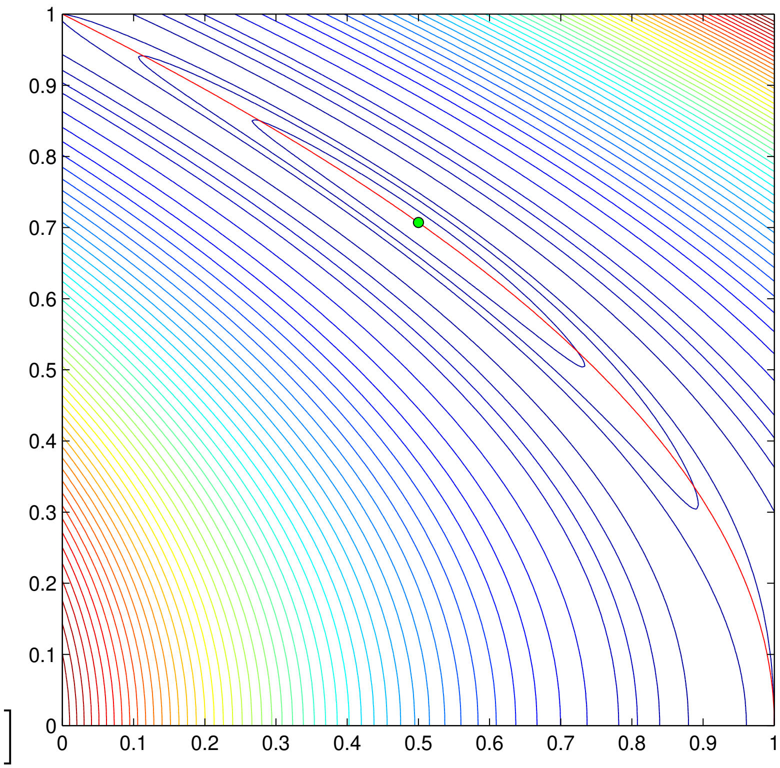
$\mu = 100$ $\mu = 1$

Quadratic penalty function for min $x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$

CONTOURS OF THE PENALTY FUNCTION (cont.)



$\mu = 0.1$



$\mu = 0.01$

Quadratic penalty function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$

BASIC QUADRATIC PENALTY FUNCTION ALGORITHM

Given $\mu_0 > 0$, set $k = 0$

Until "convergence" iterate:

Starting from x_k^{s}, use an unconstrained minimization algorithm to find an "approximate" minimizer x_k of $\Phi(x, \mu_k)$

Compute $\mu_{k+1} > 0$ smaller than μ_k such that $\lim_{k \to \infty} \mu_{k+1} = 0$ and increase k by 1

- often choose $\mu_{k+1} = 0.1\mu_k$ or even $\mu_{k+1} = \mu_k^2$
- might choose $x_{k+1}^{\mathrm{s}} = x_k$

MAIN CONVERGENCE RESULT

Theorem 3.5. Suppose that f, $c \in \mathcal{C}^2$, that

$$\|\nabla_x \Phi(x_k, \mu_k)\|_2 \leqslant \epsilon_k,$$

where ϵ_k and μ_k converge to zero as $k \longrightarrow \infty$, that

$$y_k^{\mathrm{Q}} := -\frac{c(x_k)}{\mu_k}$$

and that x_k converges to x_* for which $A(x_*)$ is full rank. Then x_* satisfies the first-order necessary optimality conditions for the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

and $\{y_k^{\mathrm{Q}}\}$ converge to the associated Lagrange multipliers y_*.

ALGORITHMS TO MINIMIZE $\Phi(x, \mu)$

Can use

- linesearch methods
  - might use specialized linesearch to cope with large quadratic term $\|c(x)\|_2^2/2\mu$
- trust-region methods
  - (ideally) need to "shape" trust region to cope with contours of the $\|c(x)\|_2^2/2\mu$ term

DERIVATIVES OF THE QUADRATIC PENALTY FUNCTION

- $\Phi(x, \mu) = f(x) + \dfrac{1}{2\mu}\|c(x)\|_2^2$
- $\nabla_x \Phi(x, \mu) = g(x) + \dfrac{1}{\mu} A^T(x) c(x) = g(x, y^{\mathrm{Q}}(x))$
- $\nabla_{xx}^2 \Phi(x, \mu) = H(x, y^{\mathrm{Q}}(x)) + \dfrac{1}{\mu} A^T(x) A(x)$

where

- $g(x, y) = g(x) - A^T(x) y$: **gradient of the Lagrangian**
- **Lagrange multiplier estimates**:

$$y^{\mathrm{Q}}(x) = -\frac{c(x)}{\mu}$$

- $H(x, y) = H(x) - \sum_{i=1}^{m} y_i H_i(x)$: **Lagrangian Hessian**

GENERIC QUADRATIC PENALTY NEWTON SYSTEM

Newton correction s from x for quadratic penalty function is

$$\left(H(x, y^{\mathrm{Q}}(x)) + \frac{1}{\mu} A^T(x) A(x) \right) s = -g(x, y^{\mathrm{Q}}(x))$$

LIMITING DERIVATIVES OF Φ

For small μ: roughly

$$\nabla_x \Phi(x, \mu) = \underbrace{g(x) - A^T(x) y^{\mathrm{Q}}(x)}_{\text{moderate}}$$

$$\nabla_{xx}^2 \Phi(x, \mu) = \underbrace{H(x, y^{\mathrm{Q}}(x))}_{\text{moderate}} + \underbrace{\frac{1}{\mu} A^T(x) A(x)}_{\text{large}} \approx \underbrace{\frac{1}{\mu} A^T(x) A(x)}_{\text{rank defficient}}$$

POTENTIAL DIFFICULTY

Ill-conditioning of the Hessian of the penalty function:

roughly speaking (non-degenerate case)

- m eigenvalues $\approx \lambda_i \left[A^T(x) A(x) \right] / \mu_k$
- $n - m$ eigenvalues $\approx \lambda_i \left[S^T(x) H(x_*, y_*) S(x) \right]$

where $S(x)$ orthogonal basis for null-space of $A(x)$

$\Longrightarrow$ condition number of $\nabla^2_{xx}\Phi(x_k, \mu_k) = O(1/\mu_k)$

$\Longrightarrow$ may not be able to find minimizer easily

THE ILL-CONDITIONING IS BENIGN

Newton system:

$$\left(H(x, y^{\mathrm{Q}}(x)) + \frac{1}{\mu} A^T(x) A(x)\right) s = -\left(g(x) + \frac{1}{\mu} A^T(x) c(x)\right)$$

Define auxiliary variables

$$w = \frac{1}{\mu}\left(A(x)s + c(x)\right)$$

$\Longrightarrow$

$$\begin{pmatrix} H(x, y^{\mathrm{Q}}(x)) & A^T(x) \\ A(x) & -\mu I \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = -\begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

- essentially independent of μ for small $\mu \Longrightarrow$ **no** inherent ill-conditioning
- thus can solve Newton equations accurately
- more sophisticated analysis $\Longrightarrow$ original system OK

PERTURBED OPTIMALITY CONDITIONS

First order optimality conditions for

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

are:

$$\begin{aligned} g(x) - A^T(x)y &= 0 \qquad \text{dual feasibility} \\ c(x) &= 0 \qquad \text{primal feasibility} \end{aligned}$$

Consider the "perturbed" problem

$$\begin{aligned} g(x) - A^T(x)y &= 0 \qquad \text{dual feasibility} \\ c(x) + \mu y &= 0 \qquad \textbf{perturbed} \text{ primal feasibility} \end{aligned}$$

where $\mu > 0$

PRIMAL-DUAL PATH-FOLLOWING METHODS

Track roots of

$$g(x) - A^T(x)y = 0 \;\text{ and }\; c(x) + \mu y = 0$$

as $0 < \mu \rightarrow 0$

- nonlinear system $\Longrightarrow$ use Newton's method

Newton correction (s, v) to (x, y) satisfies

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ A(x) & \mu I \end{pmatrix} \begin{pmatrix} s \\ v \end{pmatrix} = - \begin{pmatrix} g(x) - A^T(x)y \\ c(x) + \mu y \end{pmatrix}$$

Eliminate $v \Longrightarrow$

$$\left(H(x,y) + \frac{1}{\mu} A^T(x) A(x) \right) s = - \left(g(x) + \frac{1}{\mu} A^T(x) c(x) \right)$$

c.f. Newton method for quadratic penalty function minimization!

PRIMAL VS. PRIMAL-DUAL

Primal:

$$\left(H(x, y^{\mathrm{Q}}(x)) + \frac{1}{\mu} A^T(x) A(x) \right) s^{\mathrm{P}} = -g(x, y^{\mathrm{Q}}(x))$$

Primal-dual:

$$\left(H(x, y) + \frac{1}{\mu} A^T(x) A(x) \right) s^{\mathrm{PD}} = -g(x, y^{\mathrm{Q}}(x))$$

where

$$y^{\mathrm{Q}}(x) = -\frac{c(x)}{\mu}$$

What is the difference?

- freedom to choose y in $H(x, y)$ for primal-dual . . . vital

ANOTHER EXAMPLE FOR EQUALITY CONSTRAINTS

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \text{ subject to } \; c(x) = 0$$

Merit function **(augmented Lagrangian function)**:

$$\Phi(x, u, \mu) = f(x) - \langle y, c(x) \rangle + \frac{1}{2\mu} \|c(x)\|_2^2$$

where y and μ are auxiliary **parameters**

Two interpretations —

- shifted quadratic penalty function
- convexification of the Lagrangian function

Aim: adjust μ **and** y to encourage convergence

DERIVATIVES OF THE AUGMENTED LAGRANGIAN FUNCTION

- $\Phi(x, y, \mu) = f(x) - \langle y, c(x) \rangle + \dfrac{1}{2\mu}\|c(x)\|_2^2$
- $\nabla_x \Phi(x, y, \mu) = g(x) - A^T(x)y + \dfrac{1}{\mu}A^T(x)c(x) = g(x, y^{\mathrm{A}}(x))$
- $\nabla_{xx}^2 \Phi(x, y, \mu) = H(x, y^{\mathrm{A}}(x)) + \dfrac{1}{\mu}A^T(x)A(x)$

where

- $g(x, y) = g(x) - A^T(x)y$: gradient of the Lagrangian
- **First-order** Lagrange multiplier estimates:

$$y^{\mathrm{A}}(x) = y - \frac{c(x)}{\mu}$$

- $H(x, y) = H(x) - \sum_{i=1}^{m} y_i(x)H_i(x)$: Lagrangian Hessian

Crucially

$$c(x) = \mu[y^{\mathrm{A}}(x) - y]$$

AUGMENTED LAGRANGIAN CONVERGENCE

Theorem 3.6. Suppose that $f,\ c \in \mathcal{C}^2$, that

$$\|\nabla_x \Phi(x_k, y_k, \mu_k)\|_2 \leqslant \epsilon_k,$$

for given $\{y_k\}$, where ϵ_k converges to zero as $k \rightarrow \infty$, that

$$y_k^{\mathrm{A}} := y_k - c(x_k)/\mu_k,$$

and that x_k converges to x_* for which $A(x_*)$ is full rank. Then $\{y_k^{\mathrm{A}}\}$ converge to some y_* for which $g(x_*) = A^T(x_*)y_*$.

If additionally either

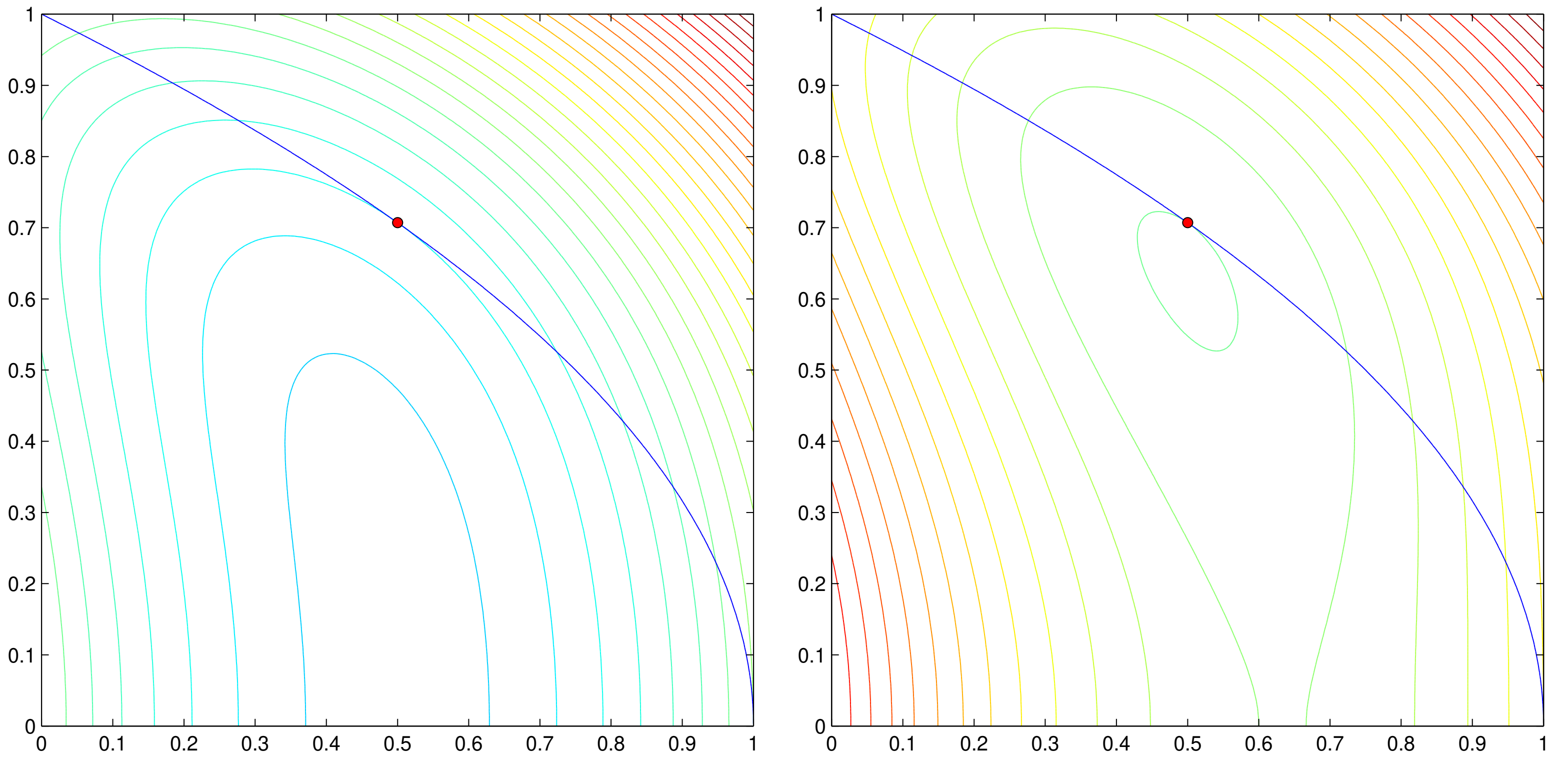
(i) μ_k converges to zero for bounded y_k or

(ii) y_k converges to y_* for bounded μ_k,

then x_* and y_* satisfy the first-order necessary optimality conditions for the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ \ f(x) \text{ subject to } \ c(x) = 0$$

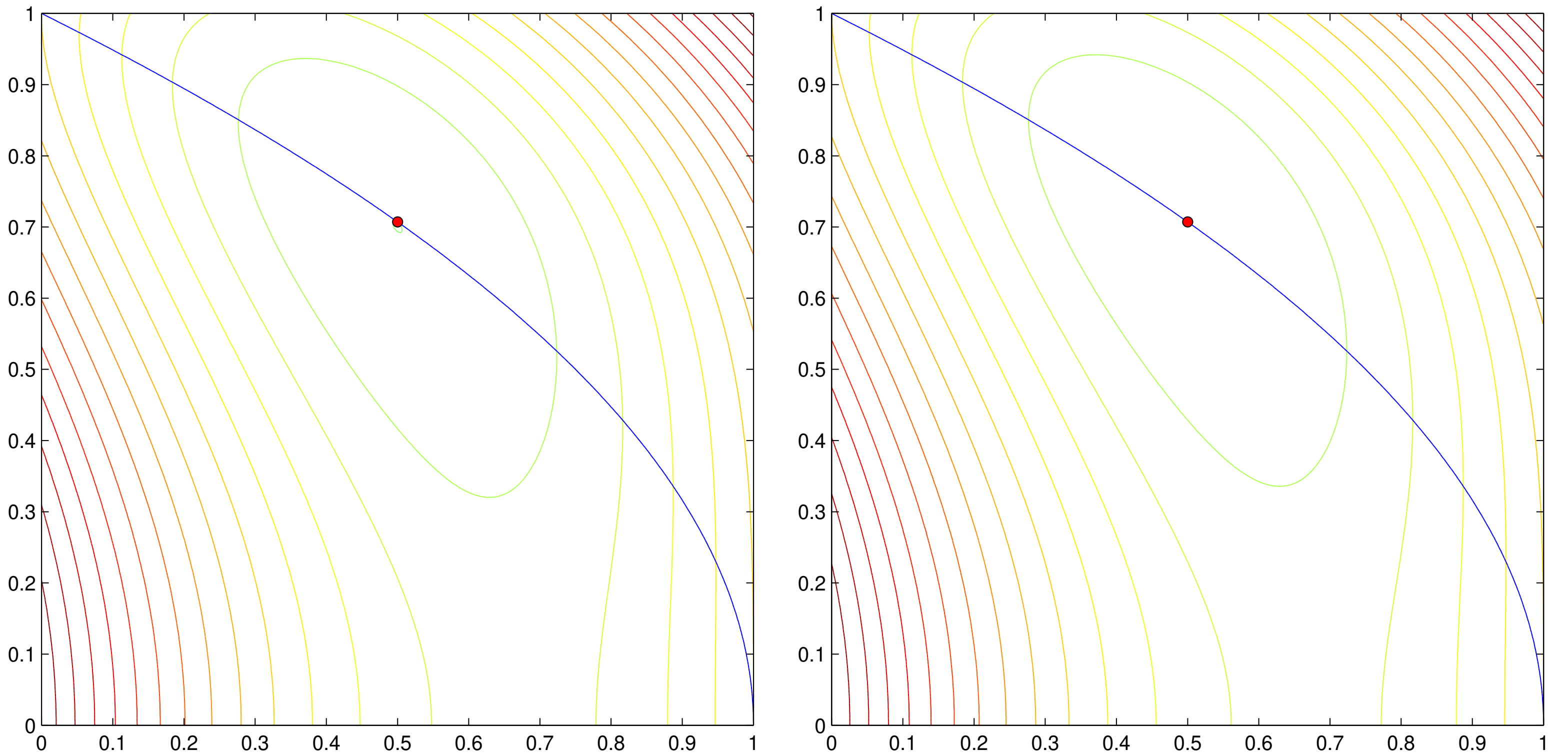
CONTOURS OF THE AUGMENTED LAGRANGIAN FUNCTION



$u = 0.5$ $\qquad$ $u = 0.9$

Augmented Lagrangian function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$ with fixed $\mu = 1$

CONTOURS OF THE AUGMENTED LAGRANGIAN FUNCTION (cont.)



$u = 0.99$ $\qquad$ $u = y_* = 1$

Augmented Lagrangian function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$ with fixed $\mu = 1$

CONVERGENCE OF AUGMENTED LAGRANGIAN METHODS

- convergence guaranteed if y_k fixed and $\mu \longrightarrow 0$
  $\Longrightarrow y_k \longrightarrow y_*$ and $c(x_k) \longrightarrow 0$
- check if $\|c(x_k)\| \leqslant \eta_k$ where $\{\eta_k\} \longrightarrow 0$
  - if so, set $y_{k+1} = y_k - c(x_k)/\mu_k$ and $\mu_{k+1} = \mu_k$
  - if not, set $y_{k+1} = y_k$ and $\mu_{k+1} \leqslant \tau\mu_k$ for some $\tau \in (0,1)$
- reasonable: $\eta_k = \mu_k^{0.1+0.9j}$ where j iterations since μ_k last changed
- under such rules, can ensure μ_k eventually unchanged under modest assumptions and (fast) linear convergence
- need also to ensure μ_k is sufficiently large that $\nabla^2_{xx}\Phi(x_k, y_k, \mu_k)$ is positive (semi-)definite

BASIC AUGMENTED LAGRANGIAN ALGORITHM

Given $\mu_0 > 0$ and y_0, set $k = 0$

Until "convergence" iterate:

Starting from x_k^{s}, use an unconstrained minimization algorithm to find an "approximate" minimizer x_k of $\Phi(x, y_k, \mu_k)$ for which $\|\nabla_x \Phi(x_k, y_k, \mu_k)\| \leqslant \epsilon_k$

If $\|c(x_k)\| \leqslant \eta_k$, set $y_{k+1} = y_k - c(x_k)/\mu_k$ and $\mu_{k+1} = \mu_k$

Otherwise set $y_{k+1} = u_k$ and $\mu_{k+1} \leqslant \tau \mu_k$

Set suitable ϵ_{k+1} and η_{k+1} and increase k by 1

- often choose $\tau = \min(0.1, \sqrt{\mu_k})$
- might choose $x_{k+1}^{\mathrm{s}} = x_k$
- reasonable: $\epsilon_k = \mu_k^{j+1}$ where j iterations since μ_k last changed

Part 3b: Interior-point methods for inequality constrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x) \ \text{ subject to }\ c(x) \geqslant 0$$

Course on continuous optimization, STFC-RAL, February 2021

CONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) \geqslant 0$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
and the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$

- assume that $f, \; c \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary

CONSTRAINTS AND MERIT FUNCTIONS

Two conflicting goals:

- minimize the objective function $f(x)$
- satisfy the constraints

Recall — overcome this by minimizing a composite **merit function** $\Phi(x, p)$ for which

- p are parameters
- (some) minimizers of $\Phi(x, p)$ wrt x approach those of $f(x)$ subject to the constraints as p approaches some set $\mathcal{P}$
- only uses **unconstrained** minimization methods

A MERIT F$^{\underline{\text{n}}}$ FOR INEQUALITY CONSTRAINTS

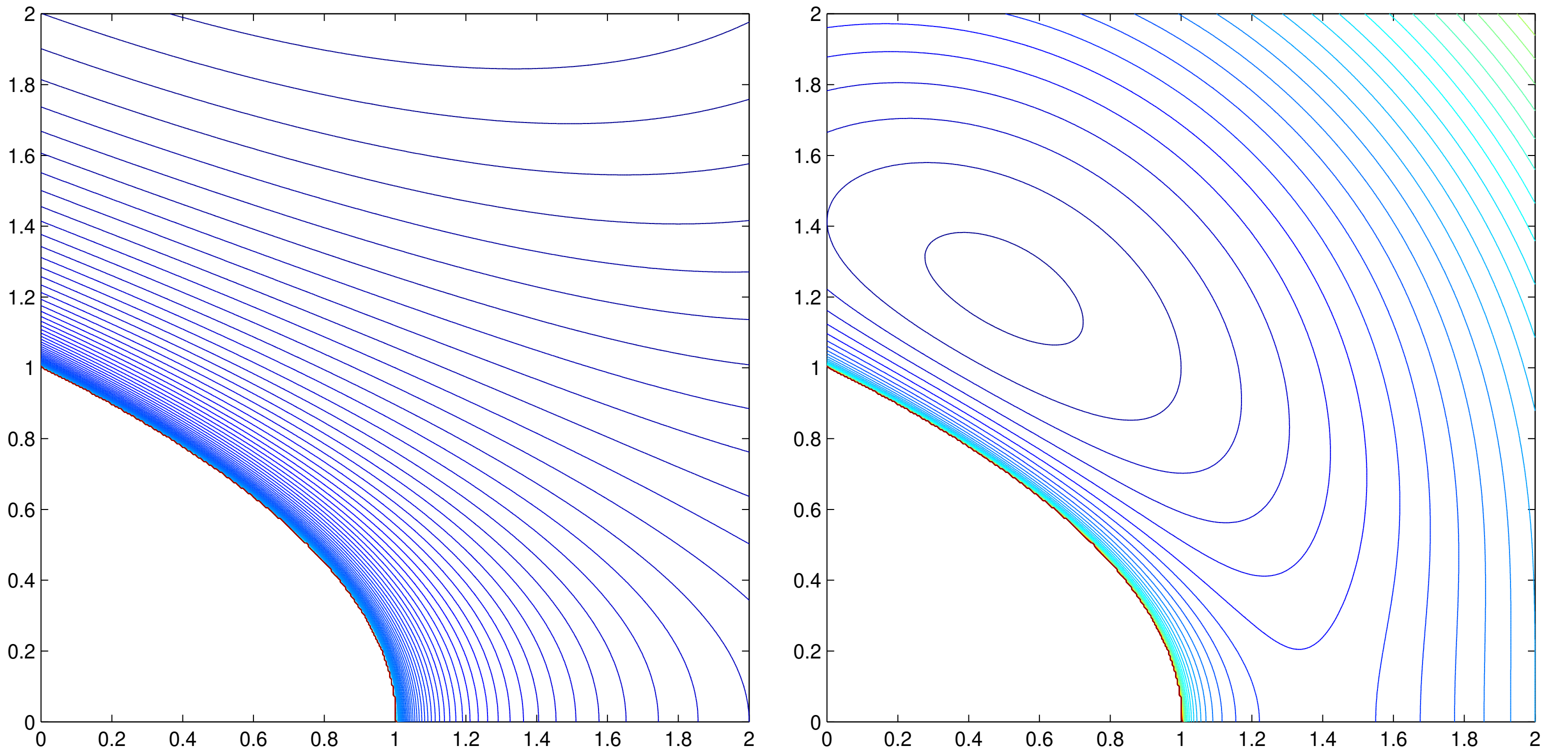
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) \geqslant 0$$

Merit function **(logarithmic barrier function)**:

$$\Phi(x, \mu) = f(x) - \mu \sum_{i=1}^{m} \log c_i(x)$$

- required solution as μ approaches $\{0\}$ from above
- may have other useless stationary points
- requires a strictly interior point to start
- consequent points are interior

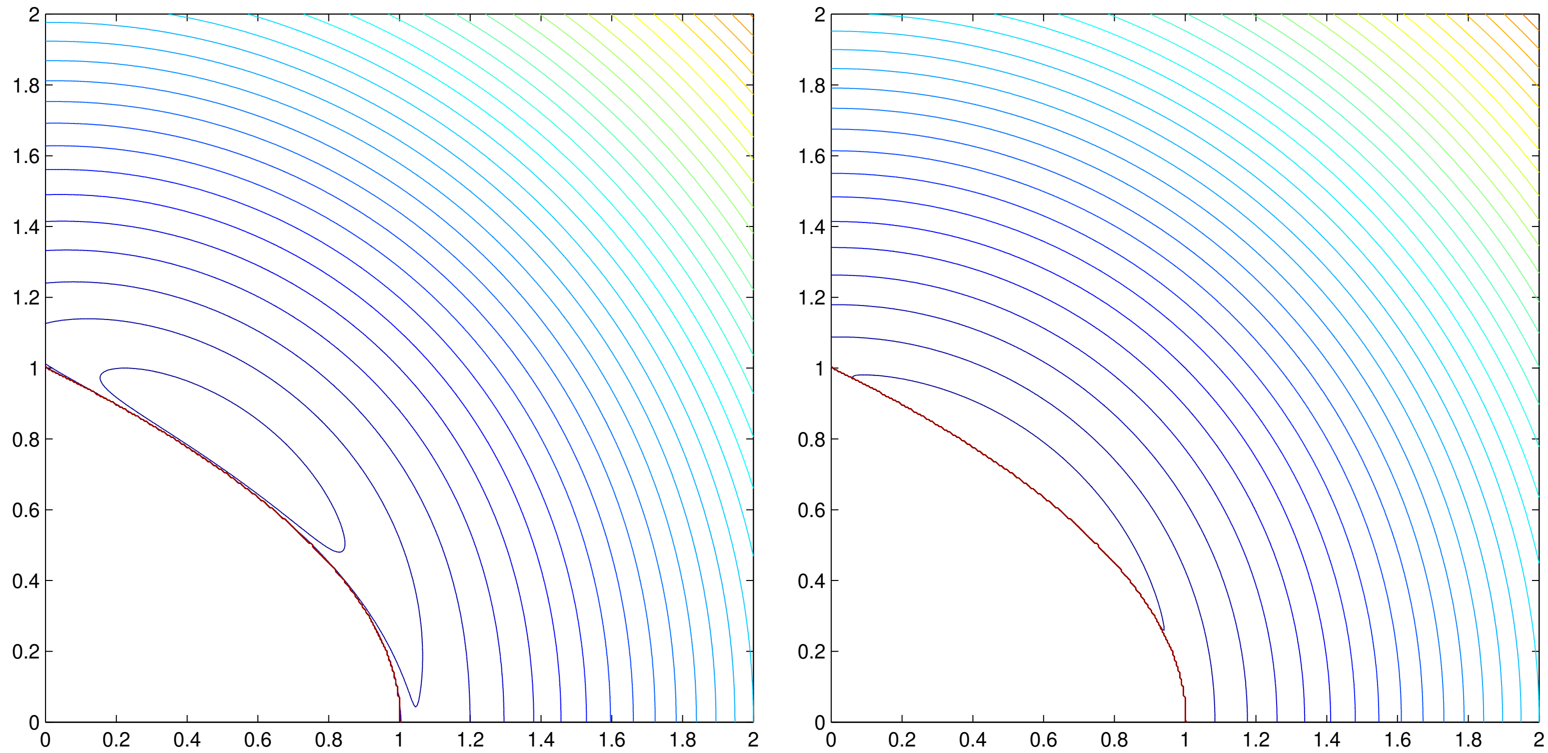
CONTOURS OF THE BARRIER FUNCTION



$\mu = 10$ $\qquad\qquad$ $\mu = 1$

Barrier function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 \geqslant 1$

CONTOURS OF THE BARRIER FUNCTION (cont.)



$\mu = 0.1$ $\mu = 0.01$

Barrier function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 \geqslant 1$

BASIC BARRIER FUNCTION ALGORITHM

Given $\mu_0 > 0$, set $k = 0$

Until "convergence" iterate:

  Find x_k^{s} for which $c(x_k^{\mathrm{s}}) > 0$

  Starting from x_k^{s}, use an unconstrained minimization algorithm to find an "approximate" minimizer x_k of $\Phi(x, \mu_k)$

  Compute $\mu_{k+1} > 0$ smaller than μ_k such that $\lim_{k\to\infty} \mu_{k+1} = 0$ and increase k by 1

- often choose $\mu_{k+1} = 0.1\mu_k$ or even $\mu_{k+1} = \mu_k^2$
- might choose $x_{k+1}^{\mathrm{s}} = x_k$

MAIN CONVERGENCE RESULT

The **active set** $\mathcal{A}(x) = \{i : c_i(x) = 0\}$

Theorem 3.7. Suppose that f, $c \in \mathcal{C}^2$, that

$$\|\nabla_x \Phi(x_k, \mu_k)\|_2 \leqslant \epsilon_k$$

where ϵ_k converges to zero as $k \to \infty$, that

$$(y_k)_i := \mu_k / c_i(x_k) \;\; \text{for} \;\; i = 1, \ldots, m,$$

and that x_k converges to x_* for which $\{a_i(x_*)\}_{i \in \mathcal{A}(x_*)}$ are linearly independent. Then x_* satisfies the first-order necessary optimality conditions for the problem

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) \geqslant 0$$

and $\{y_k\}$ converge to the associated Lagrange multipliers y_*.

ACTIVE AND INACTIVE CONSTRAINTS

Since (complementary slackness)

$$c_i(x_*)(y_*)_i = 0 \;\text{ for all }\; i = 1, \ldots m$$

Often have $\{x_k\} \to x_*$ and $\{y_k\} \to y_*$ with

- $c_i(x_k) \to 0$ and $(y_k)_i \to (y_*)_i > 0$ for $i \in \mathcal{A}(x_*)$
  active constraints
- $c_i(x_k) \to c_i(x_*) > 0$ and $(y_k)_i \to 0$ for $i \in \mathcal{I}(x_*) = \{1, \ldots, m\} \backslash \mathcal{A}(x_*)$
  inactive constraints
- sometimes **degeneracy**: $c_i(x_*) = 0$ and $(y_*)_i = 0$

ALGORITHMS TO MINIMIZE $\boldsymbol{\Phi(x, \mu)}$

Can use

- linesearch methods
  - should use specialized linesearch to cope with singularity of log
- trust-region methods
  - need to reject points for which $c(x_k + s_k) \not> 0$
  - (ideally) need to "shape" trust region to cope with contours of the singularity

DERIVATIVES OF THE BARRIER FUNCTION

- $\nabla_x \Phi(x, \mu) = g(x, y(x))$
- $$\begin{aligned} \nabla^2_{xx} \Phi(x, \mu) &= H(x, y(x)) + \mu A^T(x) C^{-2}(x) A(x) \\ &= H(x, y(x)) + A^T(x) C^{-1}(x) Y(x) A(x) \\ &= H(x, y(x)) + \tfrac{1}{\mu} A^T(x) Y^2(x) A(x) \end{aligned}$$

where

- **Lagrange multiplier estimates**: $y(x) = \mu C^{-1}(x) e$

  where e is the vector of ones
- $C(x) = \mathrm{diag}(c_1(x), \ldots, c_m(x))$
- $Y(x) = \mathrm{diag}(y_1(x), \ldots, y_m(x)) = \mu C^{-1}(x)$
- $g(x, y(x)) = g(x) - A^T(x) y(x)$: **gradient of the Lagrangian**
- $H(x, y(x)) = H(x) - \sum_{i=1}^{m} y_i(x) H_i(x)$: **Lagrangian Hessian**

LIMITING DERIVATIVES OF Φ

Let $\mathcal{I}$ = inactive set at $x_* = \{1, \ldots, m\} \backslash \mathcal{A}$

For small μ: roughly

$$\begin{aligned}
\nabla_x \Phi(x, \mu) &= g(x) - \mu A^T(x) C^{-1}(x) e \\
&= \underbrace{g(x) - A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}(x) e}_{\text{moderate}} - \underbrace{\mu A_{\mathcal{I}}^T(x) C_{\mathcal{I}}^{-1}(x) e}_{\text{small}} \\
&\approx g(x) - A_{\mathcal{A}}^T(x) y_{\mathcal{A}}(x)
\end{aligned}$$

$$\begin{aligned}
\nabla_{xx}^2 \Phi(x, \mu) &= \underbrace{H(x, y(x))}_{\text{moderate}} + \underbrace{\mu A_{\mathcal{I}}^T(x) C_{\mathcal{I}}^{-2}(x) A_{\mathcal{I}}(x)}_{\text{small}} + \underbrace{\frac{1}{\mu} A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^2(x) A_{\mathcal{A}}(x)}_{\text{large}} \\
&\approx \frac{1}{\mu} A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^2(x) A_{\mathcal{A}}(x) \\
&= A_{\mathcal{A}}^T(x) C_{\mathcal{A}}^{-1}(x) Y_{\mathcal{A}}(x) A_{\mathcal{A}}(x) \\
&= \mu A_{\mathcal{A}}^T(x) C_{\mathcal{A}}^{-2}(x) A_{\mathcal{A}}(x)
\end{aligned}$$

GENERIC BARRIER NEWTON SYSTEM

Newton correction s from x for barrier function is

$$\Big(H(x, y(x)) + A^T(x) C^{-1}(x) Y(x) A(x) \Big) \, s = -g(x, y(x))$$

LIMITING NEWTON METHOD

For small μ: roughly

$$\frac{1}{\mu} A_{\mathcal{A}}^T(x) Y_{\mathcal{A}}^2(x) A_{\mathcal{A}}(x) s \approx - \Big(g(x) - A_{\mathcal{A}}^T(x) y_{\mathcal{A}}(x) \Big)$$

POTENTIAL DIFFICULTIES I

Ill-conditioning of the Hessian of the barrier function:

roughly speaking (non-degenerate case)

- m_a eigenvalues $\approx \lambda_i \left[A_{\mathcal{A}}^T Y_{\mathcal{A}}^2 A_{\mathcal{A}} \right] / \mu_k$
- $n - m_a$ eigenvalues $\approx \lambda_i \left[N_{\mathcal{A}}^T H(x_*, y_*) N_{\mathcal{A}} \right]$

where

m_a = number of active constraints

$\mathcal{A}$ = active set at x_*

Y = diagonal matrix of Lagrange multipliers

$N_{\mathcal{A}}$ = orthogonal basis for null-space of $A_{\mathcal{A}}$

$\Longrightarrow$ condition number of $\nabla_{xx}^2 \Phi(x_k, \mu_k) = O(1/\mu_k)$

$\Longrightarrow$ may not be able to find minimizer easily

POTENTIAL DIFFICULTIES II

Value $\boldsymbol{x^{\mathrm{s}}_{k+1} = x_k}$ is a poor starting point: Suppose

$$\begin{aligned} 0 &\approx \nabla_x \Phi(x_k, \mu_k) = g(x_k) - \mu_k A^T(x_k) C^{-1}(x_k) e \\ &\approx g(x_k) - \mu_k A^T_{\mathcal{A}}(x_k) C^{-1}_{\mathcal{A}}(x_k) e \end{aligned}$$

Roughly speaking (non-degenerate case) Newton correction satisfies

$$\mu_{k+1} A^T_{\mathcal{A}}(x_k) C^{-2}_{\mathcal{A}}(x_k) A_{\mathcal{A}}(x_k) s \approx (\mu_{k+1} - \mu_k) A^T_{\mathcal{A}}(x_k) C^{-1}_{\mathcal{A}}(x_k) e$$

$\Longrightarrow$ (full rank)

$$A_{\mathcal{A}}(x_k) s \approx \left(1 - \frac{\mu_k}{\mu_{k+1}}\right) c_{\mathcal{A}}(x_k)$$

$\Longrightarrow$ (Taylor expansion)

$$c_{\mathcal{A}}(x_k + s) \approx c_{\mathcal{A}}(x_k) + A_{\mathcal{A}}(x_k) s \approx \left(2 - \frac{\mu_k}{\mu_{k+1}}\right) c_{\mathcal{A}}(x_k) < 0$$

if $\mu_{k+1} < \frac{1}{2}\mu_k \Longrightarrow$ Newton step infeasible $\Longrightarrow$ slow convergence

PERTURBED OPTIMALITY CONDITIONS

First order optimality conditions for

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \; \text{ subject to } \; c(x) \geqslant 0$$

are:

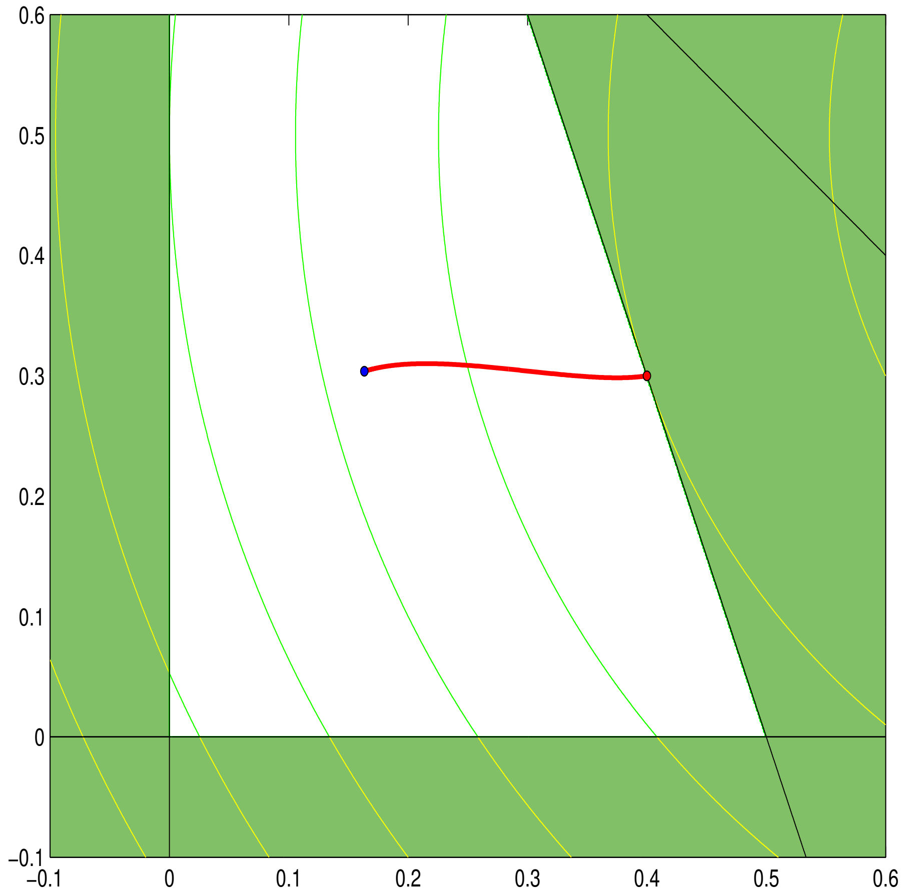
$$\begin{aligned} g(x) - A^T(x)y &= 0 && \text{dual feasibility} \\ C(x)y &= 0 && \text{complementary slackness} \\ c(x) \geqslant 0 \;\text{ and }\; y &\geqslant 0 \end{aligned}$$

Consider the "perturbed" problem

$$\begin{aligned} g(x) - A^T(x)y &= 0 && \text{dual feasibility} \\ C(x)y &= \mu e && \textbf{perturbed} \text{ comp. slkns.} \\ c(x) > 0 \;\text{ and }\; y &> 0 \end{aligned}$$

where $\mu > 0$

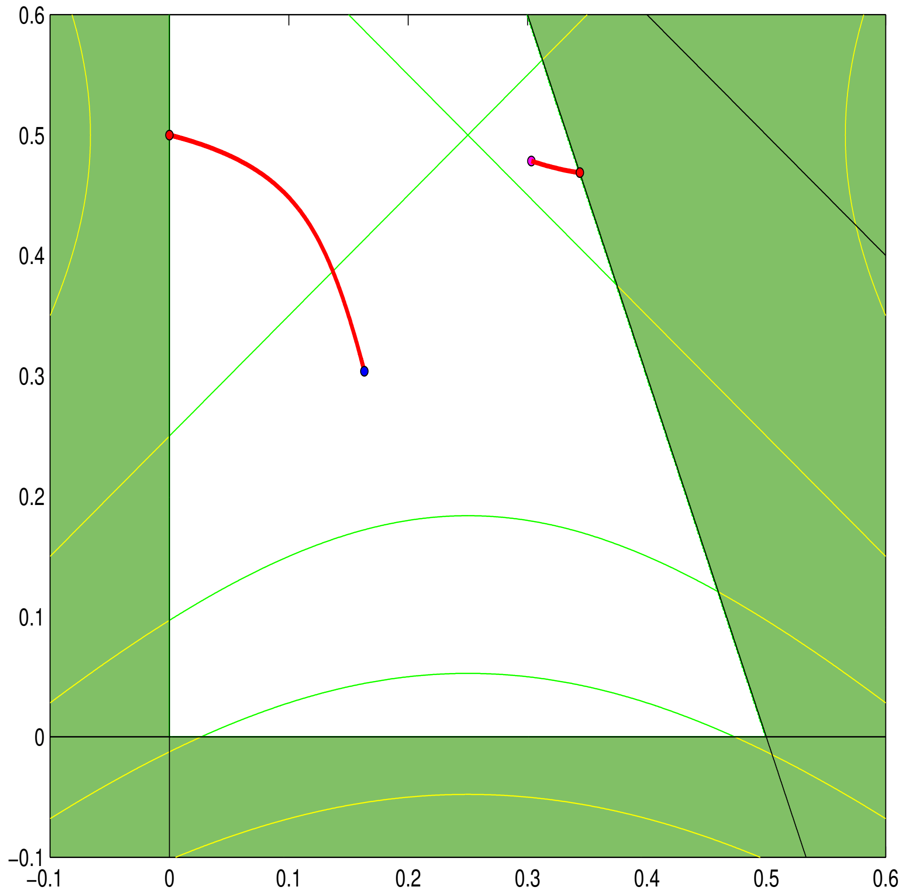
CENTRAL PATH TRAJECTORY



$$\min (x_1 - 1)^2 + (x_2 - 0.5)^2$$
$$\text{subject to } \; x_1 + x_2 \leqslant 1$$
$$3x_1 + x_2 \leqslant 1.5$$
$$(x_1, x_2) \geqslant 0$$

Trajectory $x(\mu)$ of perturbed optimality conditions as μ ranges from infinity down to zero

TRAJECTORIES FOR THE NON-CONVEX CASE



$$\min -2(x_1 - 0.25)^2 + 2(x_2 - 0.5)^2$$
$$\text{subject to } \; x_1 + x_2 \leqslant 1$$
$$3x_1 + x_2 \leqslant 1.5$$
$$(x_1, x_2) \geqslant 0$$

Trajectories $x(\mu)$ of perturbed optimality conditions as μ ranges from infinity down to zero

PRIMAL-DUAL PATH-FOLLOWING METHODS

Track roots of

$$g(x) - A^T(x)y = 0 \;\text{ and }\; C(x)y - \mu e = 0$$

as $0 < \mu \to 0$, while maintaining $c(x) > 0$ and $y > 0$

- this is a nonlinear system $\Longrightarrow$ use Newton's method

Newton correction (s, w) to (x, y) satisfies

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ YA(x) & C(x) \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x) - A^T(x)y \\ C(x)y - \mu e \end{pmatrix}$$

Eliminate $w \Longrightarrow$

$$\Big(H(x,y) + A^T(x)C^{-1}(x)YA(x)\Big)\, s = -\Big(g(x) - \mu A^T(x)C^{-1}(x)e\Big)$$

c.f. Newton method for barrier minimization!

PRIMAL VS. PRIMAL-DUAL

Primal:

$$\left(H(x,y(x)) + A^T(x)C^{-1}(x)Y(x)A(x)\right) s^{\text{P}} = -g(x,y(x))$$

Primal-dual:

$$\left(H(x,y) + A^T(x)C^{-1}(x)YA(x)\right) s^{\text{PD}} = -g(x,y(x))$$

where

$$y(x) = \mu C^{-1}(x)e$$

What is the difference?

- freedom to choose y in $H(x,y) + A^T(x)C^{-1}(x)YA(x)$ for primal-dual ... vital
- Hessian approximation for small μ

$$H(x,y) + A^T(x)C^{-1}(x)YA(x) \approx A^T_{\mathcal{A}}(x)C^{-1}_{\mathcal{A}}(x)Y_{\mathcal{A}}A_{\mathcal{A}}(x)$$

POTENTIAL DIFFICULTY II . . . REVISITED

Value $x^{\mathrm{S}}_{k+1} = x_k$ can be a good starting point:

- primal method has to choose $y = y(x^{\mathrm{S}}_k) = \mu_{k+1} C^{-1}(x_k) e$
  - factor μ_{k+1}/μ_k too small for a good Lagrange multiplier estimate
- primal-dual method can choose $y = \mu_k C^{-1}(x_k) e \rightarrow y_*$

Advantage: roughly (non-degenerate case) correction s^{PD} satisfies

$$\mu_k A^T_{\mathcal{A}}(x_k) C^{-2}_{\mathcal{A}}(x_k) A_{\mathcal{A}}(x_k) s^{\mathrm{PD}} \approx (\mu_{k+1} - \mu_k) A^T_{\mathcal{A}}(x_k) C^{-1}_{\mathcal{A}}(x_k) e$$

$\Longrightarrow$ (full rank)

$$A_{\mathcal{A}}(x_k) s^{\mathrm{PD}} \approx \left(\frac{\mu_{k+1}}{\mu_k} - 1 \right) c_{\mathcal{A}}(x_k)$$

$\Longrightarrow$ (Taylor expansion)

$$c_{\mathcal{A}}(x_k + s^{\mathrm{PD}}) \approx c_{\mathcal{A}}(x_k) + A_{\mathcal{A}}(x_k) s^{\mathrm{PD}} \approx \frac{\mu_{k+1}}{\mu_k} c_{\mathcal{A}}(x_k) > 0$$

$\Longrightarrow$ Newton step allowed $\Longrightarrow$ fast convergence

PRIMAL-DUAL BARRIER METHODS

Choose a search direction s for $\Phi(x, \mu_k)$ by (approximately) solving the problem

$$\underset{s \in \mathbb{R}^n}{\text{minimize}} \; g(x, y(x))^T s + \tfrac{1}{2} s^T \left(H(x, y) + A^T(x) C^{-1}(x) Y A(x) \right) s$$

possibly subject to a trust-region constraint

- $y(x) = \mu C^{-1}(x) e \Longrightarrow g(x, y(x)) = \nabla_x \Phi(x, \mu)$
- $y = \ldots$
  - $y(x) \Longrightarrow$ primal Newton method
  - occasionally $(\mu_{k-1}/\mu_k) y(x) \Longrightarrow$ good starting point
  - $y^{\text{OLD}} + w^{\text{OLD}} \Longrightarrow$ primal-dual Newton method
  - $\max(y^{\text{OLD}} + w^{\text{OLD}}, \epsilon(\mu_k) e)$ for "small" $\epsilon(\mu_k) > 0$ (e.g., $\epsilon(\mu_k) = \mu_k^{1.5}$) $\Longrightarrow$ practical primal-dual method

POTENTIAL DIFFICULTY I ... REVISITED

Ill-conditioning $\not\Longrightarrow$ we can't solve equations accurately:

roughly (non-degenerate case, $\mathcal{I}$ = inactive set at x_*)

$$\begin{pmatrix} H & -A^T \\ YA & C \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g - A^T y \\ Cy - \mu e \end{pmatrix} \quad \Longrightarrow$$

$$\begin{pmatrix} H & -A_{\mathcal{A}}^T & -A_{\mathcal{I}}^T \\ Y_{\mathcal{A}} A_{\mathcal{A}} & C_{\mathcal{A}} & 0 \\ Y_{\mathcal{I}} A_{\mathcal{I}} & 0 & C_{\mathcal{I}} \end{pmatrix} \begin{pmatrix} s \\ w_{\mathcal{A}} \\ w_{\mathcal{I}} \end{pmatrix} = - \begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} - A_{\mathcal{I}}^T y_{\mathcal{I}} \\ C_{\mathcal{A}} y_{\mathcal{A}} - \mu e \\ C_{\mathcal{I}} y_{\mathcal{I}} - \mu e \end{pmatrix} \quad \Longrightarrow$$

$$\begin{pmatrix} H + A_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} Y_{\mathcal{I}} A_{\mathcal{I}} & -A_{\mathcal{A}}^T \\ A_{\mathcal{A}} & C_{\mathcal{A}} Y_{\mathcal{A}}^{-1} \end{pmatrix} \begin{pmatrix} s \\ w_{\mathcal{A}} \end{pmatrix} = - \begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} - \mu A_{\mathcal{I}}^T C_{\mathcal{I}}^{-1} e \\ c_{\mathcal{A}} - \mu Y_{\mathcal{A}}^{-1} e \end{pmatrix}$$

- potentially bad terms $C_{\mathcal{I}}^{-1}$ and $Y_{\mathcal{A}}^{-1}$ bounded
- in the limit becomes well-behaved

$$\begin{pmatrix} H & -A_{\mathcal{A}}^T \\ A_{\mathcal{A}} & 0 \end{pmatrix} \begin{pmatrix} s \\ w_{\mathcal{A}} \end{pmatrix} = - \begin{pmatrix} g - A_{\mathcal{A}}^T y_{\mathcal{A}} \\ 0 \end{pmatrix}$$

PRACTICAL PRIMAL-DUAL METHOD

Given $\mu_0 > 0$ and feasible $(x_0^{\mathrm{s}}, y_0^{\mathrm{s}})$, set $k = 0$

Until "convergence" iterate:

Inner minimization: starting from $(x_k^{\mathrm{s}}, y_k^{\mathrm{s}})$, use an unconstrained minimization algorithm to find (x_k, y_k) for which $\|C(x_k)y_k - \mu_k e\| \leqslant \mu_k$ and $\|g(x_k) - A^T(x_k)y_k\| \leqslant \mu_k^{1.00005}$

Set $\mu_{k+1} = \min(0.1\mu_k, \mu_k^{1.9999})$

Find $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ using a primal-dual Newton step from (x_k, y_k)

If $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ is infeasible, reset $(x_{k+1}^{\mathrm{s}}, y_{k+1}^{\mathrm{s}})$ to (x_k, y_k)

Increase k by 1

FAST ASYMPTOTIC CONVERGENCE

Theorem 3.8. Suppose that f, $c \in \mathcal{C}^2$, that a subsequence $\{(x_k, y_k)\}$, $k \in \mathcal{K}$, of the practical primal-dual method converges to (x_*, y_*) satisfying second-order sufficiency conditions, that $A_{\mathcal{A}}(x_*)$ is full-rank, and that $(y_*)_{\mathcal{A}} > 0$. Then the starting point satisfies the inner-minimization termination test (i.e., $(x_k, y_k) = (x_k^{\mathrm{s}}, y_k^{\mathrm{s}})$) and the whole sequence $\{(x_k, y_k)\}$ converges to (x_*, y_*) at a superlinear rate (Q-factor 1.9998).

OTHER ISSUES

- polynomial algorithms for many convex problems
  - linear programming
  - quadratic programming
  - semi-definite programming ...
- excellent practical performance
- globally, need to keep away from constraint boundary until near convergence, otherwise very slow
- initial interior point:

$$\underset{(x,c)}{\text{minimize}} \ \langle e, c\rangle \ \text{subject to} \ c(x) + c \geqslant 0$$

Part 3c: SQP methods for equality constrained optimization

Nick Gould (nick.gould@stfc.ac.uk)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

Course on continuous optimization, STFC-RAL, February 2021

EQUALITY CONSTRAINED MINIMIZATION

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

where the **objective function** $f : \mathbb{R}^n \longrightarrow \mathbb{R}$
and the **constraints** $c : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(m \leqslant n)$

- assume that $f, \; c \in C^1$ (sometimes C^2) and Lipschitz
- often in practice this assumption violated, but not necessary
- easily generalized to inequality constraints ... but may be better to use interior-point methods for these

OPTIMALITY AND NEWTON'S METHOD

1st order optimality:

$$g(x,y) \equiv g(x) - A^T(x)y = 0 \;\text{ and }\; c(x) = 0$$

this is a nonlinear system (linear in y)

$\Longrightarrow$

use Newton's method to find a correction (s,w) to (x,y)

$\Longrightarrow$

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x,y) \\ c(x) \end{pmatrix}$$

ALTERNATIVE FORMULATIONS

unsymmetric:

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ w \end{pmatrix} = - \begin{pmatrix} g(x,y) \\ c(x) \end{pmatrix}$$

or symmetric:

$$\begin{pmatrix} H(x,y) & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -w \end{pmatrix} = - \begin{pmatrix} g(x,y) \\ c(x) \end{pmatrix}$$

or (with $y^+ = y + w$) unsymmetric:

$$\begin{pmatrix} H(x,y) & -A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

or symmetric:

$$\begin{pmatrix} H(x,y) & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

DETAILS

- Often approximate with symmetric $B \approx H(x,y) \implies$ e.g.

$$\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = - \begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

- solve system using
  - unsymmetric (LU) factorization of $\begin{pmatrix} B & -A^T(x) \\ A(x) & 0 \end{pmatrix}$
  - symmetric (indefinite) factorization of $\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix}$
  - symmetric factorizations of B and the Schur Complement $A(x)B^{-1}A^T(x)$
  - iterative method (GMRES(k), MINRES, CG within $\mathcal{N}(A)$,...)

AN ALTERNATIVE INTERPRETATION

$$\textbf{QP}: \ \underset{s \in \mathbb{R}^n}{\text{minimize}} \ \langle s, g(x) \rangle + \tfrac{1}{2}\langle s, Bs \rangle \ \text{ subject to } \ A(x)s = -c(x)$$

- QP = **quadratic program**
- first-order model of constraints $c(x+s)$
- second-order model of objective $f(x+s)$ ... but
  B includes curvature of constraints

solution to QP satisfies

$$\begin{pmatrix} B & A^T(x) \\ A(x) & 0 \end{pmatrix} \begin{pmatrix} s \\ -y^+ \end{pmatrix} = -\begin{pmatrix} g(x) \\ c(x) \end{pmatrix}$$

SEQUENTIAL QUADRATIC PROGRAMMING - SQP

or **successive** quadratic programming

or **recursive** quadratic programming (RQP)

Given (x_0, y_0), set $k = 0$

Until "convergence" iterate:

Compute a suitable symmetric B_k using (x_k, y_k)

Find

$$s_k = \underset{s \in \mathbb{R}^n}{\arg\min} \; \langle g_k, s \rangle + \tfrac{1}{2}\langle s, B_k s \rangle \;\text{ subject to }\; A_k s = -c_k$$

along with associated Lagrange multiplier estimates y_{k+1}

Set $x_{k+1} = x_k + s_k$ and increase k by 1

ADVANTAGES

- simple
- fast
  - quadratically convergent with $B_k = H(x_k, y_k)$
  - superlinearly convergent with good $B_k \approx H(x_k, y_k)$
    - don't actually need $B_k \longrightarrow H(x_k, y_k)$

PROBLEMS WITH PURE SQP

- how to choose B_k?
- what if QP_k is unbounded from below? and when?
- how do we globalize this iteration?