Lecture 13 and 14: Interior point methods for inequality constrained optimization

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Nonconvex inequality-constrained problems

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad c(x) \geq 0, \qquad \text{(iCP)}$$

where $f : \mathbb{R}^n \to \mathbb{R}$, $c = (c_1, \ldots, c_p) : \mathbb{R}^n \to \mathbb{R}^p$ smooth.

- ignore (linear) equality constraints for simplicity.
- $\Omega := \{x : \ c(x) \geq 0\}$ feasible set; let $\Omega^o := \{x : \ c(x) > 0\}$

- Assumption: strictly feasible set $\Omega^o \neq \emptyset$. [SCQ (Slater)]
- Attempt to find local solutions (at least KKT points) of (iCP).

For (each) $\mu > 0$, associate the logarithmic barrier subproblem

$$\min_{x \in \mathbb{R}^n} f_\mu(x) := f(x) - \mu \sum_{i=1}^{p} \log c_i(x) \ \text{ subject to } \ c(x) > 0. \qquad (\text{iCP}_\mu)$$

- (iCP$_\mu$) is essentially an unconstrained problem as each $c_i(x) > 0$ is enforced by the corresponding log barrier term of f_μ.

The logarithmic barrier function for (iCP)

Assume $x(\mu)$ minimizes the barrier problem

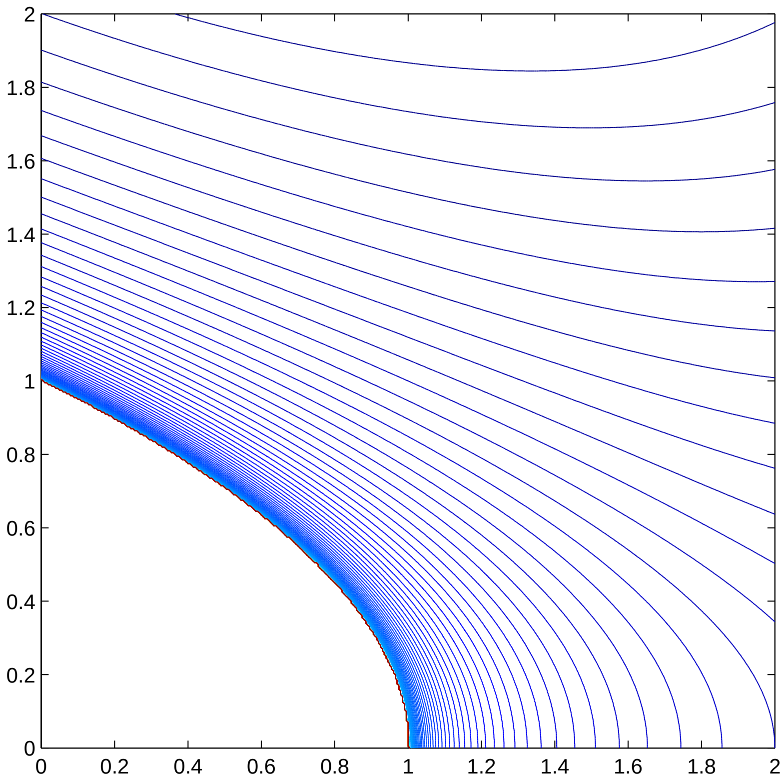
$$\min_{x \in \mathbb{R}^n} f_\mu(x) = f(x) - \mu \sum_{i=1}^{n} \log c_i(x) \text{ subject to } c(x) > 0. \quad \text{(iCP}_\mu\text{)}$$

Since ($c_i(x) \to 0 \implies -\log c_i(x) \to +\infty$), $x(\mu)$ must be "well inside" the feasible set Ω, "far" from the boundaries of Ω, especially when $\mu > 0$ is "large". Strict feasibility well-ensured!

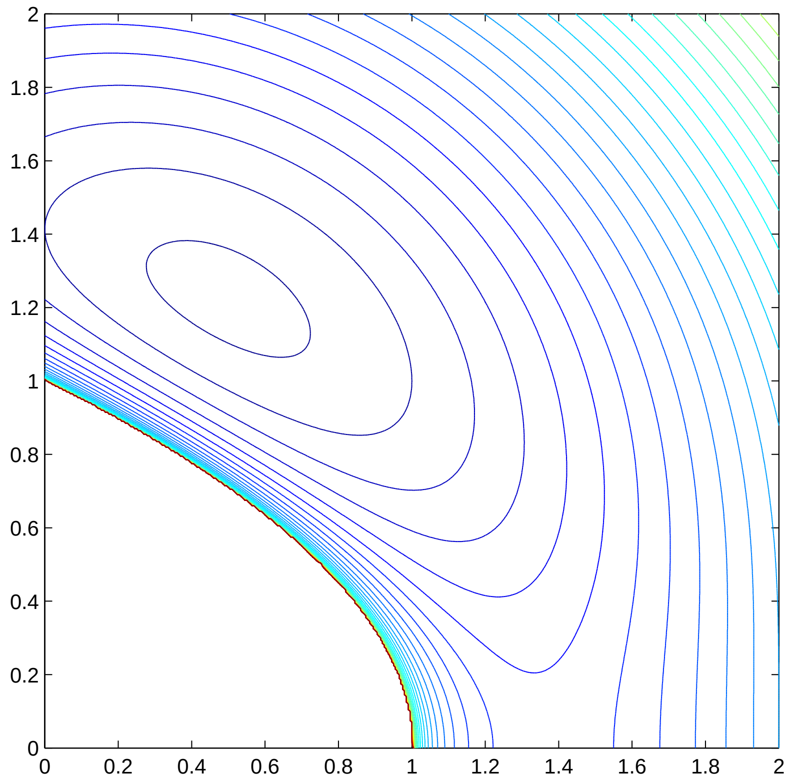
When μ "small", $\mu \to 0$: the term $f(x)$ "dominates" the log barrier terms in the objective of (iCP$_\mu$) $\implies x(\mu)$ "close" to the optimal boundary of Ω. [This also causes ill-conditioning ...]

• Subject to conditions, some minimizers of f_μ converge to local solutions of (iCP), as $\mu \to 0$. But f_μ may have other stationary points, useless for our purposes.

Contours of the barrier function f_μ - an example



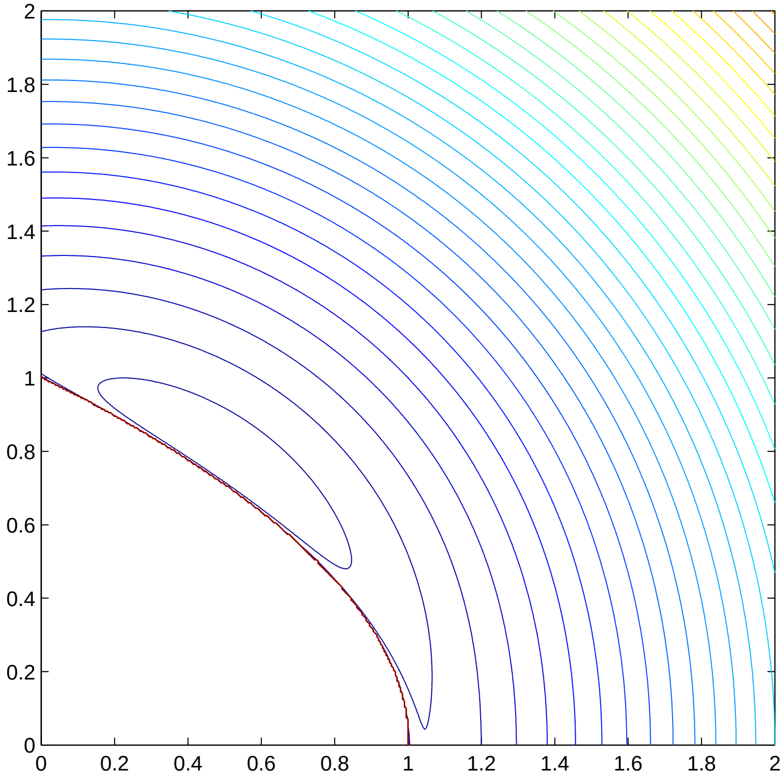
$\mu = 10$



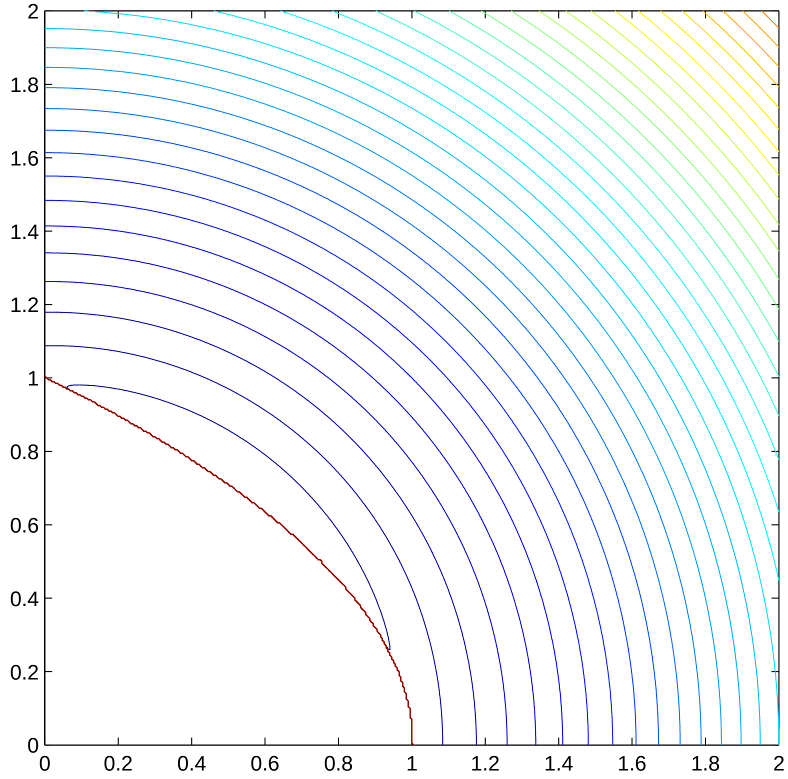
$\mu = 1$

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

Contours of the barrier function f_μ - an example...



$\mu = 0.1$



$\mu = 0.01$

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

Optimality conditions for (iCP) and (iCP$_\mu$)

$$f_\mu(x) := f(x) - \mu \sum_{i=1}^{p} \log c_i(x) \Longrightarrow$$

$$\nabla f_\mu(x) = \nabla f(x) - \sum_{i=1}^{p} \frac{\mu}{c_i(x)} \nabla c_i(x) = \nabla f(x) - \mu J(x)^\top c^{-1}(x),$$

where $J(x)$ Jacobian of $c(x)$, $c^{-1}(x) := (1/c_1(x), \ldots, 1/c_p(x))$.

First-order necessary optimality conditions for (iCP$_\mu$): [=uncons.]

$x(\mu)$ local minimizer of $f_\mu \Longrightarrow \nabla f_\mu(x(\mu)) = 0 \Longleftrightarrow$
$\nabla f(x(\mu)) = \sum_{i=1}^{p} \frac{\mu}{c_i(x(\mu))} \nabla c_i(x(\mu))$ with $\frac{\mu}{c_i(x(\mu))} > 0$, $i = \overline{1,p}$.

First-order necessary optimality conditions for (iCP): [=KKT]

Assume $\Omega^o \neq \emptyset$. If x^* local minimizer of (iCP) $\Longrightarrow$
$\nabla f(x^*) = \sum_{i=1}^{p} \lambda_i^* \nabla c_i(x^*)$, $\lambda^* \geq 0$, $\lambda_i^* c_i(x^*) = 0$, $i = \overline{1,p}$.

If x^* (nondegenerate) local min. of (iCP) (2nd order sufficient optimality conditions), $\frac{\mu}{c_i(x(\mu))} \to \lambda_i^*$, $i = \overline{1,p}$, as $\mu \to 0$.
Moreover ...

The path of barrier minimizers exists locally

... under second order sufficient optimality conditions at $x^* \in \Omega$, the central path of f_μ-minimizers $\{x(\mu) : \mu_\epsilon > \mu > 0\}$ exists, for μ_ϵ sufficiently small, and $x(\mu) \to x^*$, as $\mu \to 0$.

Theorem 27. (Local existence of central path) Assume that $\Omega^o \neq \emptyset$, and x^* is a local minimizer of (iCP) s. t.

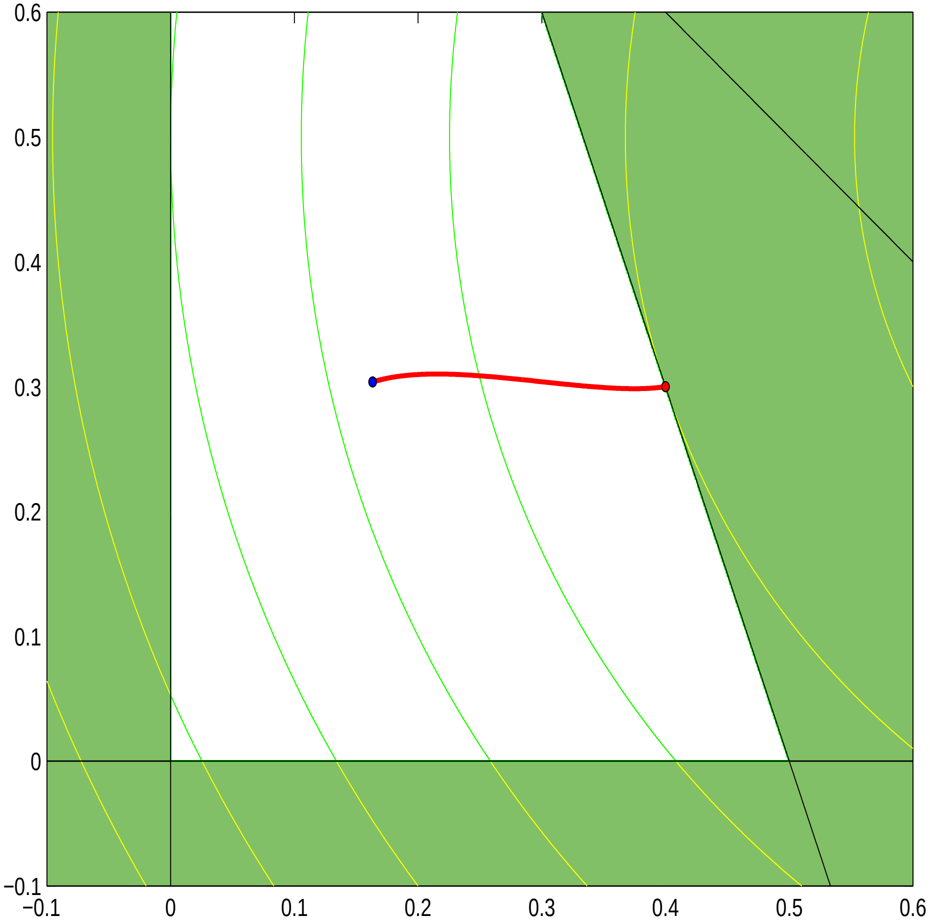
(a) $\lambda_i^* > 0$ if $c_i(x^*) = 0$.

(b) $\nabla c_i(x^*)$, $i \in \mathcal{A} := \{i \in \{1, \ldots, p\} : c_i(x^*) = 0\}$, are linearly independent. [LICQ]

(c) $\exists \alpha > 0$ such that $s^\top \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) s \geq \alpha \|s\|^2$, where s such that $J(x^*)_\mathcal{A} s = 0$, and $\nabla^2_{xx} \mathcal{L}$ is the Hessian of the Lagragian function of (iCP).

Then a unique, continuously differentiable vector function $x(\mu)$ of minimizers of f_μ exists in a neighbourhood of $\mu = 0$ and $x(\mu) \to x^*$ as $\mu \to 0$. □

Central path trajectory



$$\min (x_1 - 1)^2 + (x_2 - 0.5)^2$$

subject to $x_1 + x_2 \leq 1$

$$3x_1 + x_2 \leq 1.5$$

$$(x_1, x_2) \geq 0$$

Central path trajectory $x(\mu)$ for all $\mu > 0$.

Central path trajectory - nonconvex case

Central path trajectory $x(\mu)$ for all $\mu > 0$.

$$\min -2(x_1 - 0.25)^2 + 2(x_2 - 0.5)^2$$

$$\text{subject to } \; x_1 + x_2 \leq 1$$

$$3x_1 + x_2 \leq 1.5$$

$$(x_1, x_2) \geq 0$$

Basic barrier method (Fiacco-McCormick, 1960s)

Given $\mu^0 > 0$, let $k = 0$. Until "convergence" do:

- Choose $0 < \mu^{k+1} < \mu^k$.
- Find x_0^k such that $c(x_0^k) > 0$ (possibly, $x_0^k := x^k$).
- Starting from x_0^k, use an unconstrained minimization algorithm to find an "approximate" minimizer x^{k+1} of $f_{\mu^{k+1}}$. Let $k := k + 1$.

Must have $\mu^k \to 0$, $k \to 0$. $\mu^{k+1} := 0.1\mu^k$, $\mu^{k+1} := (\mu^k)^2$, etc.

Algorithms for minimizing f_μ: take Newton steps inside

• Linesearch methods: use special linesearch to cope with singularity of the log.

• Trust region methods: "shape" trust region to cope with contours of the singularity of the log. Reject points for which $c(x^k + s^k)$ is not positive.

A convergence result for the barrier algorithm

Theorem 28. (Global convergence of barrier algorithm)
Apply the basic barrier algorithm to the (iCP). Assume that

$f, c \in \mathcal{C}^2$, $\lambda_i^k = \dfrac{\mu^k}{c_i(x^k)}$, $i = \overline{1, p}$, and

$$\|\nabla f_{\mu^k}(x^k)\| \leq \epsilon^k, \text{ where } \epsilon^k \to 0, k \to \infty$$

and also that $\mu^k \to 0$ as $k \to \infty$. Moreover, assume that $x^k \to x^*$, where $\nabla c_i(x^*)$, $i \in \mathcal{A}$, are linearly independent, where $\mathcal{A} := \{i : c_i(x^*) = 0\}$ (ie LICQ).

Then x^* is a KKT point of (iCP) and $\lambda^k \to \lambda^*$, where λ^* is the vector of Lagrange multipliers of x^*. □

A convergence result for the barrier algorithm

Proof of Theorem 28. Let $\mathcal{A} = \{i : c_i(x^*) = 0\}$ (active constraints) and $\mathcal{I} = \{1, \ldots, p\} \setminus \mathcal{A}$ (inactive). Let $J_{\mathcal{A}}(x)$ denote the Jacobian of the active constraints and its pseudo-inverse

$$J_{\mathcal{A}}(x)^+ = (J_{\mathcal{A}}(x) J_{\mathcal{A}}(x)^T)^{-1} J_{\mathcal{A}}(x).$$

$J_{\mathcal{A}}(x^*)$ is full rank (it is $p_a \times n$ where $p_a = |\mathcal{A}|$ and so $p_a \leq n$) $\Longrightarrow J_{\mathcal{A}}(x^*)^+$ well-defined and $J_{\mathcal{A}}(x^k)^+$ well-defined and continuous for all k sufficiently large, due also to $x^k \to x^*$.
Define $\lambda^*_{\mathcal{A}} = J_{\mathcal{A}}(x^*)^+ \nabla f(x^*)$ and $\lambda^*_{\mathcal{I}} = 0$.
$x^k \to x^* \Longrightarrow c_i(x^k) \to c_i(x^*)$ and so for $i \in \mathcal{I}$, $c_i(x^k) \geq \frac{1}{2} c_i(x^*)$ for all k sufficiently large. Furthermore, for all k sufficiently large,

$$\|\lambda^k_{\mathcal{I}}\| = \sqrt{\sum_{i \in \mathcal{I}} \frac{(\mu^k)^2}{c_i(x^k)^2}} \leq \frac{2\mu^k \sqrt{|\mathcal{I}|}}{\min_{i \in \mathcal{I}} c_i(x^*)} := \mu^k \epsilon^*. \quad (\Diamond)$$

A convergence result for the barrier algorithm

Proof of Theorem 28. (continued)
Note that $J(x^k)^T = (J_{\mathcal{A}}(x^k)^T \; J_{\mathcal{I}}(x^k)^T)$ and $\lambda^k = (\lambda^k_{\mathcal{A}} \; \lambda^k_{\mathcal{I}})$ and so $J(x^k)^T\lambda^k = J_{\mathcal{A}}(x^k)^T\lambda^k_{\mathcal{A}} + J_{\mathcal{I}}(x^k)^T\lambda^k_{\mathcal{I}}$.

$$\|\nabla f(x^k) - J_{\mathcal{A}}(x^k)^T\lambda^k_{\mathcal{A}}\| \leq \|\nabla f(x^k) - J(x^k)^T\lambda^k\| + \|J_{\mathcal{I}}(x^k)^T\lambda^k_{\mathcal{I}}\|$$

$$= \|\nabla f_{\mu^k}(x^k)\| + \|J_{\mathcal{I}}(x^k)^T\lambda^k_{\mathcal{I}}\| \leq \|\nabla f_{\mu^k}(x^k)\| + 2\|J_{\mathcal{I}}(x^*)\| \cdot \|\lambda^k_{\mathcal{I}}\|$$

$$\leq \epsilon^k + 2\epsilon^*\|J_{\mathcal{I}}(x^*)\|\mu^k := \bar{\epsilon}^k, \quad (\Diamond\Diamond)$$

where in the penultimate inequality, we used $\|J_{\mathcal{I}}(x^k)^T\| \leq \|J_{\mathcal{I}}(x^k) - J_{\mathcal{I}}(x^*)\| + \|J_{\mathcal{I}}(x^*)\| \leq 2\|J_{\mathcal{I}}(x^*)\|$ since $x^k \to x^*$ and J continuous; in the last inequality, we used ($\Diamond$) and the termination condition for the inner minimization of the barrier subproblem. Thus

A convergence result for the barrier algorithm

Proof of Theorem 28. (continued)

$$\|J_{\mathcal{A}}(x^k)^+\nabla f(x^k) - \lambda^k_{\mathcal{A}}\| = \|J_{\mathcal{A}}(x^k)^+(\nabla f(x^k) - J_{\mathcal{A}}(x^k)^T\lambda^k_{\mathcal{A}})\|$$
$$\leq 2\|J_{\mathcal{A}}(x^*)^+\| \cdot \|\nabla f(x^k) - J_{\mathcal{A}}(x^k)^T\lambda^k_{\mathcal{A}}\| \leq 2\|J_{\mathcal{A}}(x^*)^+\|\bar{\epsilon}^k.$$

Finally,

$$\begin{aligned}\|\lambda^k_{\mathcal{A}} - \lambda^*_{\mathcal{A}}\| &\leq \|\lambda^k_{\mathcal{A}} - J_{\mathcal{A}}(x^k)^+\nabla f(x^k)\| \\ &\quad +\|J_{\mathcal{A}}(x^k)^+\nabla f(x^k) - J_{\mathcal{A}}(x^*)^+\nabla f(x^*)\| \\ &\leq 2\|J_{\mathcal{A}}(x^*)^+\|\bar{\epsilon}^k + \alpha^k \longrightarrow 0,\end{aligned}$$

since $\mu^k \to 0$, $\epsilon^k \to 0$, $x^k \to x^*$, J^+ and ∇f are continuous.
From $(\Diamond)$ and $\mu^k \to 0$, $\lambda^k_{\mathcal{I}} \to 0 = \lambda^*_{\mathcal{I}}$.
Passing to the limit in $(\Diamond\Diamond)$, we deduce $\nabla f(x^*) - J_{\mathcal{A}}(x^*)^T\lambda^*_{\mathcal{A}} = 0$. Since $c(x^k) > 0$, then $c(x^*) \geq 0$; from $\lambda^k > 0$, we deduce $\lambda^* \geq 0$. $\lambda^*_i c_i(x^*) = 0$ for all i by construction.

Minimizing the barrier function f_μ

Use Newton's method with linesearch or trust-region.

$$f_\mu(x) := f(x) - \mu \sum_{i=1}^p \log c_i(x) \Longrightarrow$$

$$\nabla f_\mu(x) = \nabla f(x) - \sum_{i=1}^p \frac{\mu}{c_i(x)} \nabla c_i(x) = \nabla f(x) - \mu J(x)^\top c^{-1}(x),$$

where $J(x)$ is the Jacobian of $c(x)$. Let $C^j(x) :=$ diag$(c^j(x))$.

$$\begin{aligned}\nabla^2 f_\mu(x) &= \nabla^2 f(x) - \sum_{i=1}^p \frac{\mu}{c_i(x)} \nabla^2 c_i(x) + \sum_{i=1}^p \frac{\mu}{c_i(x)^2} \nabla c_i(x) \nabla c_i(x)^\top \\ &= \nabla^2 f(x) - \sum_{i=1}^p \frac{\mu}{c_i(x)} \nabla^2 c_i(x) + \mu J(x)^\top C^{-2}(x) J(x).\end{aligned}$$

Given x such that $c(x) > 0$, the Newton direction for f_μ solves

$$\nabla^2 f_\mu(x) s = -\nabla f_\mu(x) \qquad [\mu = \mu^{k+1}]$$

Estimates of the Lagrange multipliers: $\lambda_i(x) := \mu / c_i(x)$, $i = \overline{1,p}$.

Minimizing the barrier function f_μ ...

$\Longrightarrow \nabla f_\mu(x) = \nabla f(x) - J(x)^T \lambda(x)$

$\Longrightarrow$ gradient of Lagrangian of (iCP) at $(x, \lambda(x))$.

Recall: the Lagragian function of (iCP)

$$\mathcal{L}(x, \lambda) := f(x) - \sum_{i=1}^{p} \lambda_i c_i(x).$$

$\Longrightarrow \nabla^2 f_\mu(x) = \nabla^2 \mathcal{L}(x, \lambda(x)) + \mu J(x)^\top C^{-2}(x) J(x),$

As $\mu \to 0$, $\dfrac{\mu}{c_i(x)^2} \to 0$ for all $i \in \mathcal{A}$ (active),

and so $\quad \mu J(x)^\top C^{-2}(x) J(x) \to \infty$ as $\mu \to 0$.

Potential difficulties

I. Ill-conditioning of the Hessian of f_μ

Asymptotic estimates of the eigenvalues of $\nabla^2 f_{\mu^k}(x^k)$:

'Fact' (Th 5.2, Gould Ref.) $\Longrightarrow$

- $p_a = |\mathcal{A}|$ eigenvalues of $\nabla^2 f_{\mu^k}(x^k)$ tend to infinity as $k \to \infty$.
- the condition number of $\nabla^2 f_{\mu^k}(x^k)$ is $\mathcal{O}(1/\mu^k)$
  $\Longrightarrow$ it blows up as $k \to \infty$.
  $\Longrightarrow$ may not be able to compute x^k accurately.

This is the main reason for the barrier methods falling out of favour with the nonlinear optimization community in the 1960s.

Potential difficulties ...

II. Poor starting points

Recall we need x_0^k starting point for the (approximate) minimization of $f_{\mu^{k+1}}$, after the barrier parameter μ^k has been decreased to μ^{k+1}.

It can be shown that the current computed iterate x^k appears to be a very poor choice of starting point x_0^k, in the sense that the full Newton step $x^k + s^k$ will be <u>asymptotically infeasible</u> (i. e., $c(x^k + s^k) < 0$) whenever $\mu^{k+1} < 0.5\mu^k$ (i. e., for any meaningful decrease in μ^k). Thus the barrier method is unlikely to converge fast.

Solution to troubles I & II: use primal-dual IPMs.

Perturbed optimality conditions

Recall first order necessary conditions for (iCP$_\mu$):

$x(\mu)$ local minimizer of $f_\mu \Longrightarrow \nabla f_\mu(x(\mu)) = 0 \Longleftrightarrow$
$\nabla f(x(\mu)) = \mu J(x(\mu))^\top c^{-1}(x(\mu))$. Let $\lambda(\mu) := \mu c^{-1}(x(\mu))$.

Thus $(x(\mu), \lambda(\mu))$ satisfy:

$$\begin{cases} \nabla f(x) - J(x)^\top \lambda = 0, \\ c_i(x)\lambda_i = \mu, \ i = \overline{1,p}, \end{cases} \qquad \text{(OPT}_\mu\text{)}$$

$$c(x) > 0, \quad \lambda > 0.$$

Compare with the KKT system for (iCP):

$$\begin{cases} \nabla f(x) - J(x)^\top \lambda = 0, \\ c_i(x)\lambda_i = \mu, \ i = \overline{1,p}, \end{cases} \qquad \text{(KKT)}$$

$$c(x) \geq 0, \quad \lambda \geq 0.$$

Primal-dual path-following methods (1990s)

Satisfy $c(x) > 0$ and $\lambda > 0$, and use Newton's method to solve the system $\quad e := (1, \dots, 1)^T$

$$\begin{cases} \nabla f(x) - J(x)^\top \lambda = 0, \\ C(x)\lambda = \mu e, \end{cases} \qquad (\mathrm{OPT}_\mu)$$

i. e., the Newton direction $(dx, d\lambda)$ satisfies

$$\begin{pmatrix} \nabla^2 \mathcal{L}(x,\lambda) & -J(x)^\top \\ \Lambda J(x) & C(x) \end{pmatrix} \begin{pmatrix} dx \\ d\lambda \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - J(x)^\top \lambda \\ C(x)\lambda - \mu e \end{pmatrix},$$

where $\Lambda := \mathrm{diag}(\lambda)$. Eliminating $d\lambda$, we deduce

$$(\nabla^2 \mathcal{L}(x,s) + J(x)^\top C^{-1}(x) \Lambda J(x)) dx = -(\nabla f(x) - \mu J(x)^\top c^{-1}(x)).$$

Primal-dual versus primal methods

Primal-dual:

$$(\nabla^2 \mathcal{L}(x, \lambda) + J(x)^\top C^{-1}(x) \Lambda J(x)) dx^{pd} = -\nabla \mathcal{L}(x, \lambda(x)).$$

Primal:

$$(\nabla^2 \mathcal{L}(x, \lambda(x)) + J(x)^\top C^{-1}(x) \Lambda(x) J(x)) dx^{p} = -\nabla \mathcal{L}(x, \lambda(x)),$$

where $\lambda(x) := \mu c^{-1}(x)$.

$\Longrightarrow$ In PD methods, changes to the estimates s of the Lagrange multipliers are computed explicitly on each iteration. In primal methods, they are updated from implicit information. Makes a huge difference!

• For PD IPMs, $x_0^k := x^k$ is a good starting point for the subproblem solution. Ill-conditioning of the Hessian can be 'overlooked' by solving in the right subspaces.

Ill-conditioning revisited (non-examinable)

Ill-conditioning does not imply can't solve equations accurately!
Assume $\lambda_i^* > 0$ if $c(x^*) = 0$. Let $\mathcal{I} = \{i : c_i(x^*) > 0\}$. Drop x.

$$\begin{pmatrix} \nabla^2 \mathcal{L} & -J^\top \\ \Lambda J^\top & C \end{pmatrix} \begin{pmatrix} dx \\ d\lambda \end{pmatrix} = - \begin{pmatrix} \nabla f - J^\top \lambda \\ C\lambda - \mu e \end{pmatrix} \Longrightarrow$$

$$\begin{pmatrix} \nabla^2 \mathcal{L} + J_\mathcal{I}^\top C_\mathcal{I}^{-1} \Lambda_\mathcal{I} J_\mathcal{I} & -J_\mathcal{A}^\top \\ J_\mathcal{A} & C_\mathcal{A} \Lambda_\mathcal{A}^{-1} \end{pmatrix} \begin{pmatrix} dx \\ d\lambda_\mathcal{A} \end{pmatrix} = - \begin{pmatrix} \nabla f - J_\mathcal{A}^\top s_\mathcal{A} - \mu J_\mathcal{I} c_\mathcal{I}^{-1} \\ c_\mathcal{A}(x) - \mu \lambda_\mathcal{A}^{-1} \end{pmatrix}$$

Note $C_\mathcal{I}^{-1}(x)$ and $\Lambda_\mathcal{A}^{-1}$ bounded above (as $x \to x^*$). Thus, in the limit,

$$\begin{pmatrix} \nabla^2 \mathcal{L} & -J_\mathcal{A}^\top \\ J_\mathcal{A}^\top & 0 \end{pmatrix} \begin{pmatrix} dx \\ d\lambda_\mathcal{A} \end{pmatrix} = - \begin{pmatrix} \nabla f - J_\mathcal{A}^\top \lambda_\mathcal{A} - \mu J_\mathcal{I} c_\mathcal{I}^{-1} \\ 0 \end{pmatrix}.$$

Note that this approach needs an accurate prediction of the active $\mathcal{A}$ and inactive $\mathcal{I}$ sets 'asymptotically' during the run of a primal-dual algorithm (not so easy!)

Primal-dual path-following methods

Choice of barrier parameter: $\mu^{k+1} = \mathcal{O}((\mu^k)^2)$
$\Longrightarrow$ Fast (superlinear) asymptotic convergence!

Several Newton iterations are performed for each value of μ (with linesearch or trust-region).

In implementations, it is essential to keep iterates away from boundaries early in the algorithm (else iterates may get trapped near the boundary $\Rightarrow$ slow convergence!)

The computation of initial starting point x^0 satisfying $c(x^0) > 0$ is nontrivial. Various heuristics exist.

Powerful software available: IPOPT, KNITRO etc.

Linear Programming (LP): IPMs solve LP in polynomial time!

The simplex versus interior point methods for LP

- worst-case complexity: exponential versus polynomial for LP (in problem dimension/length of input);
  - the Klee-Minty example (1972): the simplex method has exponential running time in the worst-case; linear polynomial in the average case
  - IPMs: Karmarkar (1984), A New Polynomial-Time Algorithm for Linear Programming, *Combinatorica*.
    Khachiyan (the ellipsoid method, 1979).
    Renegar (best-known worst-case complexity bound).
    Central path is unique and global; Newton's method for barrier function can be precisely quantified.
- IPMs solve very large-scale LPs;
  - numerically-observed average complexity: log(LP dimension) iterations.
- each IPM iteration more expensive than the simplex one.