Methods for Constrained Optimization
Numerical Optimization Lectures 3-4

Coralia Cartis, University of Oxford

INFOMM CDT: Modelling, Analysis and Computation of

Continuous Real-World Problems

Problems and solutions

$$\text{minimize } f(x) \quad \text{subject to } x \in \Omega \subseteq \mathbb{R}^n. \qquad (\dagger)$$

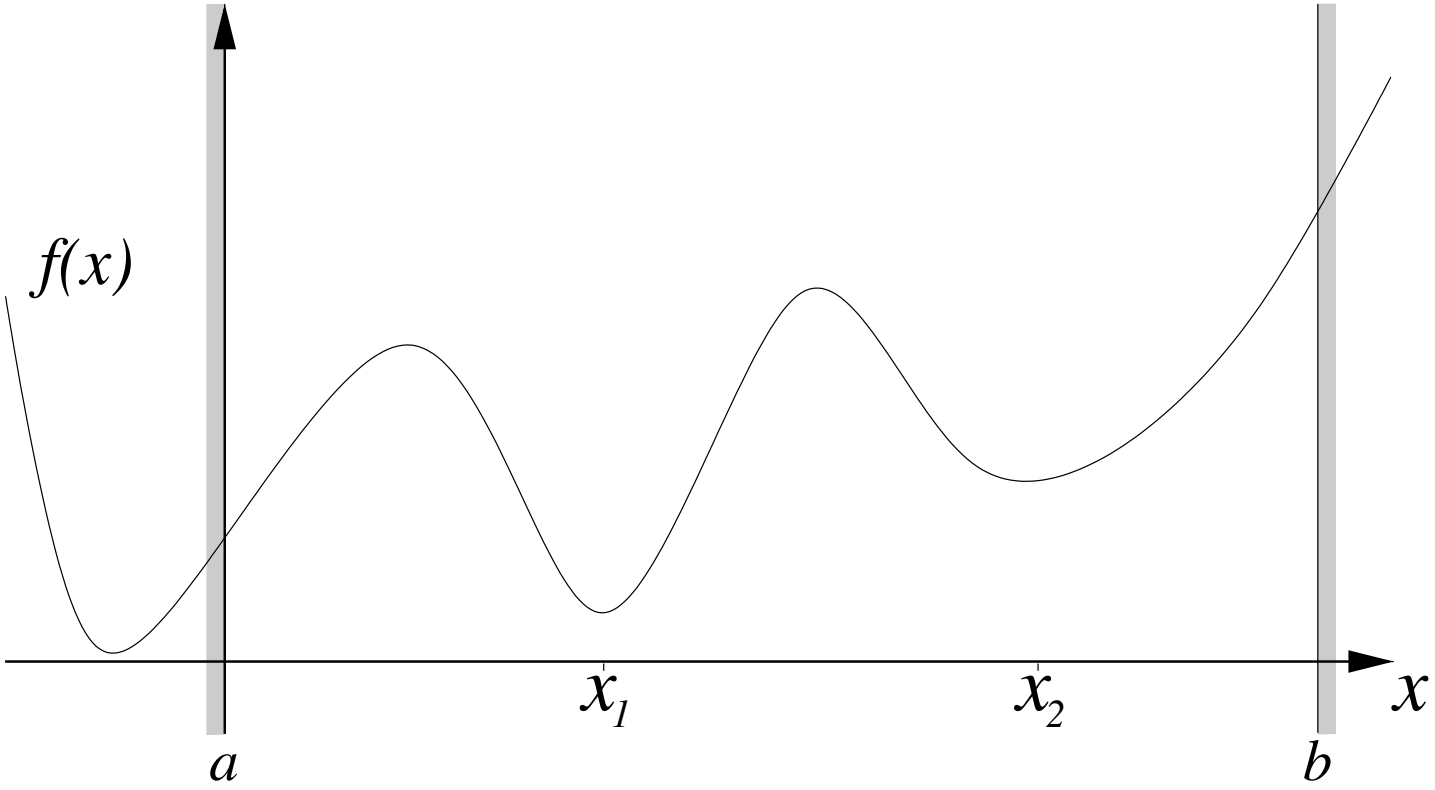
- $f : \Omega \to \mathbb{R}$ is (sufficiently) smooth.
- f objective; x variables.
- Ω feasible set determined by finitely many (equality and/or inequality) constraints.

x^* global minimizer of f over $\Omega \implies f(x) \geq f(x^*), \forall x \in \Omega$.

x^* local minimizer of f over $\Omega \implies$
$\exists N(x^*, \delta)$ such that $f(x) \geq f(x^*)$, for all $x \in \Omega \cap N(x^*, \delta)$.
- $N(x^*, \delta) := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$.

Example problem in one dimension

Example : $\min f(x)$ **subject to** $a \leq x \leq b$.



- The feasible region Ω is the interval $[a, b]$.
- The point x_1 is the global minimizer; x_2 is a local (non-global) minimizer; $x = a$ is a constrained local minimizer.

Optimality conditions for constrained problems

== algebraic characterizations of solutions $\longrightarrow$ suitable for computations.

- provide a way to guarantee that a candidate point is optimal (sufficient conditions)
- indicate when a point is not optimal (necessary conditions)

$$\text{minimize}_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad c_E(x) = 0, \quad c_I(x) \geq 0. \tag{CP}$$

- $f : \mathbb{R}^n \to \mathbb{R}$, $c_E : \mathbb{R}^n \to \mathbb{R}^m$ and $c_I : \mathbb{R}^n \to \mathbb{R}^p$ (suff.) smooth;
  - $c_I(x) \geq 0 \Leftrightarrow c_i(x) \geq 0$, $i \in I$.
  - $\Omega := \{x : \ c_E(x) = 0, \ c_I(x) \geq 0\}$ feasible set of the problem.

Optimality conditions for constrained problems

unconstrained problem $\longrightarrow \hat{x}$ stationary point ($\nabla f(\hat{x}) = 0$).

constrained problem $\longrightarrow \hat{x}$ Karush-Kuhn-Tucker (KKT) point.

Definition: $\hat{x}$ KKT point of (CP) if there exist $\hat{y} \in \mathbb{R}^m$ and $\hat{\lambda} \in \mathbb{R}^p$ such that $(\hat{x}, \hat{y}, \hat{\lambda})$ satisfies

$$\nabla f(\hat{x}) = \sum_{j \in E} \hat{y}_j \nabla c_j(\hat{x}) + \sum_{i \in I} \hat{\lambda}_i \nabla c_i(\hat{x}),$$

$$c_E(\hat{x}) = 0, \quad c_I(\hat{x}) \geq 0,$$

$$\hat{\lambda}_i \geq 0, \quad \hat{\lambda}_i c_i(\hat{x}) = 0, \quad \text{for all } i \in I.$$

- Let $\mathcal{A} := E \cup \{i \in I : c_i(\hat{x}) = 0\}$ index set of active constraints at $\hat{x}$; $c_j(\hat{x}) > 0$ inactive constraint at $\hat{x} \Rightarrow \hat{\lambda}_j = 0$. Then $\sum_{i \in I} \hat{\lambda}_i \nabla c_i(\hat{x}) = \sum_{i \in I \cap \mathcal{A}} \hat{\lambda}_i \nabla c_i(\hat{x})$.
- $J(x) = \left(\nabla c_i(x)^T\right)_i$ Jacobian matrix of constraints c. Thus $\sum_{j \in E} \hat{y}_j \nabla c_j(\hat{x}) = J_E(x)^T \hat{y}$ and $\sum_{i \in I} \hat{\lambda}_i \nabla c_i(\hat{x}) = J_I(x)^T \hat{\lambda}$.

Optimality conditions for constrained problems ...

$\hat{x}$ KKT point $\longrightarrow$ $\hat{y}$ and $\hat{\lambda}$ Lagrange multipliers of the equality and inequality constraints, respectively.

$\hat{y}$ and $\hat{\lambda} \longrightarrow$ sensitivity analysis.

$\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ Lagrangian function of (CP),

$$\mathcal{L}(x, y, \lambda) := f(x) - y^\top c_E(x) - \lambda^\top c_I(x), \quad x \in \mathbb{R}^n.$$

Thus $\nabla_x \mathcal{L}(x, y, \lambda) = \nabla f(x) - J_E(x)^\top y - J_I(x)^\top \lambda,$

and $\hat{x}$ KKT point of (CP) $\implies$ $\nabla_x \mathcal{L}(\hat{x}, \hat{y}, \hat{\lambda}) = 0$
(i. e., $\hat{x}$ is a stationary point of $\mathcal{L}(\cdot, \hat{y}, \hat{\lambda})$).

- duality theory...

An illustration of the KKT conditions

$$\min_{x \in \mathbb{R}^2} (x_1 - 2)^2 + (x_2 - 0.5(3 - \sqrt{5}))^2 \quad \text{subject to}$$

$$-x_1 - x_2 + 1 \geq 0, \;\; x_2 - x_1^2 \geq 0. \qquad (*)$$

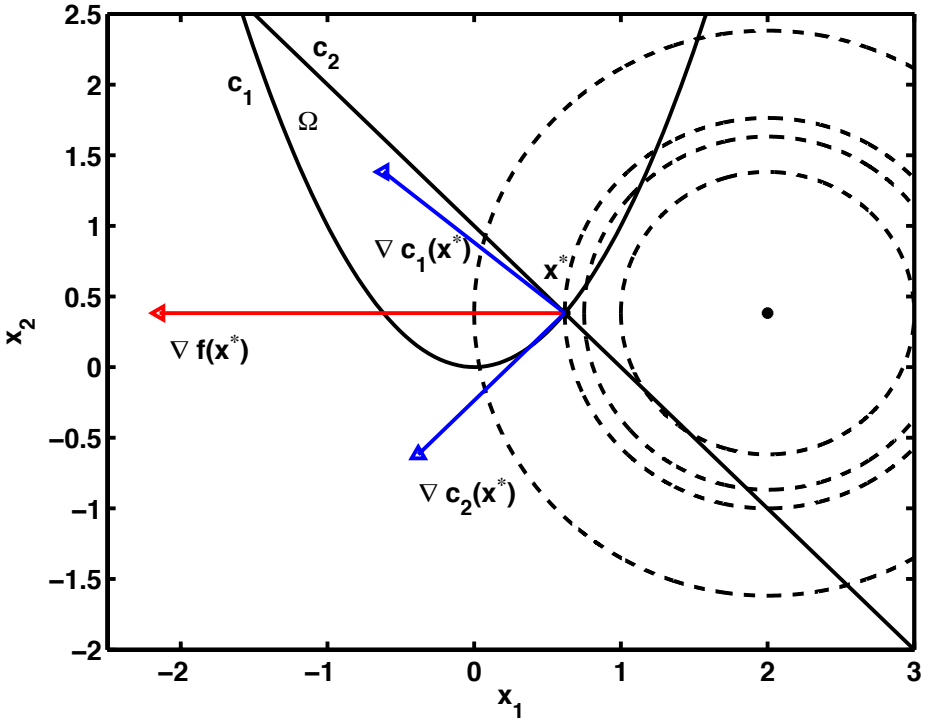
$x^* = \frac{1}{2}(-1+\sqrt{5}, 3-\sqrt{5})^\top$:

- global solution of $(*)$,
- KKT point of $(*)$.

$\nabla f(x^*) = (-5 + \sqrt{5}, 0)^\top,$

$\nabla c_1(x^*) = (1 - \sqrt{5}, 1)^\top,$

$\nabla c_2(x^*) = (-1, -1)^\top.$



$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*) + \lambda_2^* \nabla c_2(x^*)$, with $\lambda_1^* = \lambda_2^* = \sqrt{5} - 1 > 0$.

$c_1(x^*) = c_2(x^*) = 0$: constraints are active at x^*.

An illustration of the KKT conditions ...

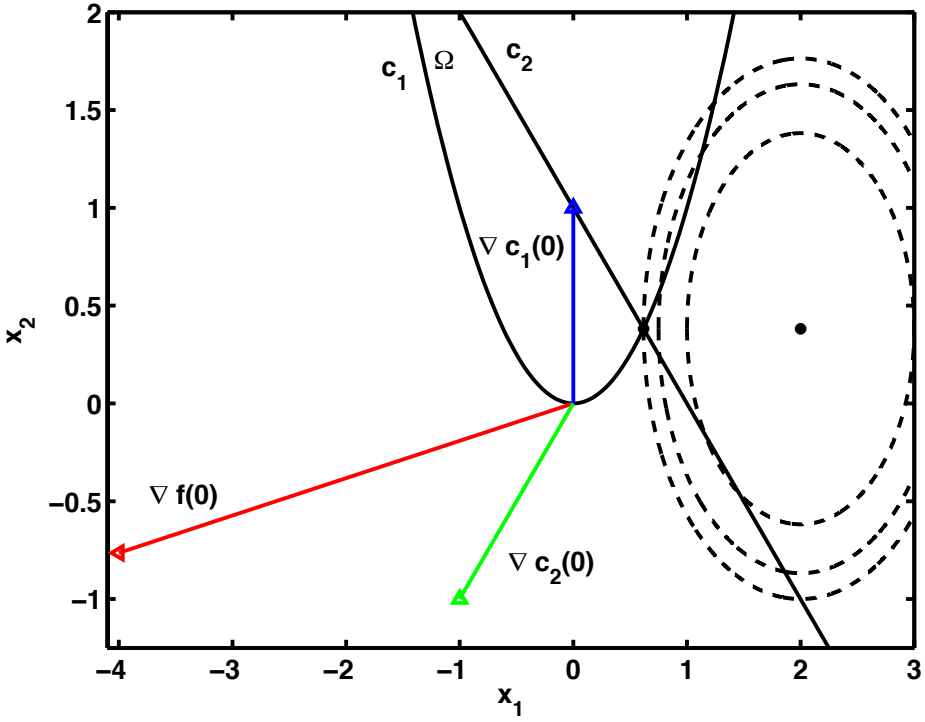
$$\min_{x \in \mathbb{R}^2} (x_1 - 2)^2 + (x_2 - 0.5(3 - \sqrt{5}))^2 \quad \text{subject to}$$

$$-x_1 - x_2 + 1 \geq 0, \;\; x_2 - x_1^2 \geq 0. \qquad (*)$$

$x := (0, 0)^\top$
is NOT a KKT point of $(*)$!

$c_1(x) = 0$: active at x.
$c_2(x) = 1$: inactive at x.
$\Longrightarrow \lambda_2 = 0$ and
$\nabla f(x) = \lambda_1 \nabla c_1(x)$,
with $\lambda_1 \geq 0$.

$\Downarrow$



Contradiction with $\nabla f(x) = (-4, \sqrt{5} - 3)^\top$ and
$\nabla c_1(x) = (0, 1)^\top$.

Optimality conditions for constrained problems ...

In general, need constraints/feasible set of (CP) to satisfy regularity assumption called constraint qualification in order to derive optimality conditions.

Theorem (First order necessary conditions) Under suitable constraint qualifications,
x^* local minimizer of (CP) $\Longrightarrow$ x^* KKT point of (CP).

- Let (CP) with equalities only ($I = \emptyset$). Then feasible descent direction s at $x \in \Omega$ if $\nabla f(x)^T s < 0$ and $J_E(x)s = 0$.
- Let (CP). Then feasible descent direction s at $x \in \Omega$ if $\nabla f(x)^T s < 0$, $J_E(x)s = 0$ and $\nabla c_i(x)^T s \geq 0$ for all $i \in I \cap \mathcal{A}(x)$.

Constraint qualifications

- Proof of theorem needs (first-order) Taylor to linearize f and c_i along feasible paths/perturbations $x(\alpha)$ etc. Only correct if linearized approximation covers the essential geometry of the feasible set. CQs ensure this is the case.

  Examples:

- (CP) satisfies the Slater Constraint Qualification (SCQ) $\Longleftrightarrow$ if $\exists\, x$ s.t. $c_E(x) = 0$ and $c_I(x) > 0$ (i.e., $c_i(x) > 0$, $i \in I$).
- (CP) satisfies the Linear Independence Constraint Qualification (LICQ) $\Longleftrightarrow$ $\nabla c_i(x)$, $i \in \mathcal{A}(x)$, are linearly independent (at relevant x).

Both SCQ and LICQ fail for

$$\Omega = \{(x_1, x_2) : c_1(x) = 1 - x_1^2 - (x_2 - 1)^2 \geq 0;\ c_2(x) = -x_2 \geq 0\}.$$

[see Nocedal & Wright, Numerical Optimization for full details]

Optimality conditions for constrained problems ...

If the constraints of (CP) are linear in the variables, no constraint qualification is required.

Theorem (First order necessary conditions for linearly constrained problems) Let $(c_E, c_I)(x) := Ax - b$ in (CP). Then x^* local minimizer of (CP) $\implies$ x^* KKT point of (CP).

Let $A = (A_E, A_I)$ and $b = (b_E, b_I)$ corresponding to equality and inequality constraints.
KKT conditions for linearly-constrained (CP): x^* KKT point $\Leftrightarrow$ there exists (y^*, λ^*) such that

$$\nabla f(x^*) = A_E^T y^* + A_I^T \lambda^*,$$
$$A_E x^* - b_E = 0, \quad A_I x^* - b_I \geq 0,$$
$$\lambda^* \geq 0, \quad (\lambda^*)^T (A_I x^* - b_I) = 0.$$

Optimality conditions for convex problems

(CP) is a convex programming problem if and only if $f(x)$ is a convex function, $c_i(x)$ is a concave function for all $i \in I$ and $c_E(x) = Ax - b$.

- c_i is a concave function $\Leftrightarrow$ $(-c_i)$ is a convex function.
- (CP) convex problem $\Rightarrow$ Ω is a convex set.
- (CP) convex problem $\Rightarrow$ any local minimizer of (CP) is global.

First order necessary conditions are also sufficient for optimality when (CP) is convex.

Theorem. (Sufficient optimality conditions for convex problems: Let (CP) be a convex programming problem.
$\hat{x}$ KKT point of (CP) $\Longrightarrow$ $\hat{x}$ is a (global) minimizer of (CP). $\square$

Optimality conditions for nonconvex problems

• When (CP) is not convex, the KKT conditions are not in general sufficient for optimality
$\longrightarrow$ need positive definite Hessian of the Lagragian function along “feasible” directions.

Penalty methods

Nonlinear equality-constrained problems

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

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

- attempt to find local solutions (at least KKT points).
- constrained optimization $\longrightarrow$ conflict of requirements: objective minimization & feasibility of the solution.
  - easier to generate feasible iterates for linear equality and general inequality constrained problems;
  - very hard, even impossible, in general, when general equality constraints are present.

$\Longrightarrow$ form a single, parametrized and unconstrained objective, whose minimizers approach initial problem solutions as parameters vary (eg: barrier methods for (iCP)).

A penalty function for (eCP)

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad c(x) = 0. \qquad \text{(eCP)}$$

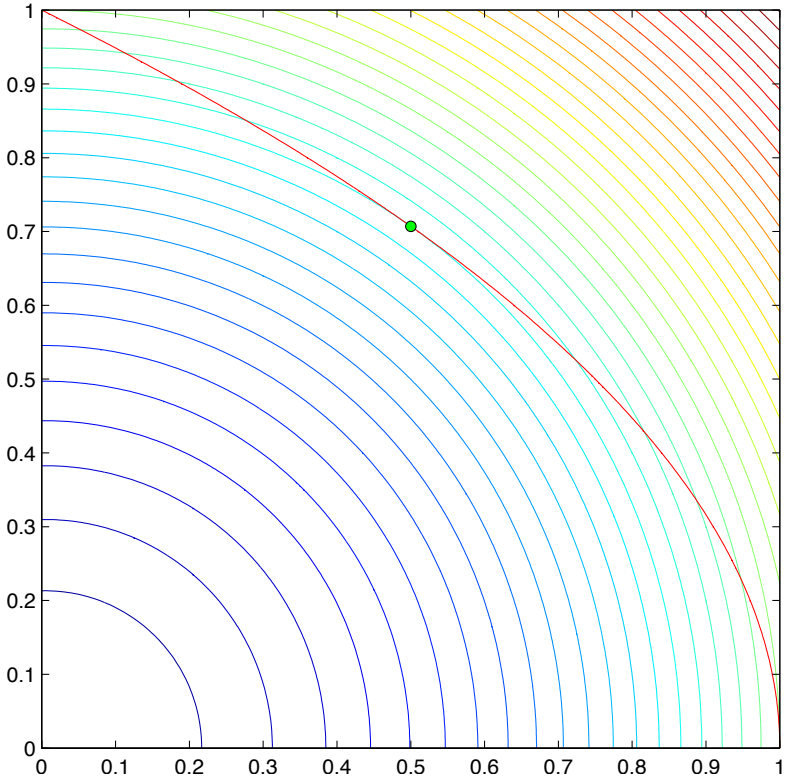
The quadratic penalty function:

$$\min_{x \in \mathbb{R}^n} \quad \Phi_\sigma(x) = f(x) + \frac{1}{2\sigma}\|c(x)\|^2, \qquad \text{(eCP}_\sigma\text{)}$$

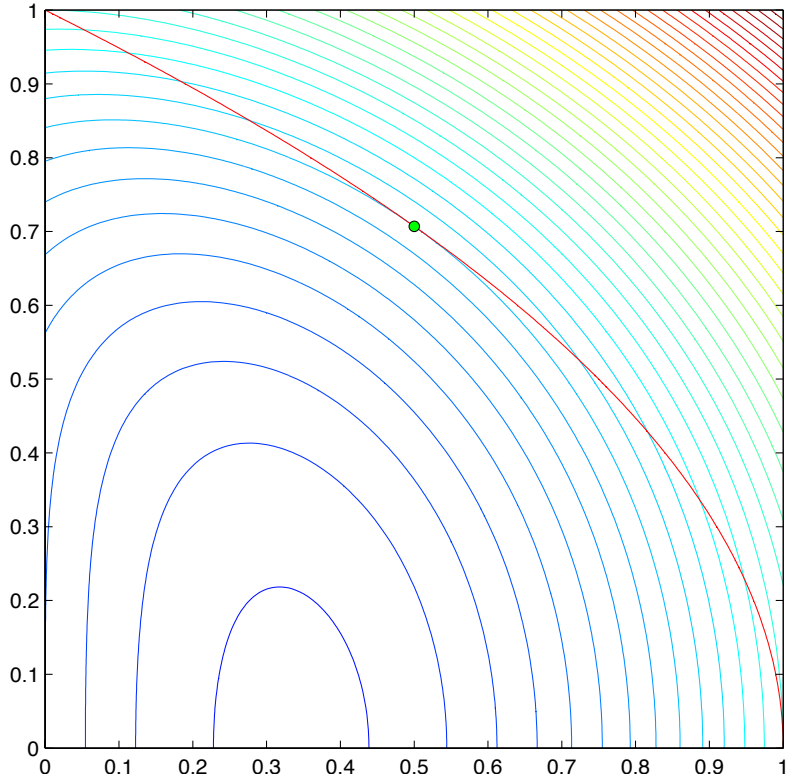
where $\sigma > 0$ penalty parameter.

- σ: penalty on infeasibility;
- $\sigma \longrightarrow 0$: 'forces' constraint to be satisfied and achieve optimality for f.
- Φ_σ may have other stationary points that are not solutions for (eCP); eg., when $c(x) = 0$ is inconsistent.

Contours of the penalty function Φ_σ - an example



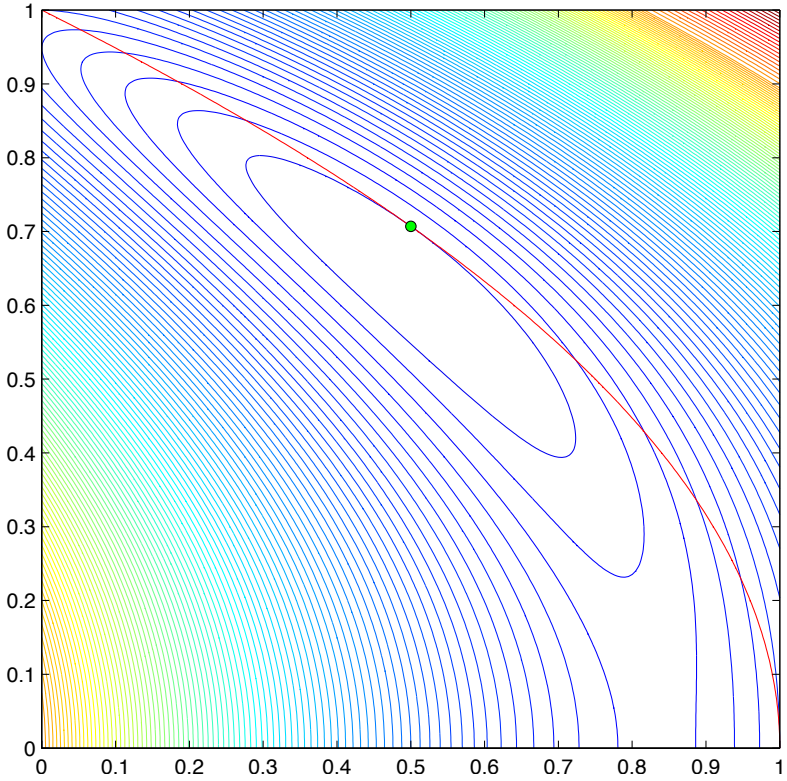
$\sigma = 100$



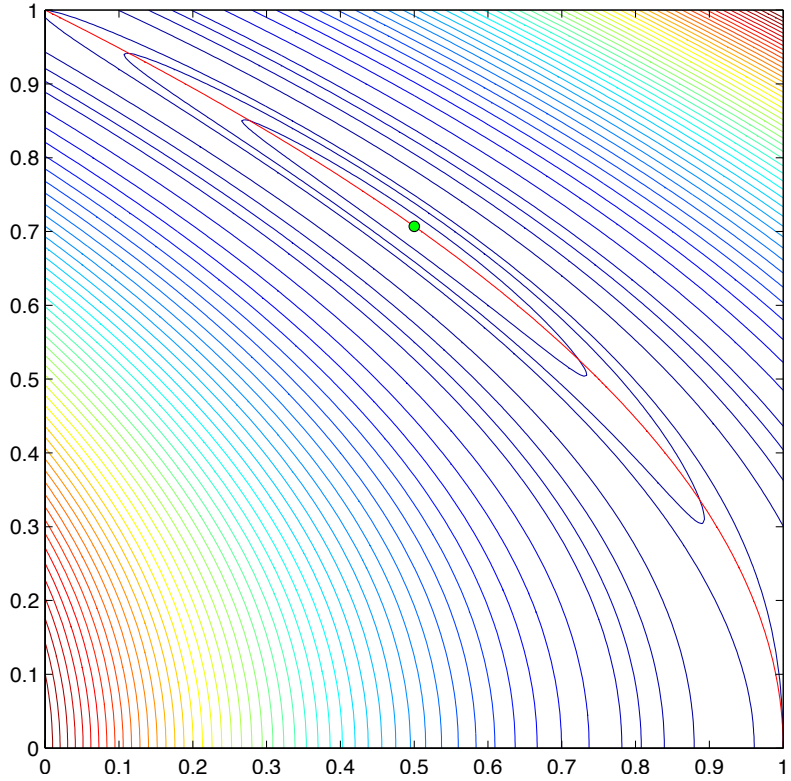
$\sigma = 1$

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

Contours of the penalty function Φ_σ - an example...



$\sigma = 0.1$



$\sigma = 0.01$

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

A quadratic penalty method

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

- Choose $0 < \sigma^{k+1} < \sigma^k$.
- Starting from x_0^k (possibly, $x_0^k := x^k$), use an unconstrained minimization algorithm to find an "approximate" minimizer x^{k+1} of $\Phi_{\sigma^{k+1}}$.
  Let $k := k + 1$. $\Diamond$

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

Algorithms for minimizing Φ_σ:

• Linesearch, trust-region methods.

• σ small: Φ_σ very steep in the direction of constraints' gradients, and so rapid change in Φ_σ for steps in such directions; implications for "shape" of trust region.

A convergence result for the penalty method

Theorem. (Global convergence of penalty method) Apply the basic quadratic penalty method to the (eCP). Assume that $f, c \in \mathcal{C}^1$, $y_i^k = -c_i(x^k)/\sigma^k$, $i = \overline{1, m}$, and

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

and also $\sigma^k \to 0$, as $k \to \infty$. Moreover, assume that $x^k \to x^*$, where $\nabla c_i(x^*)$, $i = \overline{1, m}$, are linearly independent.

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

Derivatives of the penalty function

- Let $y(\sigma) := -c(x)/\sigma$: estimates of Lagrange multipliers.
- Let L be the Lagrangian function of (eCP),

$$L(x, y) := f(x) - y^T c(x).$$

- $\Phi_\sigma(x) = f(x) + \frac{1}{2\sigma}\|c(x)\|^2$. Then

$$\nabla\Phi_\sigma(x) = \nabla f(x) + \frac{1}{\sigma} J(x)^T c(x) = \nabla_x L(x, y(\sigma)),$$

where $J(x)$ Jacobian $m \times n$ matrix of constraints $c(x)$.

$$\begin{aligned} \nabla^2\Phi_\sigma(x) &= \nabla^2 f(x) + \frac{1}{\sigma}\sum_{i=1}^m c_i(x)\nabla^2 c_i(x) + \frac{1}{\sigma} J(x)^T J(x) \\ &= \nabla^2_{xx} L(x, y(\sigma)) + \frac{1}{\sigma} J(x)^T J(x). \end{aligned}$$

- $\sigma \longrightarrow 0$: generally, $c_i(x) \to 0$ or $\nabla^2 c_i(x) \to 0$ at the same rate with σ for all i. Thus usually, $\nabla^2_{xx} L(x, y(\sigma))$ well-behaved.
- $\sigma \to 0$: $J(x)^T J(x)/\sigma \to J(x^*)^T J(x^*)/0 = \infty$.

Ill-conditioning of the penalty's Hessian

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

m eigenvalues of $\nabla^2 \Phi_{\sigma^k}(x^k)$ are $\mathcal{O}(1/\sigma^k)$ and hence, tend to infinity as $k \to \infty$ (ie, $\sigma^k \to 0$); remaining $n - m$ are $\mathcal{O}(1)$.

• Hence, the condition number (ie, largest/smallest eigenvalue) of $\nabla^2 \Phi_{\sigma^k}(x^k)$ is $\mathcal{O}(1/\sigma^k)$

$\Longrightarrow$ it blows up as $k \to \infty$.

$\Longrightarrow$ worried that we may not be able to compute changes to x^k accurately. Namely, whether using linesearch or trust-region methods, asymptotically, we want to minimize $\Phi_{\sigma^{k+1}}(x)$ by taking Newton steps, i.e., solve the system

$$\nabla^2 \Phi_\sigma(x) dx = \nabla \Phi_\sigma(x), \qquad (*)$$

for dx from some current $x = x^{k,i}$ and $\sigma = \sigma^{k+1}$.

Despite ill-conditioning present, we can still solve for dx accurately!

Solving accurately for the Newton direction

Due to computed formulas for derivatives, (*) is equivalent to
$\left(\nabla^2_{xx} L(x, y(\sigma)) + \frac{1}{\sigma} J(x)^T J(x)\right) dx = -\left(\nabla f(x) + \frac{1}{\sigma} J(x)^T c(x)\right)$,
where $y(\sigma) = -c(x)/\sigma$. Define auxiliary variable w

$$w = \frac{1}{\sigma}\left(J(x)dx + c(x)\right).$$

Then the Newton system (*) can be re-written as

$$\begin{pmatrix} \nabla^2 L(x, y(\sigma)) & J(x)^\top \\ J(x) & -\sigma I \end{pmatrix} \begin{pmatrix} dx \\ w \end{pmatrix} = - \begin{pmatrix} \nabla f(x) \\ c(x) \end{pmatrix}$$

This system is essentially independent of σ for small $\sigma \Longrightarrow$ cannot suffer from ill-conditioning due to $\sigma \to 0$.

- Still need to be careful about minimizing Φ_σ for small σ. Eg, when using TR methods, use $\|dx\|_B \leq \Delta$ for TR constraint. B takes into account ill-conditioned terms of Hessian so as to encourage equal model decrease in all directions.

Perturbed optimality conditions

$$\min_{x \in \mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad c(x) = 0. \tag{eCP}$$

(eCP) satisfies the KKT conditions

$$(\text{dual feasibility})\ \nabla f(x) = J(x)^T y \ \text{and}\ (\text{primal feasibility})\ c(x) = 0.$$

Consider the perturbed problem

$$\begin{cases} \nabla f(x) - J(x)^T y = 0 \\ c(x) + \sigma y = 0 \end{cases} \tag{eCP$_p$}$$

Find roots of nonlinear system (eCP$_p$) as $\sigma \longrightarrow 0$ ($\sigma > 0$); use Newton's method for root finding.

Perturbed optimality conditions...

Newton's method for system (eCP$_p$) computes change (dx, dy) to (x, y) from

$$\begin{pmatrix} \nabla^2 \mathcal{L}(x,y) & -J(x)^\top \\ J(x) & \sigma I \end{pmatrix} \begin{pmatrix} dx \\ dy \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - J(x)^\top y \\ c(x) + \sigma y \end{pmatrix}$$

Eliminating dy, gives

$$\left(\nabla^2_{xx} L(x,y) + \frac{1}{\sigma} J(x)^T J(x) \right) dx = - \left(\nabla f(x) + \frac{1}{\sigma} J(x)^T c(x) \right)$$

$\Longrightarrow$ 'same' as Newton for quadratic penalty ! what's different?

Perturbed optimality conditions...

Primal:

$$\left(\nabla^2_{xx} L(x, y(\sigma)) + \frac{1}{\sigma} J(x)^T J(x)\right) dx^p = -\left(\nabla f(x) + \frac{1}{\sigma} J(x)^T c(x)\right)$$

where $y(\sigma) = -c(x)/\sigma$.

Primal-dual:

$$\left(\nabla^2_{xx} L(x, y) + \frac{1}{\sigma} J(x)^T J(x)\right) dx^{pd} = -\left(\nabla f(x) + \frac{1}{\sigma} J(x)^T c(x)\right)$$

The difference is in freedom to choose y in $\nabla^2 L(x, y)$ in primal-dual methods - it makes a big difference computationally.

Other penalty functions

Consider the general (CP) problem

$$\text{minimize}_{x\in\mathbb{R}^n} \quad f(x) \quad \text{subject to} \quad c_E(x)=0, \quad c_I(x)\geq 0. \qquad \text{(CP)}$$

Exact penalty function: $\Phi(x,\sigma)$ is exact if there is $\sigma_*>0$ such that if $\sigma<\sigma_*$, any local solution of (CP) is a local minimizer of $\Phi(x,\sigma)$. (Quadratic penalty is inexact.)

Examples:

- l_2-penalty function: $\Phi(x,\sigma)=f(x)+\frac{1}{\sigma}\|c_E(x)\|$
- l_1-penalty function: let $z^-=\min\{z,0\}$,
  $\Phi(x,\sigma)=f(x)+\frac{1}{\sigma}\sum_{i\in E}|c_i(x)|+\frac{1}{\sigma}\sum_{i\in I}[c_i(x)]^-$.

Extension of quadratic penalty to (CP):

$\Phi(x,\sigma)=f(x)+\frac{1}{2\sigma}\|c_E(x)\|^2+\frac{1}{2\sigma}\sum_{i\in I}\left([c_i(x)]^-\right)^2$

(may no longer be suff. smooth; it is inexact)

Interior point methods

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

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

- (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

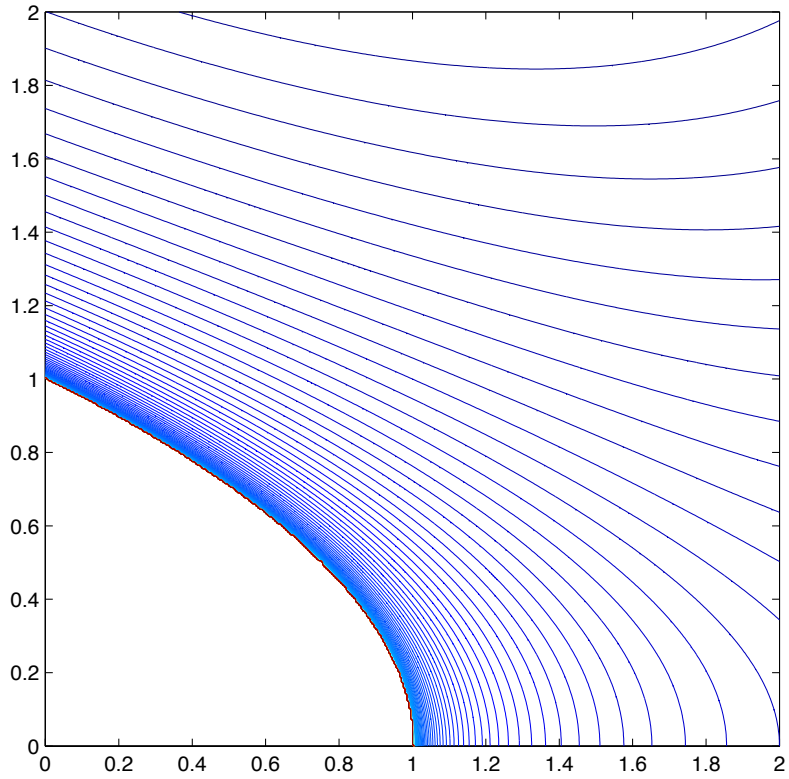
$$\min_{x \in \mathbb{R}^n} f_\mu(x) = f(x) - \mu \sum_{i=1}^{p} \log c_i(x) \text{ subject to } c(x) > 0. \quad \text{(iCP}_\mu\text{)}$$

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!

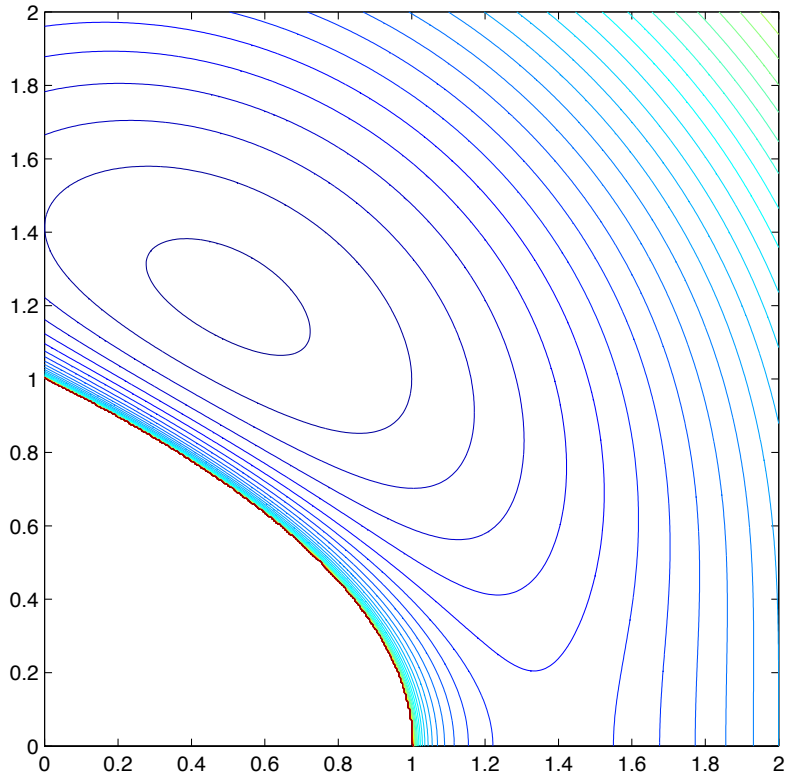
When μ "small", $\mu \to 0$: the term $f(x)$ "dominates" the log barrier terms in the objective of (iCP$_\mu$) $\implies x(\mu)$ "close" to the optimal boundary of Ω. [This also causes ill-conditioning ...]

• 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



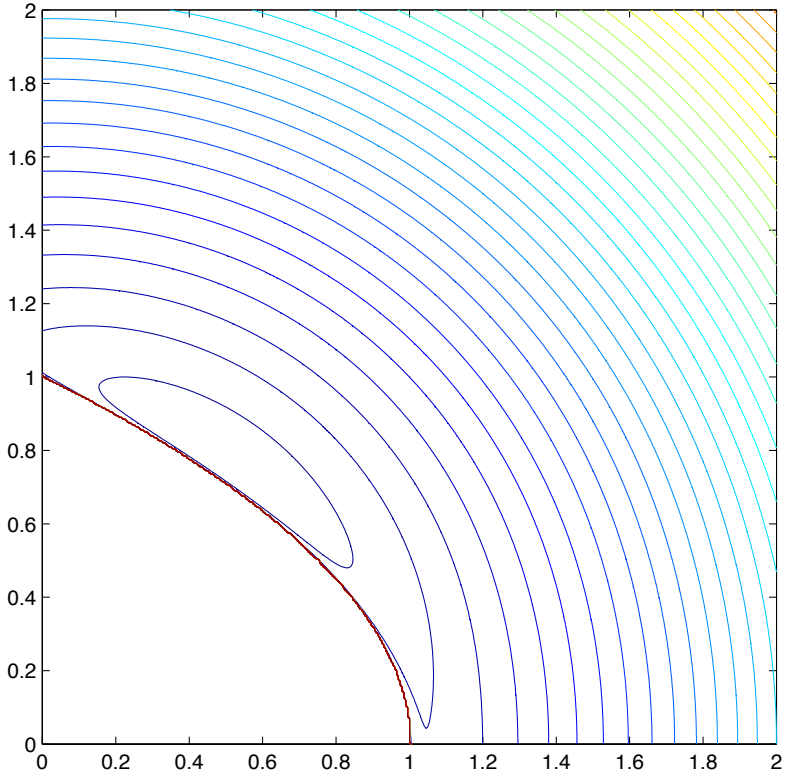
$\mu = 10$



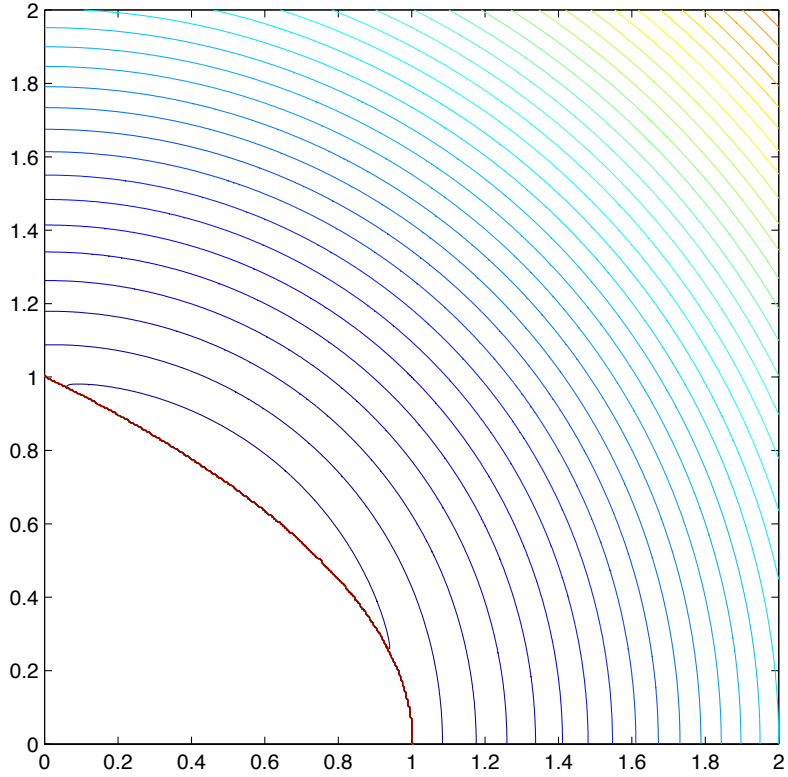
$\mu = 1$

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$_\mu$)

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

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

where $J(x)$ Jacobian of $c(x)$, $C(x) :=$ diag$(c(x))$, $e = (1, .., 1)$.

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) $\Longrightarrow$
$\nabla f(x^*) = \sum_{i=1}^p s_i^* \nabla c_i(x^*)$, $s^* \geq 0$, $s_i^* c_i(x^*) = 0$, $i = \overline{1, p}$.

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

Moreover ...

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

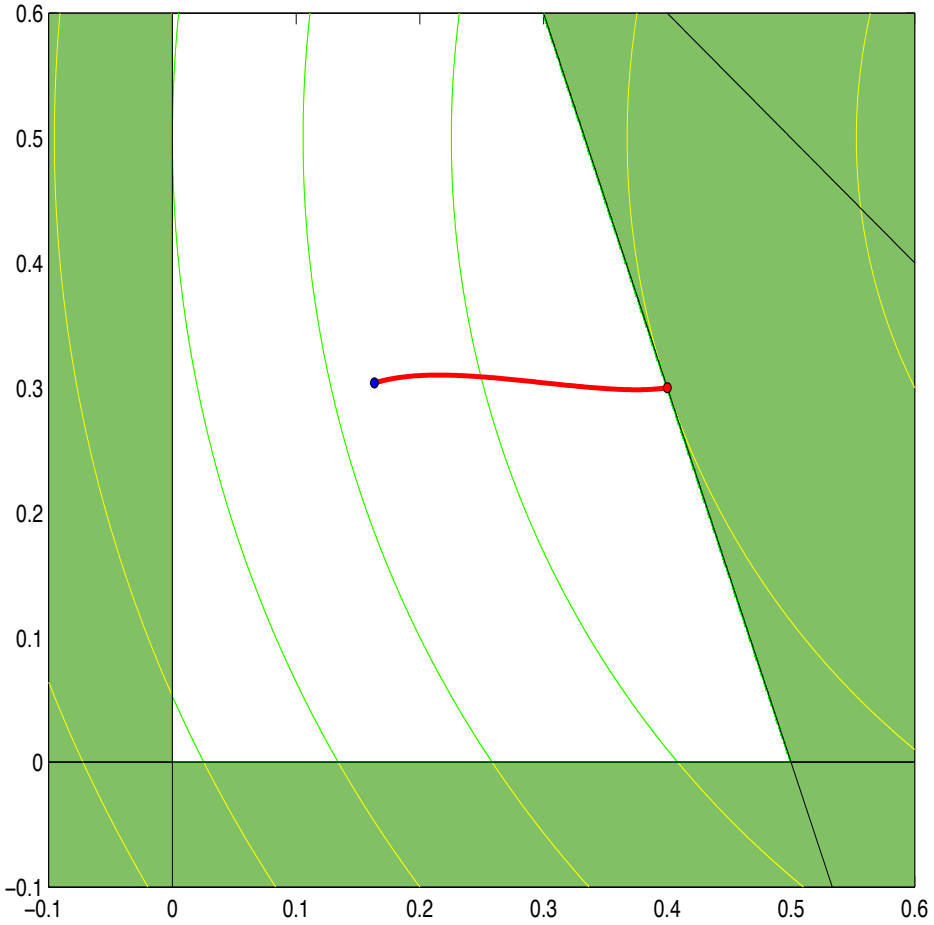
(a) $s_i^* > 0$ if $c_i(x^*) = 0$.

(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 $d^\top \nabla^2_{xx} \mathcal{L}(x^*, s^*) d \geq \alpha \|d\|^2$, where d such that $J(x^*)_{\mathcal{A}} d = 0$, 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



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

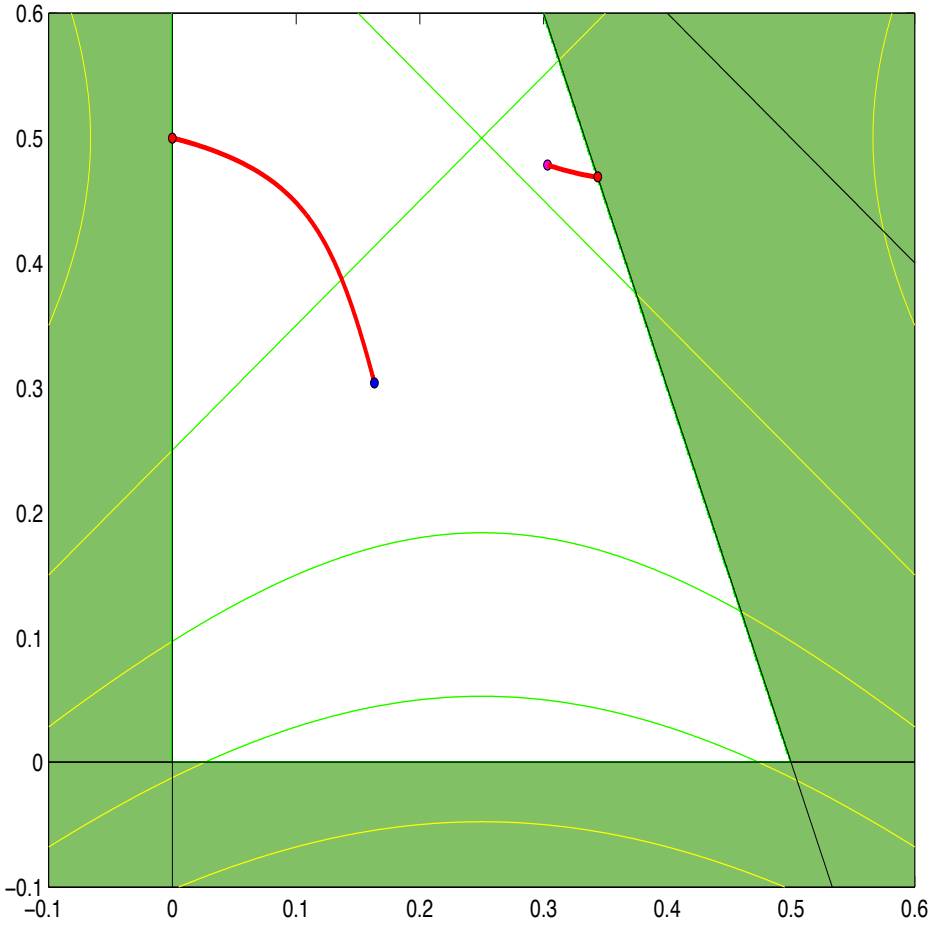
subject to $x_1 + x_2 \leq 1$

$$3x_1 + x_2 \leq 1.5$$

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

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

Central path trajectory - nonconvex case



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

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

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 $c(x^k + d^k)$ is not positive.

A convergence result for the barrier algorithm

Theorem. (Global convergence of barrier algorithm)
Apply the basic barrier algorithm to the (iCP). Assume that $f, c \in \mathcal{C}^2$, $s_i^k = \mu^k / c_i(x^k)$, $i = \overline{1, p}$, 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 $s^k \to s^*$, where s^* is the vector of Lagrange multipliers of (iCP). □

Minimizing the barrier function f_μ

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

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

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

where $J(x)$ is the Jacobian of $c(x)$, and $C(x) := \text{diag}(c(x))$.

$$\nabla^2 f_\mu(x) = \nabla^2 f(x) - \sum_{i=1}^{p} \frac{\mu}{c_i(x)} \nabla^2 c_i(x) + \sum_{i=1}^{p} \frac{\mu}{c_i(x)^2} \nabla c_i(x) \nabla c_i(x)^\top$$

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

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

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

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

Minimizing the barrier function f_μ ...

$\Longrightarrow \nabla f_\mu(x) = \nabla f(x) - J(x)s(x)$

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

Recall: the Lagragian function of (iCP)

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

$\Longrightarrow \nabla^2 f_\mu(x) = \nabla^2 \mathcal{L}(x, s(x)) + \mu J(x)^\top C^{-1}(x) S(x) J(x),$

or $\nabla^2 f_\mu(x) = \nabla^2 \mathcal{L}(x, s(x)) + \frac{1}{\mu} J(x)^\top S^2(x) J(x),$

where $S(x) := \text{diag}(s(x)) = \text{diag}(\mu / c_i(x) : i = \overline{1, n})$.

Potential difficulties

I. Ill-conditioning of the Hessian of f_μ

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

- some eigenvalues of $\nabla^2 f_{\mu^k}(x^k)$ tend to infinity as $k \to \infty$, while the rest stay bounded.
- 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 $x^k + d^k$ will be asymptotically infeasible (i. e., $c(x^k + d^k) < 0$) 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 $f_\mu \Longrightarrow \nabla f_\mu(x(\mu)) = 0 \Longleftrightarrow$
$\nabla f(x(\mu)) = \mu J(x(\mu))^\top C^{-1}(x(\mu))e$. Let $s(\mu) := \mu C^{-1}(x(\mu))$.

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

$$\begin{cases} \nabla f(x) - J(x)^\top s = 0, \\ C(x)Se = \mu e, \qquad\qquad \text{(OPT}_\mu\text{)} \end{cases}$$
$$c(x) > 0, \quad s > 0.$$

Compare with the KKT system for (iCP):

$$\begin{cases} \nabla f(x) - J(x)^\top s = 0, \\ C(x)Se = 0, \qquad\qquad \text{(KKT)} \end{cases}$$
$$c(x) \geq 0, \quad s \geq 0.$$

Primal-dual path-following methods (1990s)

Satisfy $c(x) > 0$ and $s > 0$, and use Newton's method to solve the system

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

i. e., the Newton direction (dx, ds) satisfies

$$\begin{pmatrix} \nabla^2 \mathcal{L}(x,s) & -J(x)^\top \\ SJ(x) & C(x) \end{pmatrix} \begin{pmatrix} dx \\ ds \end{pmatrix} = - \begin{pmatrix} \nabla f(x) - J(x)^\top s \\ C(x)s - \mu e \end{pmatrix}.$$

Eliminating ds, we deduce

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

Primal-dual versus primal methods

Primal-dual:

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

Primal:

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

where $s(x) := \mu C^{-1}(x) e$.

$\Longrightarrow$ 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.

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.

What we have not covered

- methods for smooth constrained optimization: augmented Lagrangian, SQP, active-set, filter
- special structure smooth optimization: linear programming, etc. (see discrete course, MT ?)
- derivative-free optimization methods (see HT course)
- global optimization
- integer (linear and nonlinear) programming (see discrete course, MT ?)
- stochastic programming