Lecture 11 and 12: Penalty methods and augmented Lagrangian methods for nonlinear programming

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

Penalty methods for nonlinear programming

Nonlinear equality-constrained problems

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

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

A penalty function for (eCP)

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

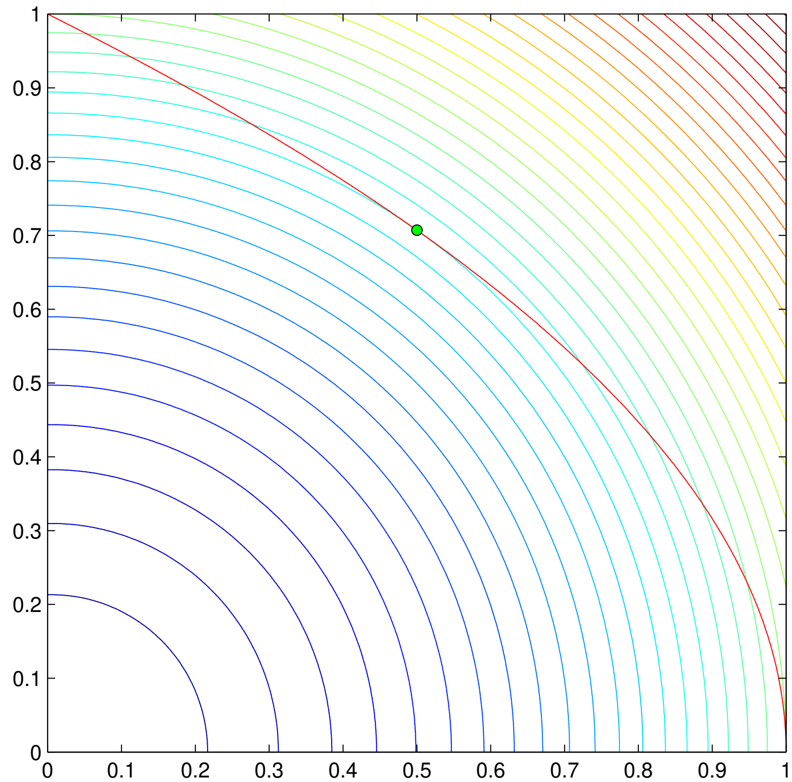
The quadratic penalty function:

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

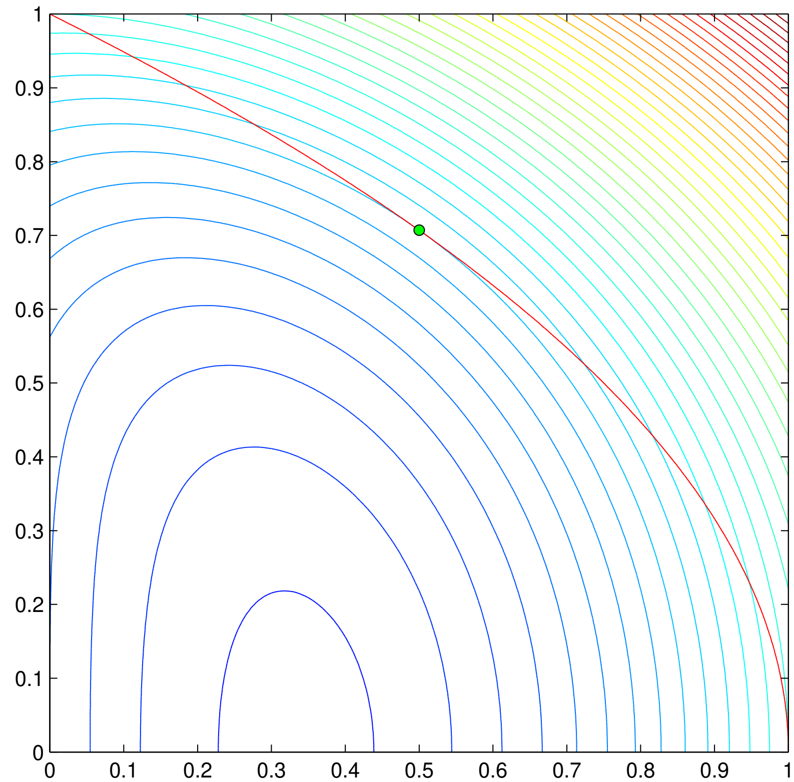
where $\sigma > 0$ penalty parameter.

- 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



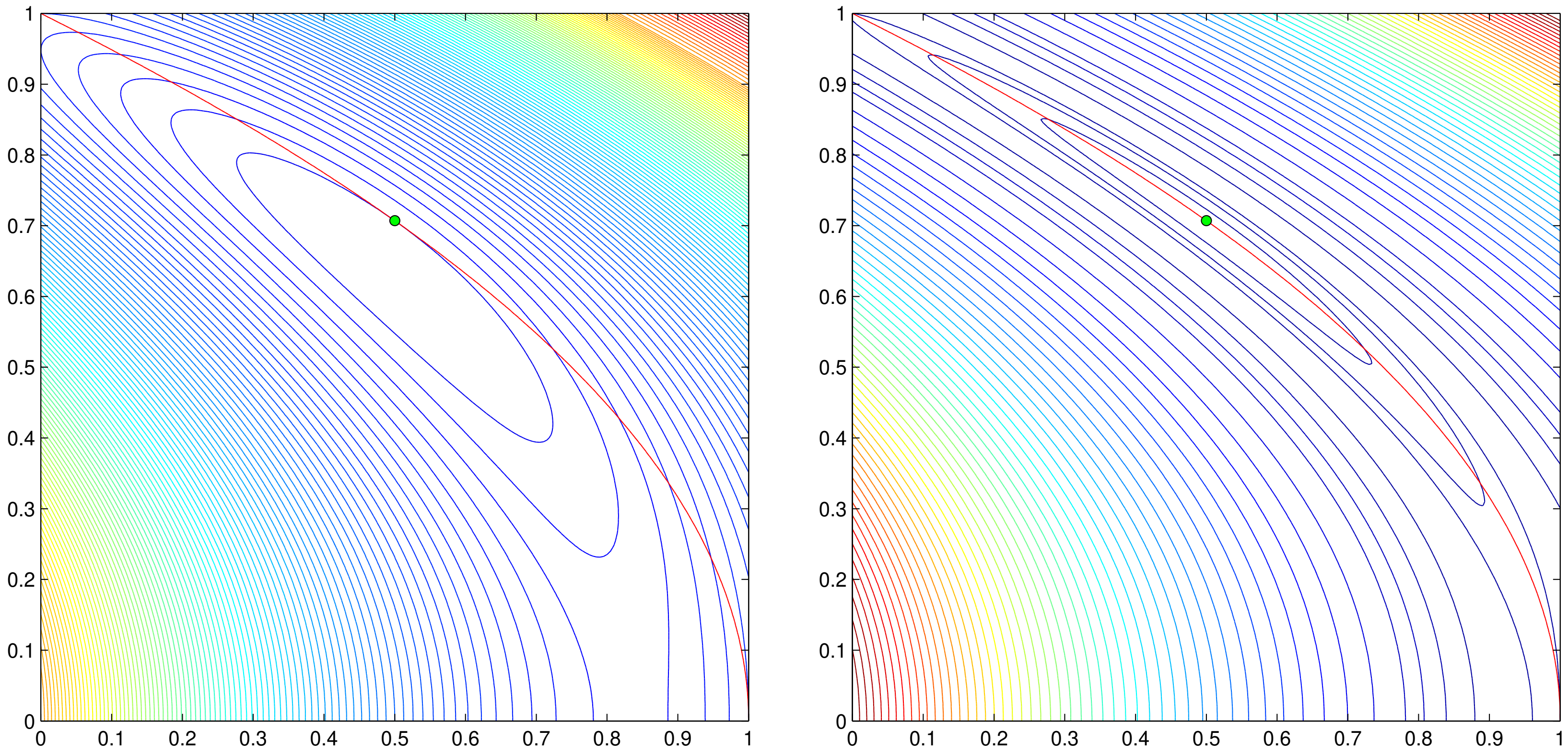
$\sigma = 100$



$\sigma = 1$

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 21. (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. $\square$

- $\nabla c_i(x^*)$, $i = \overline{1,m}$, lin. indep. $\Leftrightarrow$ the Jacobian matrix $J(x^*)$ of the constraints is full row rank and so $m \leq n$.
- $J(x^*)$ not full rank, then x^* (locally) minimizes the infeasibility $\|c(x)\|$. [let $y^k \to \infty$ in ($\Diamond$) on the next slide]

A convergence result for the penalty method

Proof of Theorem 21. $J(x^*)$ full rank $\Longrightarrow$ $\exists J(x^*)^+ = (J(x^*)J(x^*)^T)^{-1}J(x^*)$ pseudo-inverse. As $x^k \to x^*$ and J cont. $\Rightarrow \exists J(x^k)^+$ bounded above and cont. for all suff. large k. Let $y^k = -c(x^k)/\sigma^k$ and $y^* = J(x^*)^+\nabla f(x^*)$.

$$\|\nabla\Phi_{\sigma^k}(x^k)\| = \|\nabla f(x^k) - J(x^k)^T y^k\| \le \epsilon_k \quad (\Diamond)$$

$$\|J(x^k)^+\nabla f(x^k) - y^k\| = \|J(x^k)^+(\nabla f(x^k) - J(x^k)^T y^k)\| \le$$
$$\|J(x^k)^+\| \cdot \|\nabla f(x^k) - J(x^k)^T y^k\| \le$$
$$\left\{\|J(x^k)^+ - J(x^*)^+\| + \|J(x^*)^+\|\right\}\epsilon_k \le 2\|J(x^*)^+\|\epsilon_k \quad (\bullet)$$

where in the last $\le$ we used $x^k \to x^*$ and J^+ continuous.

Triangle inequality (add and subtr $J^+\nabla f$) and def of y^* give

$$\|y^k - y^*\| \le \|J(x^k)^+\nabla f(x^k) - J(x^*)^+\nabla f(x^*)\| + \|J(x^k)^+\nabla f(x^k) - y^k\|$$

Thus $y^k \to y^*$ since $x^k \to x^*$, J^+ and ∇f cont., $(\bullet)$ and $\epsilon_k \to 0$. Using all these again in $(\Diamond)$ as $k \to \infty$: $\nabla f(x^*) - J(x^*)^T y^* = 0$. As $c(x^k) = -\sigma^k y^k$, $\sigma^k \to 0$, $y^k \to y^* \Rightarrow c(x^*) = 0$. Thus x^* KKT.

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, $c_i(x) \to 0$ 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 ...

'Fact' [cf. Th 5.2, Gould ref.] $\Longrightarrow$ 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 $\mathcal{O}(1)$ in the limit.

- Hence, the condition number 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

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

(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

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

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)

Augmented Lagrangian methods for nonlinear programming

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.

Another example of merit function and method for (eCP):

augmented Lagrangian function

$$\Phi(x, u, \sigma) = f(x) - u^T c(x) + \frac{1}{2\sigma} \|c(x)\|^2$$

where $u \in \mathbb{R}^m$ and $\sigma > 0$ are auxiliary parameters.

Two interpretations:

- shifted quadratic penalty function
- convexification of the Lagrangian function

Aim: adjust u and σ to encourage convergence.

Derivatives of the augmented Lagrangian function

Let $J(x)$ Jacobian of constraints $c(x) = (c_1(x), \ldots, c_m(x))$.

- $\nabla_x \Phi(x, u, \sigma) = \nabla f(x) - J(x)^T u + \frac{1}{\sigma} J(x)^T c(x)$

  $\Longrightarrow$

  $\nabla_x \Phi(x, u, \sigma) = \nabla f(x) - J(x)^T y(x) = \nabla_x \mathcal{L}(x, y(x))$

  where $y(x) = u - \dfrac{c(x)}{\sigma}$ Lagrange multiplier estimates

- $\nabla^2 \Phi(x, u, \sigma) = \nabla^2 f(x) - \sum_{i=1}^m u_i \nabla^2 c_i(x) +$
  $\frac{1}{\sigma} \sum_{i=1}^m c_i(x) \nabla^2 c_i(x) + \frac{1}{\sigma} J(x)^T J(x)$

  $\Longrightarrow$

  $\nabla^2 \Phi(x, u, \sigma) = \nabla^2 f(x) - \sum_{i=1}^m y_i \nabla^2 c_i(x) + \frac{1}{\sigma} J(x)^T J(x)$

  $\Longrightarrow \nabla^2 \Phi(x, u, \sigma) = \nabla^2 \mathcal{L}(x, y(x)) + \frac{1}{\sigma} J(x)^T J(x)$

- Lagrangian: $\mathcal{L}(x, y) = f(x) - y^T c(x)$

A convergence result for the augmented Lagrangian

Theorem 22. (Global convergence of augmented Lagrangian) Assume that $f, c \in \mathcal{C}^1$ in (eCP) and let

$$y^k = u^k - \frac{c(x^k)}{\sigma^k},$$

for given $u^k \in \mathbb{R}^m$, and assume that

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

Moreover, assume that $x^k \to x^*$, where $\nabla c_i(x^*)$, $i = \overline{1, m}$, are linearly independent. Then $y^k \longrightarrow y^*$ as $k \longrightarrow \infty$ with y^* satisfying $\nabla f(x^*) - J(x^*)^T y^* = 0$.

If additionally, either $\sigma^k \to 0$ for bounded u^k or $u^k \to y^*$ for bounded σ^k then x^* is a KKT point of (eCP) with associated Lagrange multipliers y^*. □

A convergence result for the augmented Lagrangian

Proof of Theorem 22. The first part of Th 22, namely, convergence of y^k to $y^* = J(x^*)^+ \nabla f(x^*)$ follows exactly as in the proof of Theorem 21 (penalty method convergence). (Note that the assumption $\sigma^k \to 0$ is not needed for this part of the proof of Th 21.)

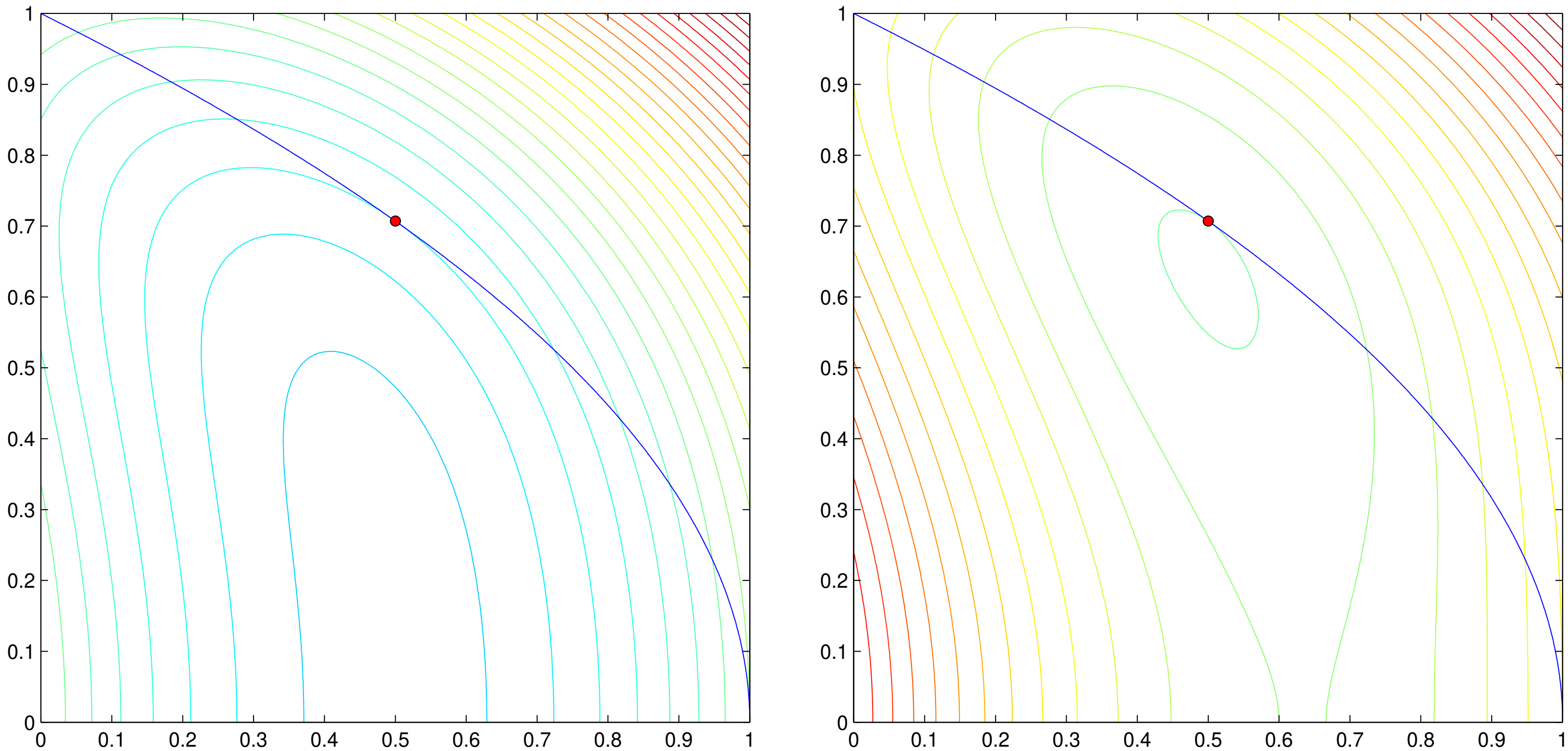
It remains to show that under the additional assumptions on u^k and σ^k, x^* is feasible for the constraints. To see this, use the definition of y^k to deduce $c(x^k) = \sigma^k(u^k - y^k)$ and so

$$\|c(x^k)\| = \sigma^k \|u^k - y^k\| \leq \sigma^k \|y^k - y^*\| + \sigma^k \|u^k - y^*\|$$

Thus $c(x^k) \longrightarrow 0$ as $k \to \infty$ due to $y^k \to y^*$ (cf. first part of theorem) and the additional assumptions on u^k and σ^k. As $x^k \to x^*$ and c is continuous, we deduce that $c(x^*) = 0$. $\square$

Note that Augmented Lagrangian may converge to KKT points without $\sigma^k \to 0$, which limits the ill-conditioning.

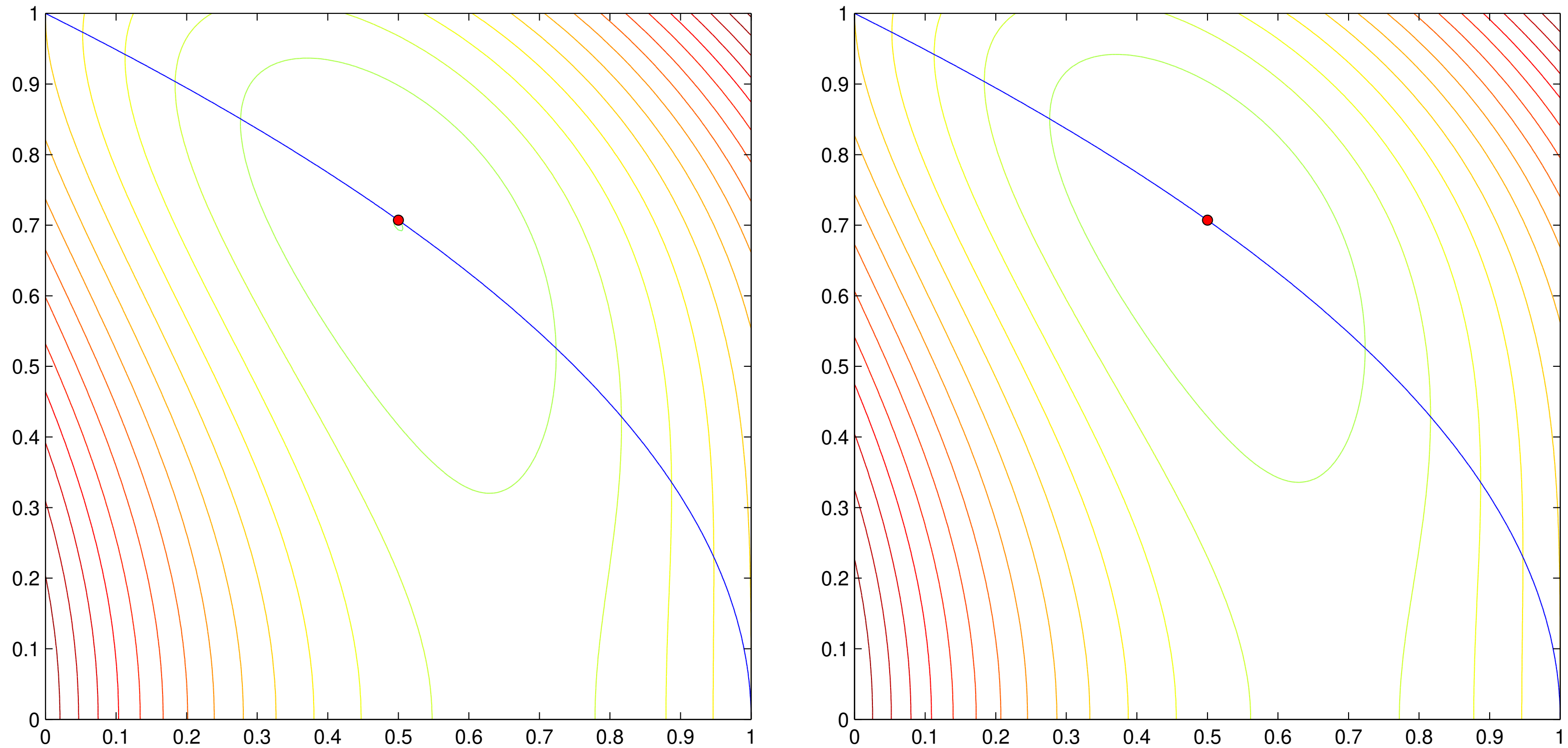
Contours of the augmented Lagrangian - an example



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

The augmented Lagrangian function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$ for fixed $\sigma = 1$

Contours of the augmented Lagrangian - an example...



$u = 0.99$ $\qquad$ $u = y^* = 1$

The augmented Lagrangian function for $\min x_1^2 + x_2^2$ subject to $x_1 + x_2^2 = 1$ for fixed $\sigma = 1$

Augmented Lagrangian methods

Th 22 $\Longrightarrow$ convergence guaranteed if u^k fixed and $\sigma^k \longrightarrow 0$

[similar to quadratic penalty methods]

$\Longrightarrow y^k \longrightarrow y^*$ and $c(x^k) \longrightarrow 0$

- check if $\|c(x^k)\| \leq \eta^k$ where $\eta^k \longrightarrow 0$
  - if so, set $u^{k+1} = y^k$ and $\sigma^{k+1} = \sigma^k$

    [recall expression of y^k in Th 22]
  - if not, set $u^{k+1} = u^k$ and $\sigma^{k+1} \leq \tau\sigma^k$ for some $\tau \in (0, 1)$
- reasonable: $\eta^k = (\sigma^k)^{0.1+0.9j}$ where j iterations since σ^k last changed

Under such rules, can ensure that σ^k is eventually unchanged under modest assumptions, and (fast) linear convergence.

Need also to ensure that σ^k is sufficiently large that the Hessian $\nabla^2\Phi(x^k, u^k, \sigma^k)$ is positive (semi-)definite.

A basic augmented Lagrangian method

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

- Set η^k and ϵ^{k+1}.
  If $\|c(x^k)\| \le \eta^k$, set $u^{k+1} = y^k$ and $\sigma^{k+1} = \sigma^k$.
  Otherwise, set $u^{k+1} = u^k$ and $\sigma^{k+1} \le \tau\sigma^k$.
- 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 $\Phi(\cdot, u^{k+1}, \sigma^{k+1})$ for which $\|\nabla_x \Phi(x^{k+1}, u^{k+1}, \sigma^{k+1})\| \le \epsilon^{k+1}$.
  Let $k := k + 1$. $\Diamond$

- Often choose $\tau = \min(0.1, \sqrt{\sigma^k})$
- Reasonable: $\epsilon^k = (\sigma^k)^{j+1}$, where j iterations since σ^k last changed