Part 5: Penalty and augmented Lagrangian methods for equality constrained optimization

Nick Gould (RAL)

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

Part C course on continuoue optimization

CONSTRAINED MINIMIZATION

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

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

- $\odot$ assume that $f,\ c \in C^1$ (sometimes C^2) and Lipschitz
- $\odot$ 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

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

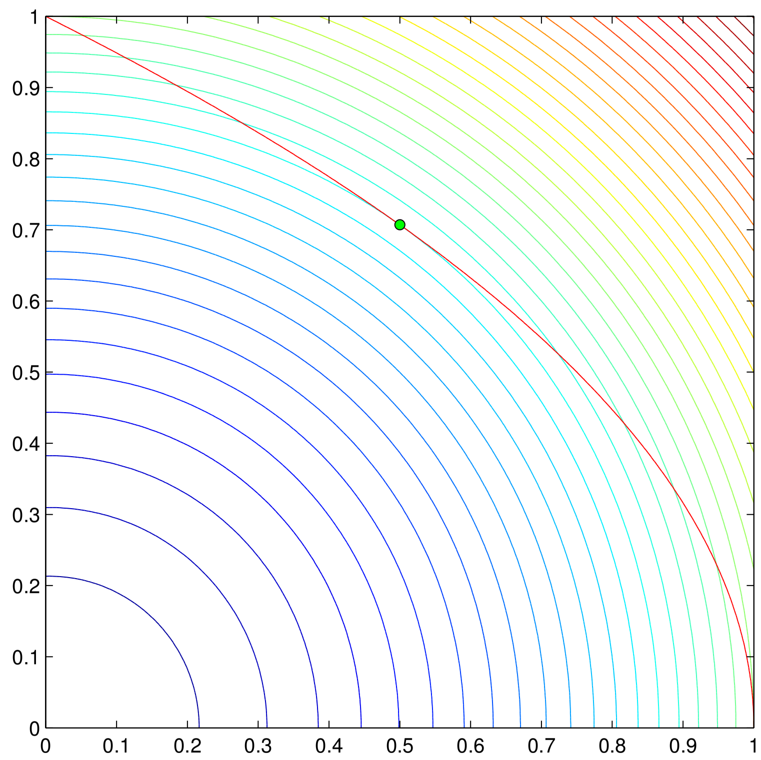
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \;\; f(x) \;\; \text{subject to} \;\; c(x) = 0$$

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

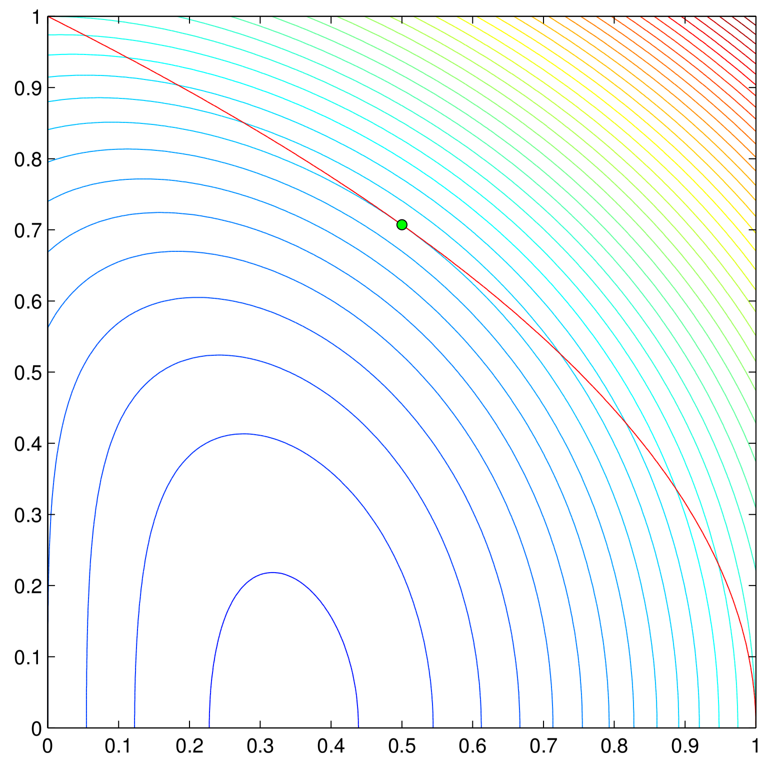
$$\Phi(x, \mu) = f(x) + \frac{1}{2\mu} \|c(x)\|_2^2$$

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

CONTOURS OF THE PENALTY FUNCTION



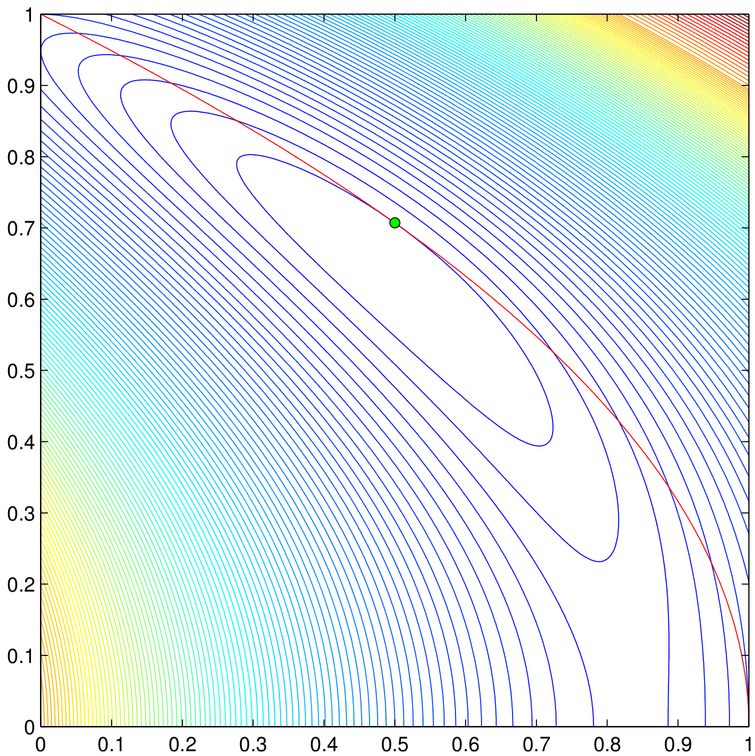
$\mu = 100$



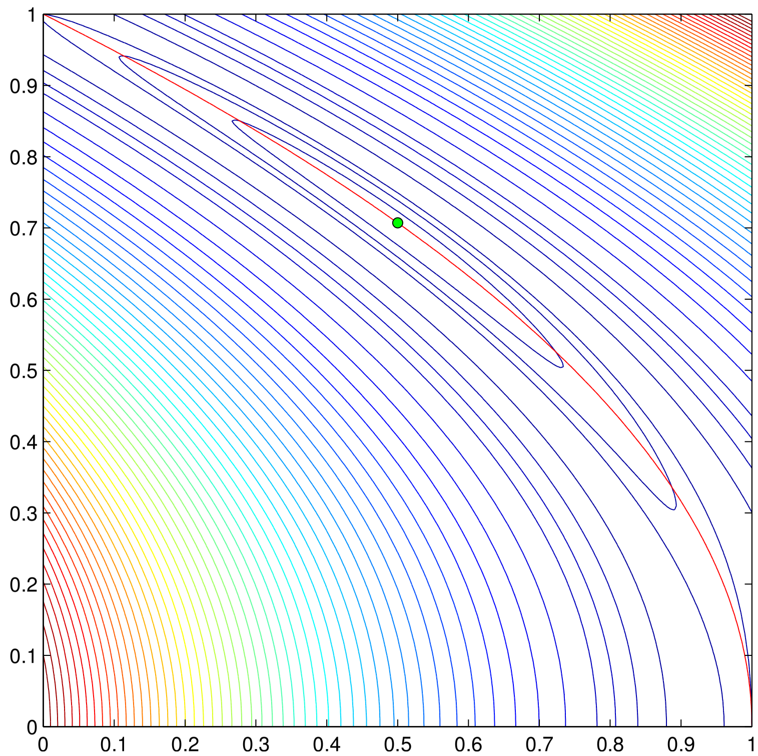
$\mu = 1$

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

$$y_k \stackrel{\text{def}}{=} -\frac{c(x_k)}{\mu_k},$$

that

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

where ϵ_k converges to zero as $k \rightarrow \infty$, 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 $\{y_k\}$ converge to the associated Lagrange multipliers y_*.

PROOF OF THEOREM 5.1

Generalized inv. $A^+(x) \stackrel{\text{def}}{=} \left(A(x)A^T(x)\right)^{-1} A(x)$ bounded near x_*.

Define

$$y_k \stackrel{\text{def}}{=} -\frac{c(x_k)}{\mu_k} \quad \text{and} \quad y_* \stackrel{\text{def}}{=} A^+(x_*)g(x_*). \tag{1}$$

Inner-iteration termination rule

$$\|g(x_k) - A^T(x_k)y_k\| \le \epsilon_k \tag{2}$$

$$\begin{aligned}
\Longrightarrow \ \|A^+(x_k)g(x_k) - y_k\|_2 &= \left\|A^+(x_k)\left(g(x_k) - A^T(x_k)y_k\right)\right\|_2 \\
&\le 2\|A^+(x_*)\|_2\epsilon_k
\end{aligned}$$

$$\begin{aligned}
\Longrightarrow \ \|y_k - y_*\|_2 \le \ & \|A^+(x_*)g(x_*) - A^+(x_k)g(x_k)\|_2 + \\
& \|A^+(x_k)g(x_k) - y_k\|_2
\end{aligned}$$

$\Longrightarrow \ \{y_k\} \longrightarrow y_*$. Continuity of gradients + (2) $\Longrightarrow$

$$g(x_*) - A^T(x_*)y_* = 0.$$

(1) implies $c(x_k) = -\mu_k y_k$ + continuity of constraints $\Longrightarrow$ $c(x_*) = 0$.

$\Longrightarrow$ (x_*, y_*) satisfies the first-order optimality conditions.

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

- $\nabla_x \Phi(x, \mu) = g(x, y(x))$
- $\nabla_{xx} \Phi(x, \mu) = H(x, y(x)) + \dfrac{1}{\mu} A^T(x) A(x)$

where

- **Lagrange multiplier estimates**:

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

- $g(x, y(x)) = g(x) - A^T(x) y(x)$: **gradient of the Lagrangian**
- $H(x, y(x)) = H(x) - \sum_{i=1}^{m} y_i(x) H_i(x)$: **Lagrangian Hessian**

GENERIC QUADRATIC PENALTY NEWTON SYSTEM

Newton correction s from x for quadratic penalty function is

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

LIMITING DERIVATIVES OF Φ

For small μ: roughly

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

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

POTENTIAL DIFFICULTY

Ill-conditioning of the Hessian of the penalty function:

roughly speaking (non-degenerate case)

- $\odot$ m eigenvalues $\approx \lambda_i \left[A^T(x)A(x)\right]/\mu_k$
- $\odot$ $n-m$ 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 $\nabla_{xx}\Phi(x_k,\mu_k) = O(1/\mu_k)$

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

THE ILL-CONDITIONING IS BENIGN

Newton system:

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

Define auxiliary variables

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

$\Longrightarrow$

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

- ⊙ 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:

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

Consider the "perturbed" problem

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

where $\mu > 0$

PRIMAL-DUAL PATH-FOLLOWING METHODS

Track roots of

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

as $0 < \mu \rightarrow 0$

$\odot$ 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 $w \Longrightarrow$

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

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

PRIMAL VS. PRIMAL-DUAL

Primal:

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

Primal-dual:

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

where

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

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)**:

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

where u and μ are auxiliary **parameters**

Two interpretations —

- ⊙ shifted quadratic penalty function
- ⊙ convexification of the Lagrangian function

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

DERIVATIVES OF THE AUGMENTED LAGRANGIAN FUNCTION

- $\nabla_x \Phi(x, u, \mu) = g(x, y^{\mathrm{F}}(x))$
- $\nabla_{xx} \Phi(x, u, \mu) = H(x, y^{\mathrm{F}}(x)) + \dfrac{1}{\mu} A^T(x) A(x)$

where

- **First-order** Lagrange multiplier estimates:

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

- $g(x, y^{\mathrm{F}}(x)) = g(x) - A^T(x) y^{\mathrm{F}}(x)$: gradient of the Lagrangian
- $H(x, y^{\mathrm{F}}(x)) = H(x) - \sum_{i=1}^{m} y_i^{\mathrm{F}}(x) H_i(x)$: Lagrangian Hessian

AUGMENTED LAGRANGIAN CONVERGENCE

Theorem 5.2. Suppose that f, $c \in \mathcal{C}^2$, that

$$y_k \stackrel{\text{def}}{=} u_k - c(x_k)/\mu_k,$$

for given $\{u_k\}$, that

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

where ϵ_k converges to zero as $k \rightarrow \infty$, and that x_k converges to x_* for which $A(x_*)$ is full rank. Then $\{y_k\}$ converge to some y_* for which $g(x_*) = A^T(x_*)y_*$.

If additionally either μ_k converges to zero for bounded u_k or u_k converges to y_* for bounded μ_k, x_* and y_* satisfy the first-order necessary optimality conditions for the problem

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

PROOF OF THEOREM 5.2

Convergence of y_k to $y_* \stackrel{\text{def}}{=} A^+(x_*)g(x_*)$ for which $g(x_*) = A^T(x_*)y_*$ is exactly as for Theorem 5.1.

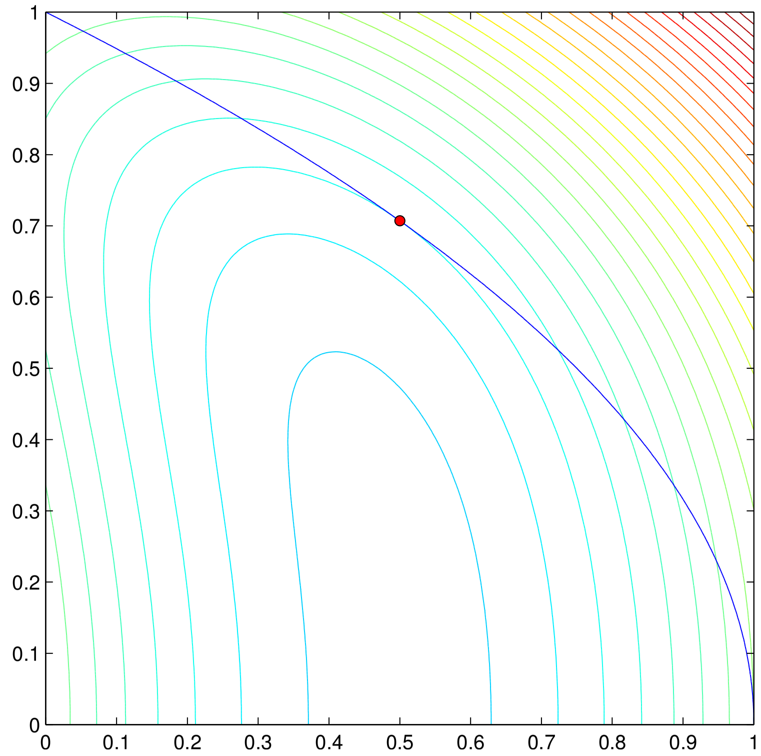
Definition of $y_k \Longrightarrow$

$$\|c(x_k)\| = \mu_k\|u_k - y_k\| \le \mu_k\|y_k - y_*\| + \mu_k\|u_k - y_*\|$$

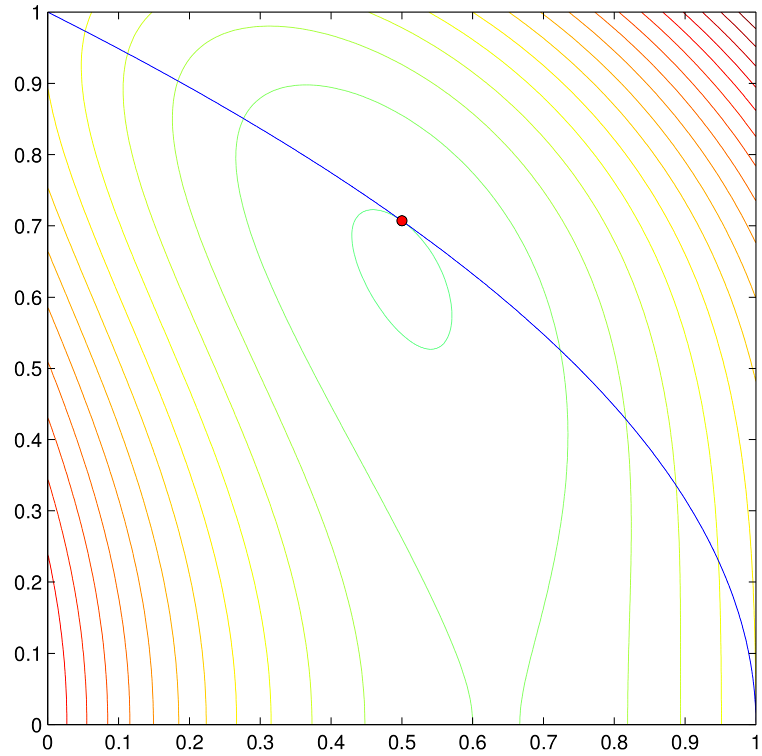
$\Longrightarrow$ $c(x_*) = 0$ from assumptions.

$\Longrightarrow$ (x_*, y_*) satisfies the first-order optimality conditions.

CONTOURS OF THE AUGMENTED LAGRANGIAN FUNCTION



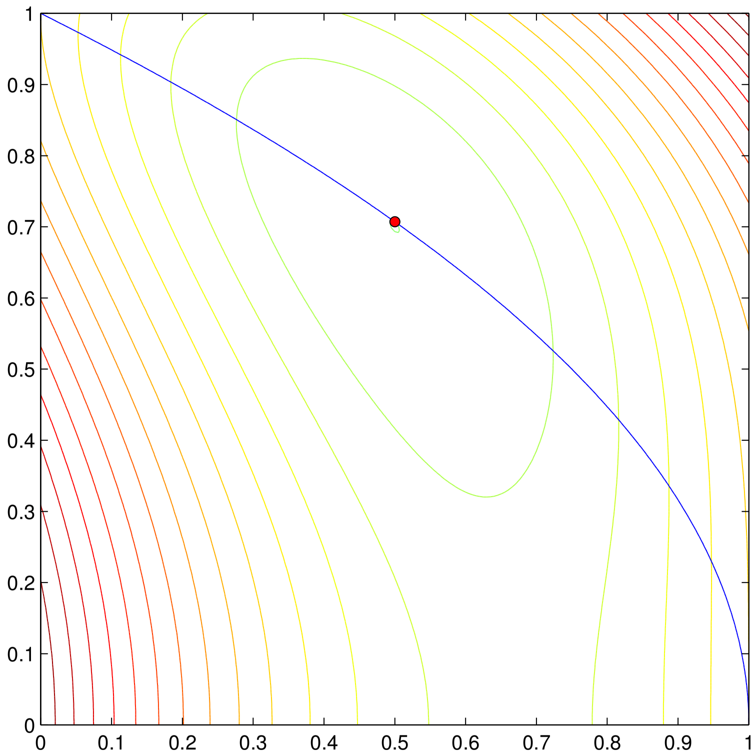
$u = 0.5$



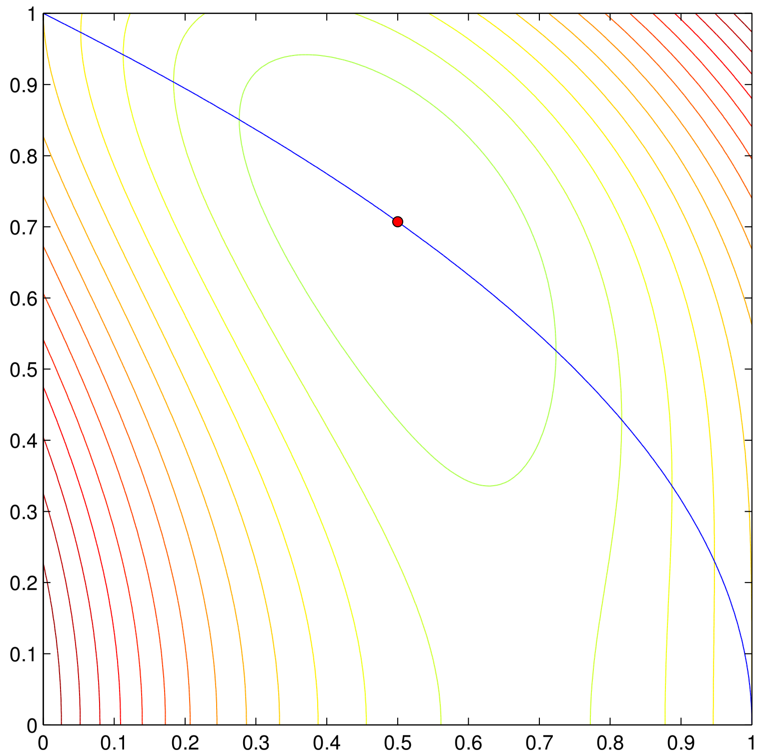
$u = 0.9$

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.)



$u = 0.99$



$u = y_* = 1$

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 u_k fixed and $\mu \longrightarrow 0$
  $\Longrightarrow y_k \longrightarrow y_*$ 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 $\mu_{k+1} = \mu_k$
  - if not, set $u_{k+1} = u_k$ and $\mu_{k+1} \leq \tau \mu_k$ 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 $\nabla_{xx}\Phi(x_k, u_k, \mu_k)$ is positive (semi-)definite

BASIC AUGMENTED LAGRANGIAN ALGORITHM

Given $\mu_0 > 0$ and u_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 $\Phi(x, u_k, \mu_k)$ for which $\|\nabla_x \Phi(x_k, u_k, \mu_k)\| \leq \epsilon_k$

  If $\|c(x_k)\| \leq \eta_k$, set $u_{k+1} = y_k$ and $\mu_{k+1} = \mu_k$

  Otherwise set $u_{k+1} = u_k$ and $\mu_{k+1} \leq \tau \mu_k$

  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