13. Constrained nonlinear least squares

- Lagrange multipliers
- constrained nonlinear least squares
- penalty method
- augmented Lagrangian method
- nonlinear control example

Constrained nonlinear least squares

Nonlinear least squares

$$\text{minimize} \quad f_1(x)^2 + \cdots + f_m(x)^2 = \|f(x)\|^2$$

- variable is n-vector x
- $f_i(x)$ is ith (scalar) *residual*
- $f : \mathbf{R}^n \to \mathbf{R}^m$ is the vector function $f(x) = (f_1(x), \ldots, f_m(x))$

Algorithms (see 133A)

- Newton's method
- Gauss–Newton method
- Levenberg–Marquardt method

This lecture: add p equality constraints

$$g_1(x) = 0, \qquad g_2(x) = 0, \qquad \ldots, \qquad g_p(x) = 0$$

Outline

- **Lagrange multipliers**
- constrained nonlinear least squares
- penalty method
- augmented Lagrangian method
- nonlinear control example

Notation

we use the same derivative notation as in 133A

- gradient and Hessian of a scalar function $h : \mathbf{R}^n \to \mathbf{R}$ are denoted by

$$\nabla h(\tilde{x}) = \begin{bmatrix} \frac{\partial h}{\partial x_1}(\tilde{x}) \\ \vdots \\ \frac{\partial h}{\partial x_n}(\tilde{x}) \end{bmatrix}, \qquad \nabla^2 h(\tilde{x}) = \begin{bmatrix} \frac{\partial^2 h}{\partial x_1^2}(\tilde{x}) & \cdots & \frac{\partial^2 h}{\partial x_1 \partial x_n}(\tilde{x}) \\ \vdots & & \vdots \\ \frac{\partial^2 h}{\partial x_n \partial x_1}(\tilde{x}) & \cdots & \frac{\partial^2 h}{\partial x_n^2}(\tilde{x}) \end{bmatrix}$$

- Jacobian of vector function $f : \mathbf{R}^n \to \mathbf{R}^m$ is denoted by

$$Df(\tilde{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\tilde{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\tilde{x}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\tilde{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\tilde{x}) \end{bmatrix} = \begin{bmatrix} \nabla f_1(\tilde{x})^T \\ \vdots \\ \nabla f_m(\tilde{x})^T \end{bmatrix}$$

Minimization with equality constraints

$$\begin{array}{ll} \text{minimize} & h(x) \\ \text{subject to} & g_1(x) = 0 \\ & \cdots \\ & g_p(x) = 0 \end{array}$$

$h, g_1, \ldots, g_p$ are functions from $\mathbf{R}^n$ to $\mathbf{R}$

- x is *feasible* if it satisfies the constraints:

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_p(x) \end{bmatrix} = 0$$

- feasible $\hat{x}$ is *optimal* (or a *minimum*) if $h(\hat{x}) \leq h(x)$ for all feasible x
- feasible $\hat{x}$ is *locally optimal* (a *local minimum*) if there exists an $R > 0$ such that

$$h(\hat{x}) \leq h(x) \quad \text{for all feasible } x \text{ with } \|x - \hat{x}\| \leq R$$

Lagrange multipliers

Lagrangian: the *Lagrangian* is the function

$$\begin{aligned} L(x, z) &= h(x) + z^T g(x) \\ &= h(x) + z_1 g_1(x) + \cdots + z_p g_p(x) \end{aligned}$$

the p-vector $z = (z_1, \ldots, z_p)$ is the vector of *Lagrange multipliers* $z_1, \ldots, z_p$

Gradient of Lagrangian

$$\nabla L(\tilde{x}, \tilde{z}) = \begin{bmatrix} \nabla_x L(\tilde{x}, \tilde{z}) \\ \nabla_z L(\tilde{x}, \tilde{z}) \end{bmatrix}$$

where

$$\begin{aligned} \nabla_x L(\tilde{x}, \tilde{z}) &= \nabla h(\tilde{x}) + \tilde{z}_1 \nabla g_1(\tilde{x}) + \cdots + \tilde{z}_p \nabla g_p(\tilde{x}) \\ &= \nabla h(\tilde{x}) + Dg(\tilde{x})^T \tilde{z} \\ \nabla_z L(\tilde{x}, \tilde{z}) &= g(\tilde{x}) \end{aligned}$$

First-order optimality conditions

$$
\begin{array}{ll}
\text{minimize} & h(x) \\
\text{subject to} & g(x) = 0
\end{array}
$$

h is a function from $\mathbf{R}^n$ to $\mathbf{R}$, g is a function from $\mathbf{R}^n$ to $\mathbf{R}^p$

First-order necessary optimality conditions

if $\hat{x}$ is locally optimal and $\operatorname{rank}(Dg(\hat{x})) = p$, then there exist multipliers $\hat{z}$ with

$$\nabla L_x(\hat{x}, \hat{z}) = \nabla h(\hat{x}) + Dg(\hat{x})^T \hat{z} = 0$$

- together with $g(\hat{x}) = 0$, this forms a set of $n + p$ equations in $n + p$ variables $\hat{x}$, $\hat{z}$
- gradient $\nabla h(\hat{x})$ is a linear combination of gradients $\nabla g_1(\hat{x}), \ldots, \nabla g_p(\hat{x})$

Regular feasible point

- a feasible x with if $\operatorname{rank}(Dg(x)) = p$ is called a *regular* feasible point
- at a regular feasible point, $\nabla g_1(x), \ldots, \nabla g_p(x)$ are linearly independent

Example

suppose A is a symmetric $n \times n$ matrix

$$\begin{array}{ll} \text{minimize} & x^T A x \\ \text{subject to} & x^T x = 1 \end{array}$$

- Lagrangian is

$$L(x, z) = x^T A x + z(x^T x - 1)$$

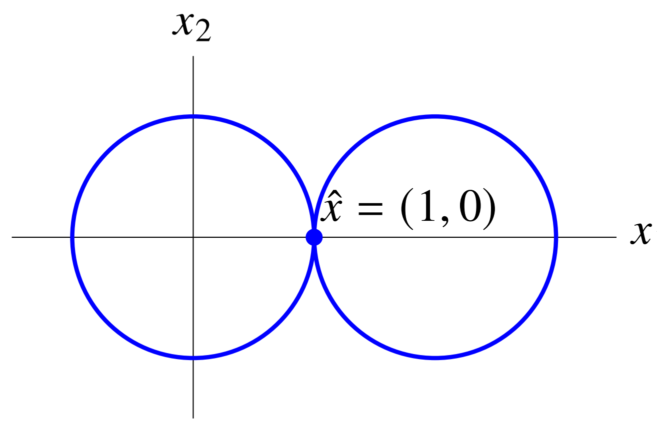
- first-order necessary optimality condition:

$$\hat{x}^T \hat{x} = 1, \quad \nabla_x L(\hat{x}, \hat{z}) = 0 \qquad \Longleftrightarrow \qquad \hat{x}^T \hat{x} = 1, \quad A\hat{x} = -\hat{z}\hat{x}$$

- hence optimal $\hat{x}$ must be an eigenvector

Example

$$\begin{array}{ll} \text{minimize} & x_2 \\ \text{subject to} & x_1^2 + x_2^2 = 1 \\ & (x_1 - 2)^2 + x_2^2 = 1 \end{array}$$



- $\hat{x} = (1, 0)$ is the only feasible point, hence optimal
- Lagrangian is $L(x, z) = x_2 + z_1(x_1^2 + x_2^2 - 1) + z_2((x_1 - 2)^2 + x_2^2 - 1)$
- 1st order optimality condition at $\hat{x} = (1, 0)$:

$$\begin{aligned} 0 = \nabla_x L(\hat{x}, \hat{z}) &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 2\hat{z}_1 \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \end{bmatrix} + 2\hat{z}_2 \begin{bmatrix} \hat{x}_1 - 2 \\ \hat{x}_2 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 1 \end{bmatrix} + 2\hat{z}_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + 2\hat{z}_2 \begin{bmatrix} -1 \\ 0 \end{bmatrix} \end{aligned}$$

- this does not hold for any $\hat{z}_1$, $\hat{z}_2$
- $\hat{x}$ is not a regular point: gradients $(2, 0)$ and $(-2, 0)$ are linearly dependent

Outline

- Lagrange multipliers
- **constrained nonlinear least squares**
- penalty method
- augmented Lagrangian method
- nonlinear control example

Constrained nonlinear least squares

$$\begin{array}{ll}\text{minimize} & f_1(x)^2 + \cdots + f_m(x)^2 \\ \text{subject to} & g_1(x) = 0, \ \ldots, \ g_p(x) = 0\end{array}$$

- variable is n-vector x
- $f_i(x)$ is ith (scalar) *residual*
- $g_i(x) = 0$ is ith (scalar) equality constraint

Vector notation

$$\begin{array}{ll}\text{minimize} & \|f(x)\|^2 \\ \text{subject to} & g(x) = 0\end{array}$$

- $f : \mathbf{R}^n \to \mathbf{R}^m$ is vector function $f(x) = (f_1(x), \ldots, f_m(x))$
- $g : \mathbf{R}^n \to \mathbf{R}^p$ is vector function $g(x) = (g_1(x), \ldots, g_p(x))$

First-order necessary optimality condition

Lagrangian

$$\begin{aligned} L(x,z) &= f_1(x)^2 + \cdots + f_m(x)^2 + z_1 g_1(x) + \cdots + z_p g_p(x) \\ &= \|f(x)\|^2 + z^T g(x) \end{aligned}$$

Gradients of Lagrangian: $\nabla_z L(\hat{x},\hat{z}) = g(\hat{x})$ and

$$\begin{aligned} \nabla_x L(\hat{x},\hat{z}) &= 2Df(\hat{x})^T f(\hat{x}) + Dg(\hat{x})^T \hat{z} \\ &= 2\begin{bmatrix} \nabla f_1(\hat{x}) & \cdots & \nabla f_m(\hat{x}) \end{bmatrix} \begin{bmatrix} f_1(\hat{x}) \\ \vdots \\ f_m(\hat{x}) \end{bmatrix} + \begin{bmatrix} \nabla g_1(\hat{x}) & \cdots & \nabla g_p(\hat{x}) \end{bmatrix} \begin{bmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_p \end{bmatrix} \end{aligned}$$

Optimality condition: if $\hat{x}$ is locally optimal, then there exists $\hat{z}$ such that

$$2Df(\hat{x})^T f(\hat{x}) + Dg(\hat{x})^T \hat{z} = 0, \qquad g(\hat{x}) = 0$$

(provided the rows of $Dg(\hat{x})$ are linearly independent)

Constrained (linear) least squares

$$\begin{array}{ll} \text{minimize} & \|Ax - b\|^2 \\ \text{subject to} & Cx = d \end{array}$$

- a special case of the nonlinear problem with

$$f(x) = Ax - b, \qquad g(x) = Cx - d$$

- apply general optimality condition:

$$2Df(\hat{x})^T f(\hat{x}) + Dg(\hat{x})^T \hat{z} = 2A^T(A\hat{x} - b) + C^T \hat{z} = 0, \qquad g(\hat{x}) = C\hat{x} - d = 0$$

- these are the Karush–Kuhn–Tucker (KKT) equations

$$\left[\begin{array}{cc} 2A^TA & C^T \\ C & 0 \end{array}\right] \left[\begin{array}{c} \hat{x} \\ \hat{z} \end{array}\right] = \left[\begin{array}{c} 2A^Tb \\ d \end{array}\right]$$

Outline

- Lagrange multipliers
- constrained nonlinear least squares
- **penalty method**
- augmented Lagrangian method
- nonlinear control example

Penalty method

solve a sequence of (unconstrained) nonlinear least squares problems

$$\text{minimize} \quad \|f(x)\|^2 + \mu\|g(x)\|^2 = \left\| \begin{bmatrix} f(x) \\ \sqrt{\mu}g(x) \end{bmatrix} \right\|^2$$

- μ is a positive *penalty parameter*
- instead of insisting on $g(x) = 0$ we assign a penalty to deviations from zero
- for increasing sequence $\mu^{(1)}, \mu^{(2)}, \ldots$, we compute $x^{(k+1)}$ by minimizing

$$\|f(x)\|^2 + \mu^{(k)}\|g(x)\|^2$$

- $x^{(k+1)}$ is computed by Levenberg–Marquardt algorithm started at $x^{(k)}$

Termination

optimality condition for constrained nonlinear least squares problem:

$$2Df(\hat{x})^T f(\hat{x}) + Dg(\hat{x})^T \hat{z} = 0, \qquad g(\hat{x}) = 0 \tag{1}$$

- $x^{(k)}$ in penalty method satisfies normal equations for linear least squares:

$$2Df(x^{(k)})^T f(x^{(k)}) + 2\mu^{(k-1)} Dg(x^{(k)})^T g(x^{(k)}) = 0$$

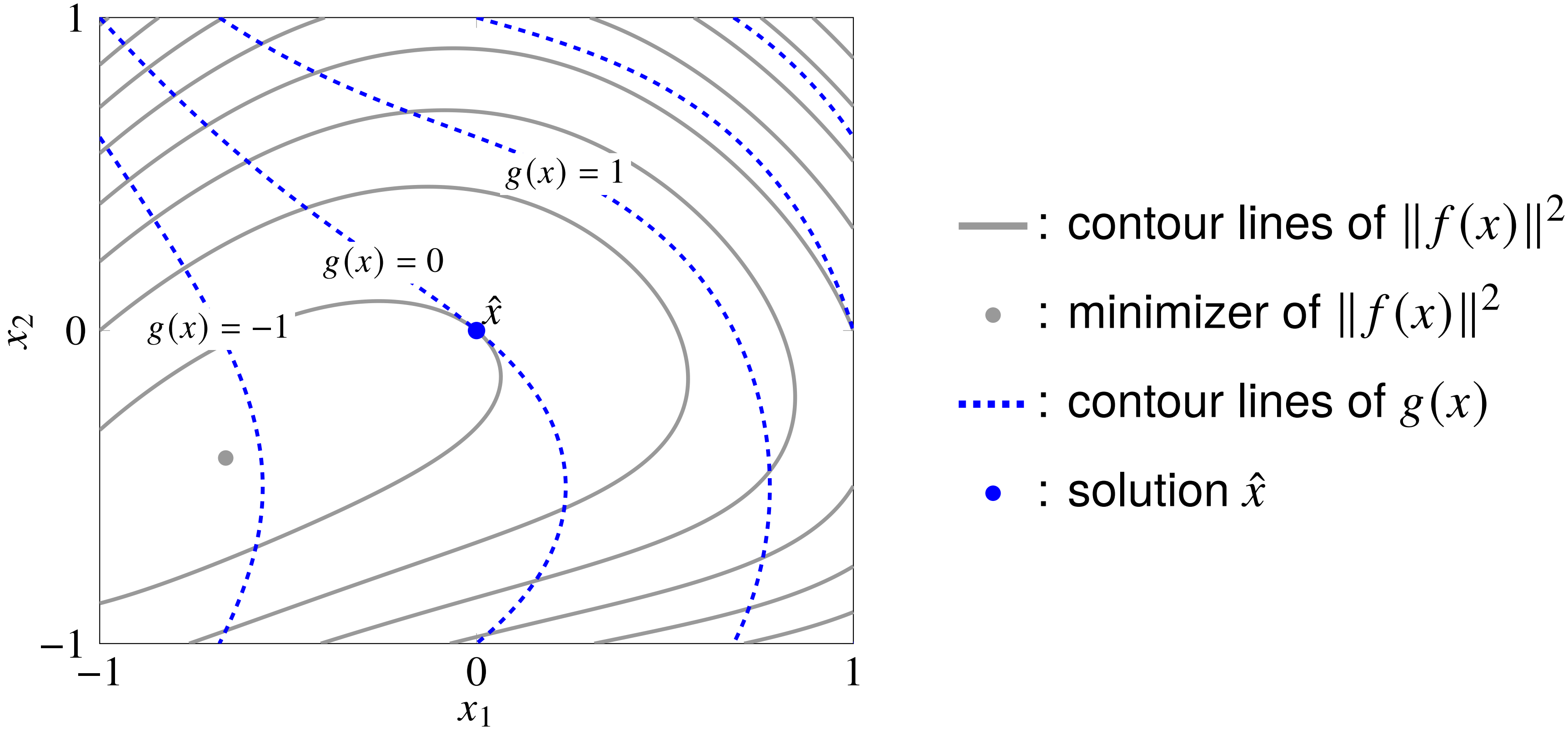
- if we define $z^{(k)} = 2\mu^{(k-1)} g(x^{(k)})$, this can be written as

$$2Df(x^{(k)})^T f(x^{(k)}) + Dg(x^{(k)})^T z^{(k)} = 0$$

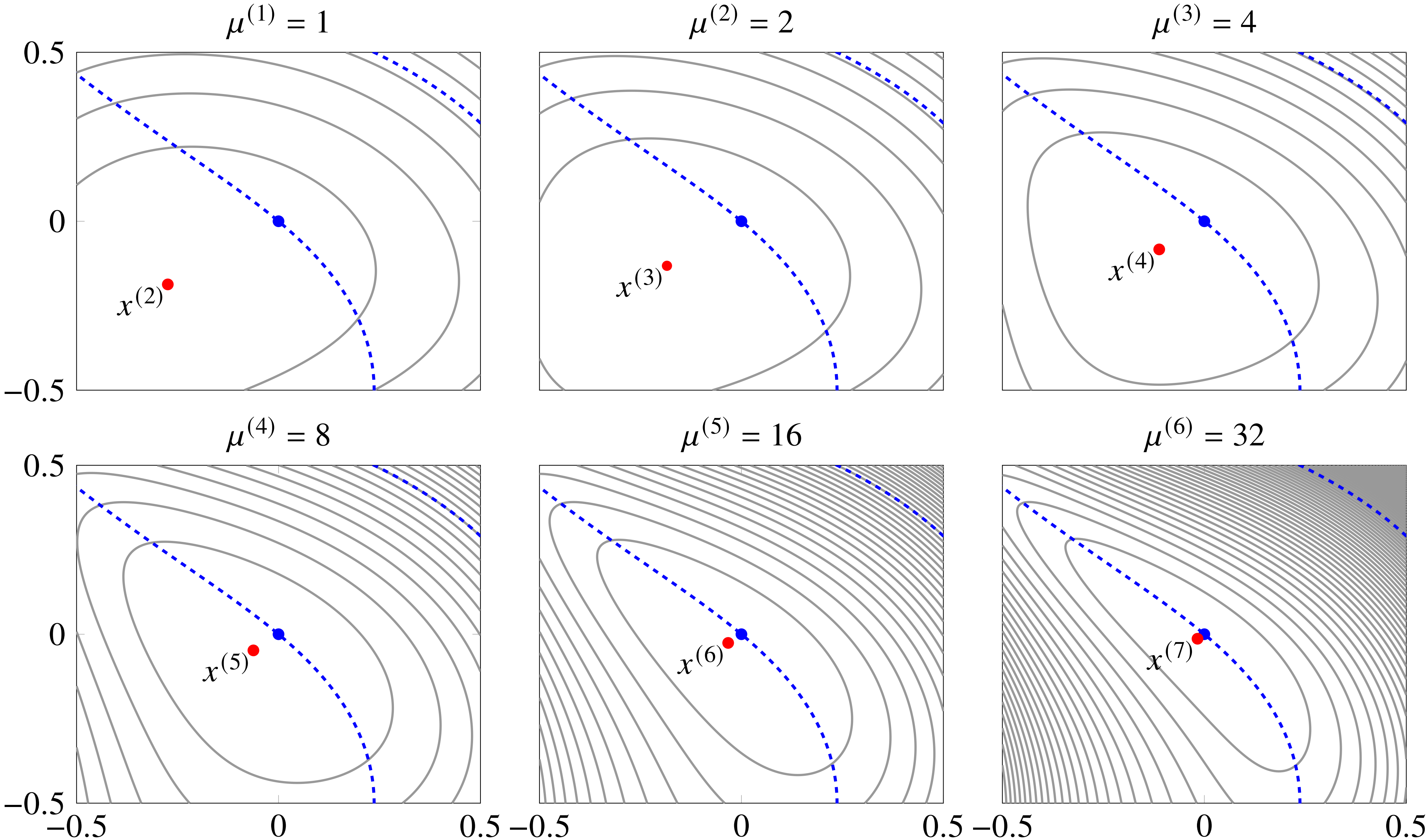
- we see that $x^{(k)}$, $z^{(k)}$ satisfy first equation in optimality condition (1)
- feasibility $g(x^{(k)}) = 0$ is only satisfied approximately for $\mu^{(k-1)}$ large enough
- penalty method is terminated when $\|g(x^{(k)})\|$ becomes sufficiently small

Example

$$f(x_1, x_2) = \begin{bmatrix} x_1 + \exp(-x_2) \\ x_1^2 + 2x_2 + 1 \end{bmatrix}, \qquad g(x_1, x_2) = x_1 + x_1^3 + x_2 + x_2^2$$

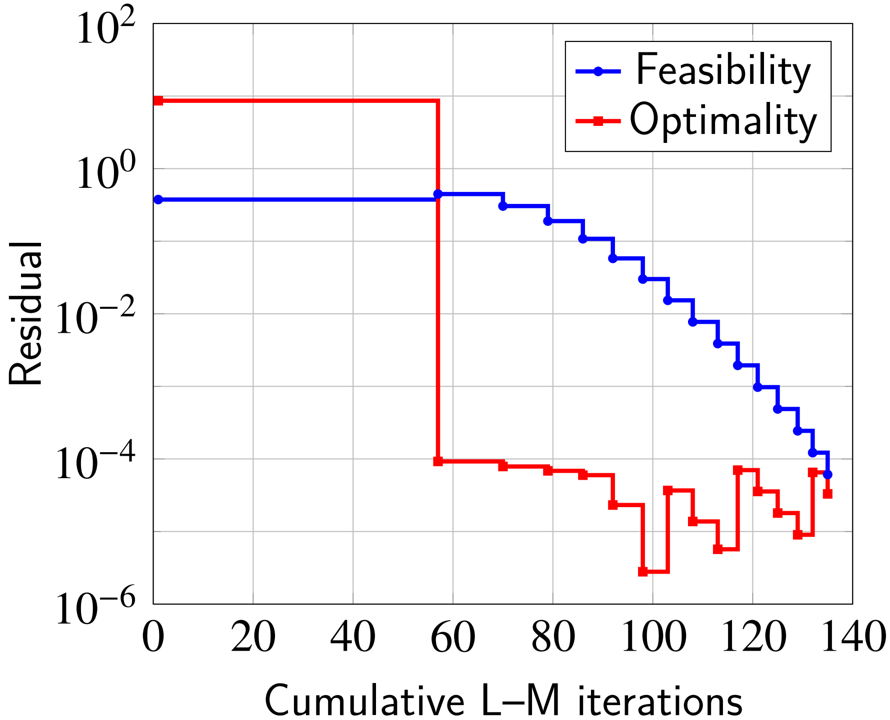


First six iterations



—— : contour lines of $\|f(x)\|^2 + \mu^{(k)}\|g(x)\|^2$

Convergence



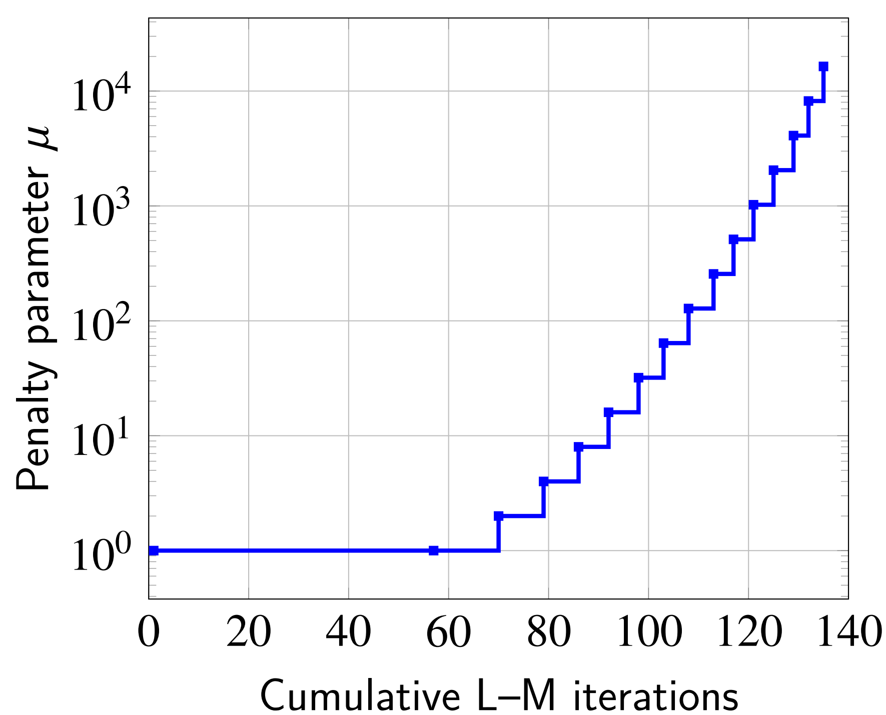


figure on the left shows the two residuals in optimality condition:

- blue curve is norm of $g(x^{(k)})$
- red curve is norm of $2Df(x^{(k)})^T f(x^{(k)}) + Dg(x^{(k)})^T z^{(k)}$

Drawback of penalty method

- $\mu^{(k)}$ increases rapidly and must become large to drive $g(x)$ to (near) zero
- for large $\mu^{(k)}$, nonlinear least squares subproblem becomes harder
- for large $\mu^{(k)}$, Levenberg–Marquardt method can take many iterations, or fail

Outline

- Lagrange multipliers
- constrained nonlinear least squares
- penalty method
- **augmented Lagrangian method**
- nonlinear control example

Augmented Lagrangian

the *augmented Lagrangian* for the constrained NLLS problem is

$$\begin{aligned} L_\mu(x,z) &= L(x,z) + \mu\|g(x)\|^2 \\ &= \|f(x)\|^2 + g(x)^T z + \mu\|g(x)\|^2 \end{aligned}$$

- this is the Lagrangian $L(x,z)$ augmented with a quadratic penalty
- μ is a positive penalty parameter
- augmented Lagrangian is the Lagrangian of the equivalent problem

$$\begin{array}{ll} \text{minimize} & \|f(x)\|^2 + \mu\|g(x)\|^2 \\ \text{subject to} & g(x) = 0 \end{array}$$

Minimizing augmented Lagrangian

- equivalent expressions for augmented Lagrangian

$$
\begin{aligned}
L_\mu(x,z) &= \|f(x)\|^2 + g(x)^T z + \mu\,\|g(x)\|^2 \\
&= \|f(x)\|^2 + \mu\,\left\|g(x) + \frac{1}{2\mu} z\right\|^2 - \frac{1}{2\mu}\,\|z\|^2 \\
&= \left\|\begin{bmatrix} f(x) \\ \sqrt{\mu}\,g(x) + \frac{1}{2\sqrt{\mu}}\,z \end{bmatrix}\right\|^2 - \frac{1}{2\mu}\,\|z\|^2
\end{aligned}
$$

- can be minimized over x (for fixed μ, z) by Levenberg–Marquardt method:

$$
\text{minimize} \quad \left\|\begin{bmatrix} f(x) \\ \sqrt{\mu}\,g(x) + \frac{1}{2\sqrt{\mu}}\,z \end{bmatrix}\right\|^2
$$

Lagrange multiplier update

optimality conditions for constrained nonlinear least squares problem:

$$2Df(\hat{x})^T f(\hat{x}) + Dg(\hat{x})^T \hat{z} = 0, \qquad g(\hat{x}) = 0$$

- minimizer $\tilde{x}$ of augmented Lagrangian $L_\mu(x, z)$ satisfies

$$2Df(\tilde{x})^T f(\tilde{x}) + Dg(\tilde{x})^T (2\mu g(\tilde{x}) + z) = 0$$

- first equation in optimality condition is satisfied if we define

$$\tilde{z} = z + 2\mu g(\tilde{x})$$

- this shows that if $g(\tilde{x}) = 0$, then $\tilde{x}$ is optimal
- if $g(\tilde{x})$ is not small, suggests $\tilde{z}$ is a good update for z

Augmented Lagrangian algorithm

1. set $x^{(k+1)}$ to be the (approximate) minimizer of

$$\|f(x)\|^2 + \mu^{(k)} \left\| g(x) + \frac{1}{2\mu^{(k)}} z^{(k)} \right\|^2$$

   $x^{(k+1)}$ is computed using Levenberg–Marquardt algorithm, starting from $x^{(k)}$

2. *multiplier update:*

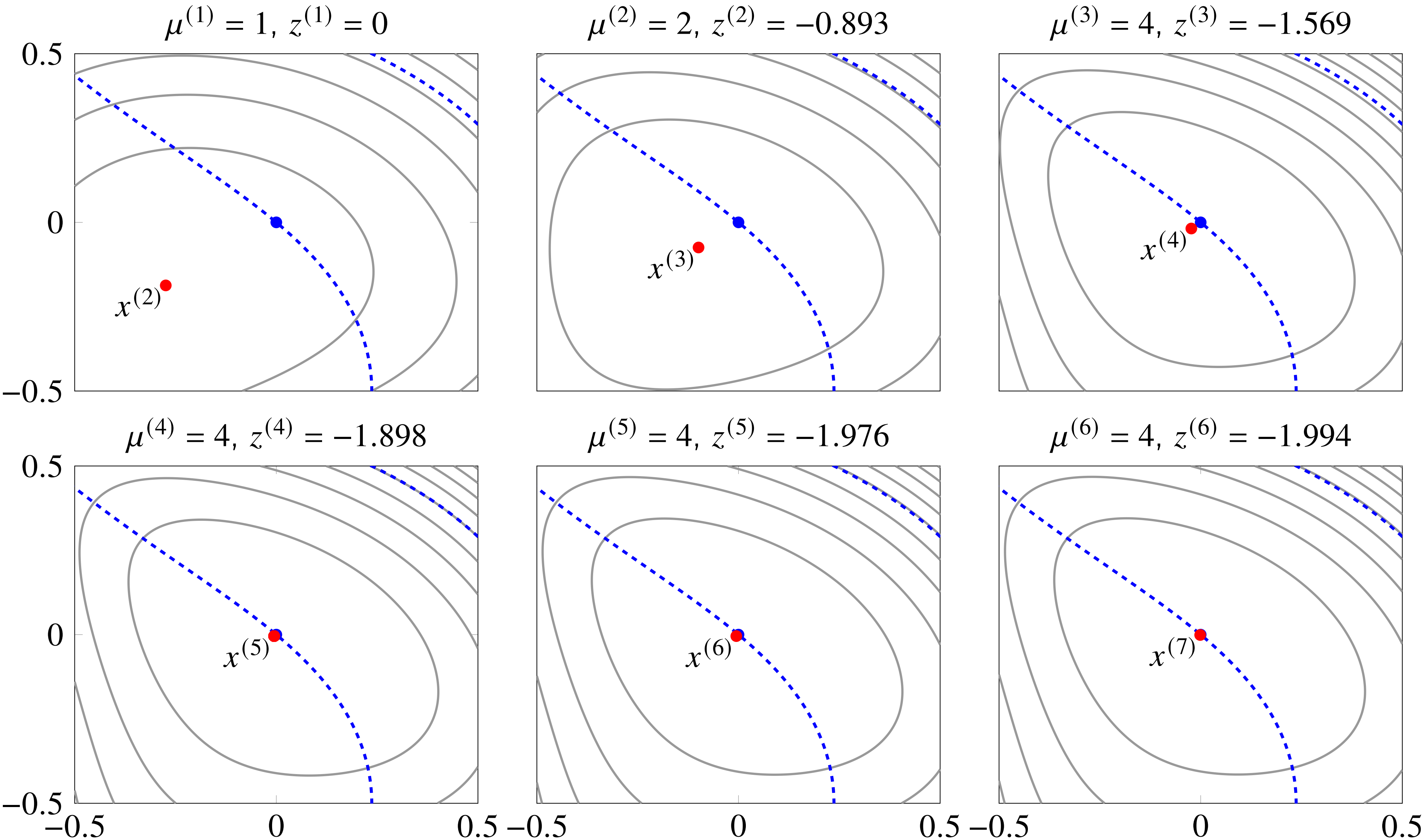
$$z^{(k+1)} = z^{(k)} + 2\mu^{(k)} g(x^{(k+1)})$$

3. *penalty parameter update:*

$$\mu^{(k+1)} = \begin{cases} \mu^{(k)} & \text{if } \|g(x^{(k+1)})\| < 0.25\|g(x^{(k)})\| \\ 2\mu^{(k)} & \text{otherwise} \end{cases}$$

- iteration starts at $z^{(1)} = 0$, $\mu^{(1)} = 1$, some initial $x^{(1)}$
- μ is increased only when needed, more slowly than in penalty method
- continues until $g(x^{(k)})$ is sufficiently small (or iteration limit is reached)

Example of slide 13.14



: contour lines of augmented Lagrangian $L_{\mu^{(k)}}(x, z^{(k)})$

Convergence

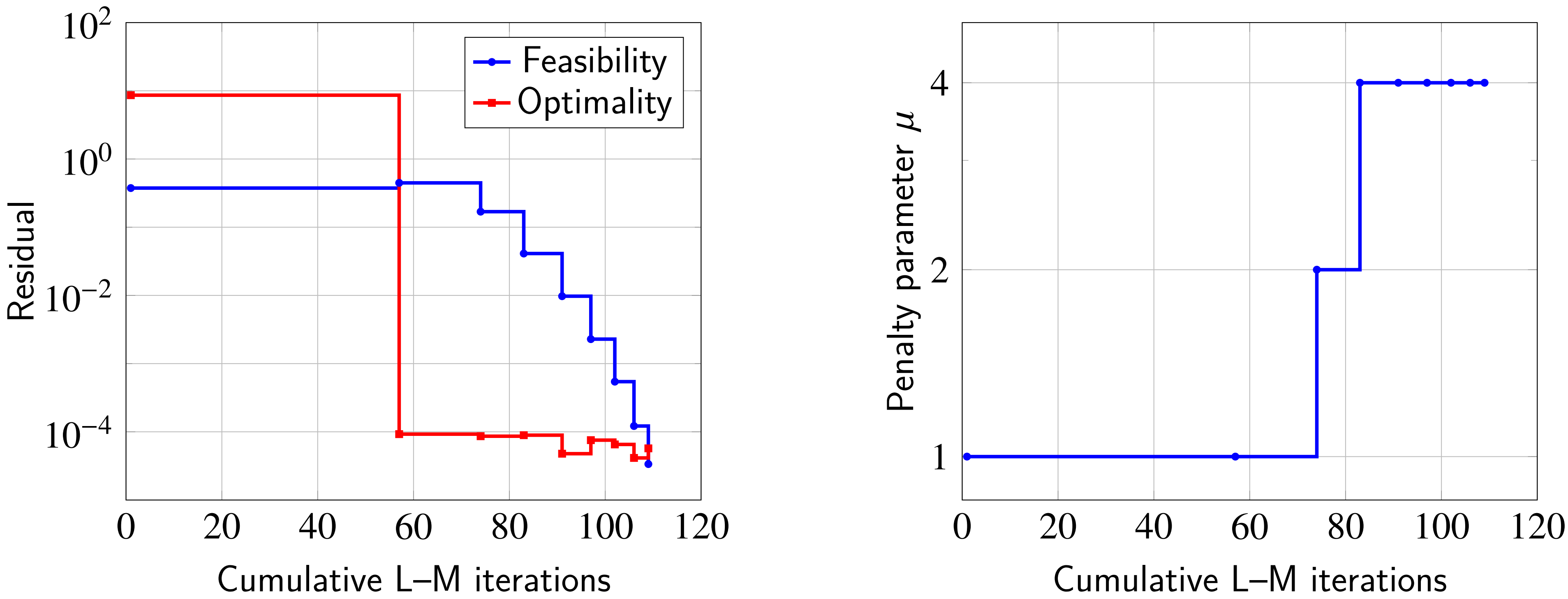


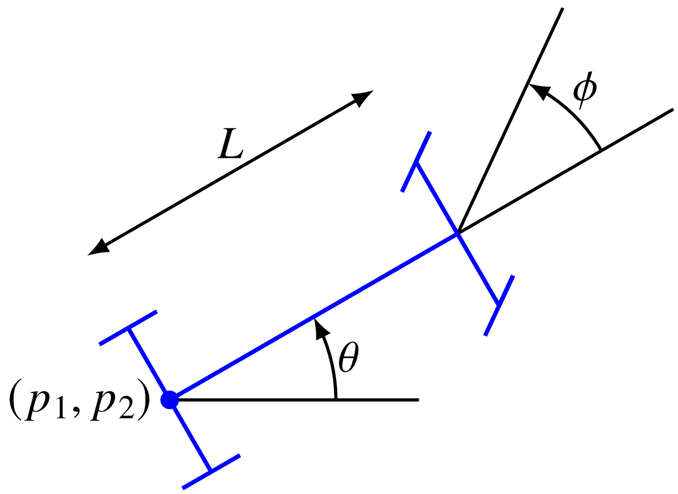
figure on the left shows residuals in optimality condition:

- blue curve is norm of $g(x^{(k)})$
- red curve is norm of $2Df(x^{(k)})^T f(x^{(k)}) + Dg(x^{(k)})^T z^{(k)}$

Outline

- Lagrange multipliers
- constrained nonlinear least squares
- penalty method
- augmented Lagrangian method
- **nonlinear control example**

Simple model of a car



$$\begin{aligned}
\frac{dp_1}{dt} &= s(t)\cos\theta(t) \\
\frac{dp_2}{dt} &= s(t)\sin\theta(t) \\
\frac{d\theta}{dt} &= \frac{s(t)}{L}\tan\phi(t)
\end{aligned}$$

- $s(t)$ is speed of vehicle
- $\phi(t)$ is steering angle
- $p(t)$ is position
- $\theta(t)$ is orientation

Discretized model

- discretized model (for small time interval h):

$$\begin{aligned} p_1(t+h) &\approx p_1(t) + hs(t)\cos(\theta(t)) \\ p_2(t+h) &\approx p_2(t) + hs(t)\sin(\theta(t)) \\ \theta(t+h) &\approx \theta(t) + h\frac{s(t)}{L}\tan(\phi(t)) \end{aligned}$$

- define input vector $u_k = (s(kh), \phi(kh))$
- define state vector $x_k = (p_1(kh), p_2(kh), \theta(kh))$
- discretized model is $x_{k+1} = f(x_k, u_k)$ with

$$f(x_k, u_k) = \left[\begin{array}{c} (x_k)_1 + h(u_k)_1\cos((x_k)_3) \\ (x_k)_2 + h(u_k)_1\sin((x_k)_3) \\ (x_k)_3 + h(u_k)_1\tan((u_k)_2)/L \end{array}\right]$$

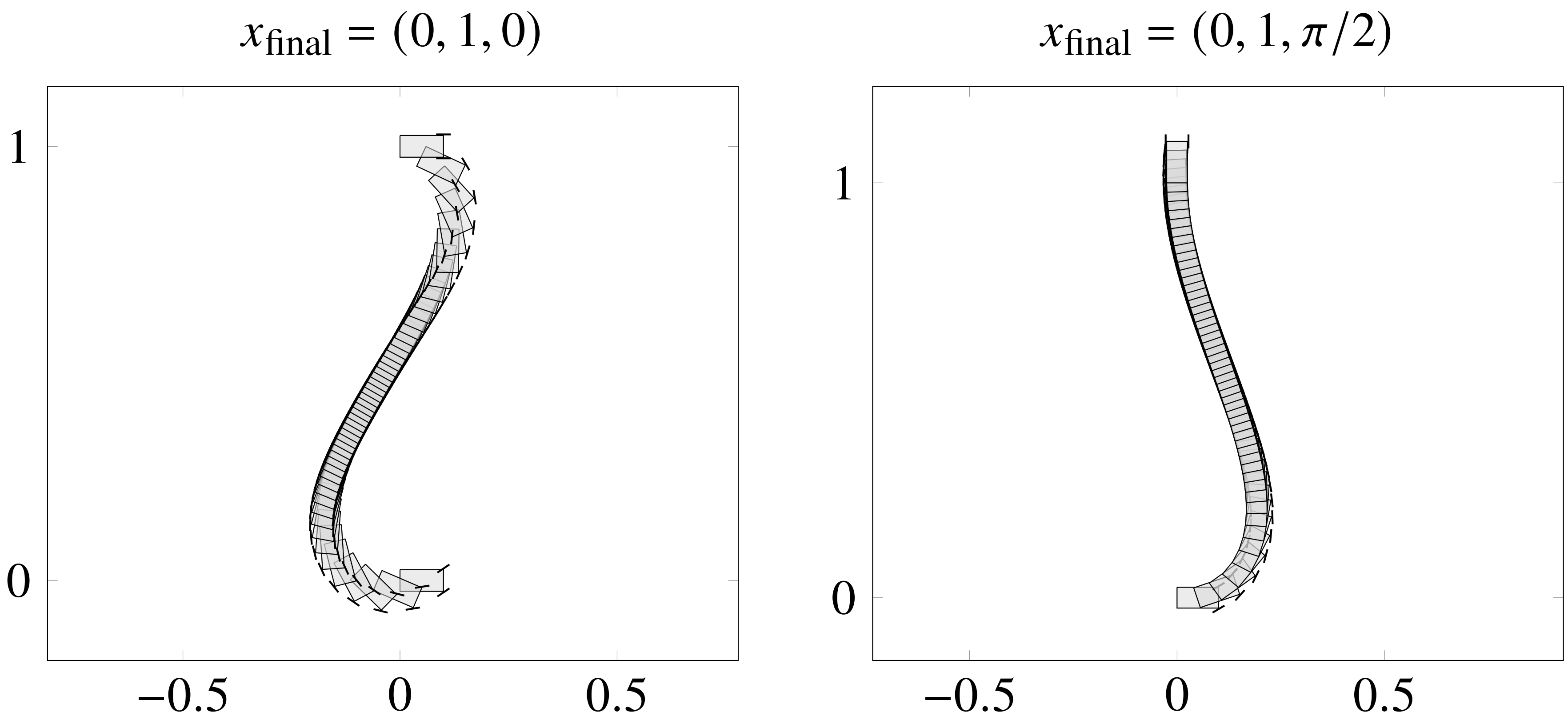
Control problem

- move car from given initial to desired final position and orientation
- using a small and slowly varying input sequence
- this is a constrained nonlinear least squares problem:

$$
\begin{array}{ll}
\text{minimize} & \sum_{k=1}^{N} \|u_k\|^2 + \gamma \sum_{k=1}^{N-1} \|u_{k+1} - u_k\|^2 \\
\text{subject to} & x_2 = f(0, u_1) \\
& x_{k+1} = f(x_k, u_k), \quad k = 2, \ldots, N-1 \\
& x_{\text{final}} = f(x_N, u_N)
\end{array}
$$

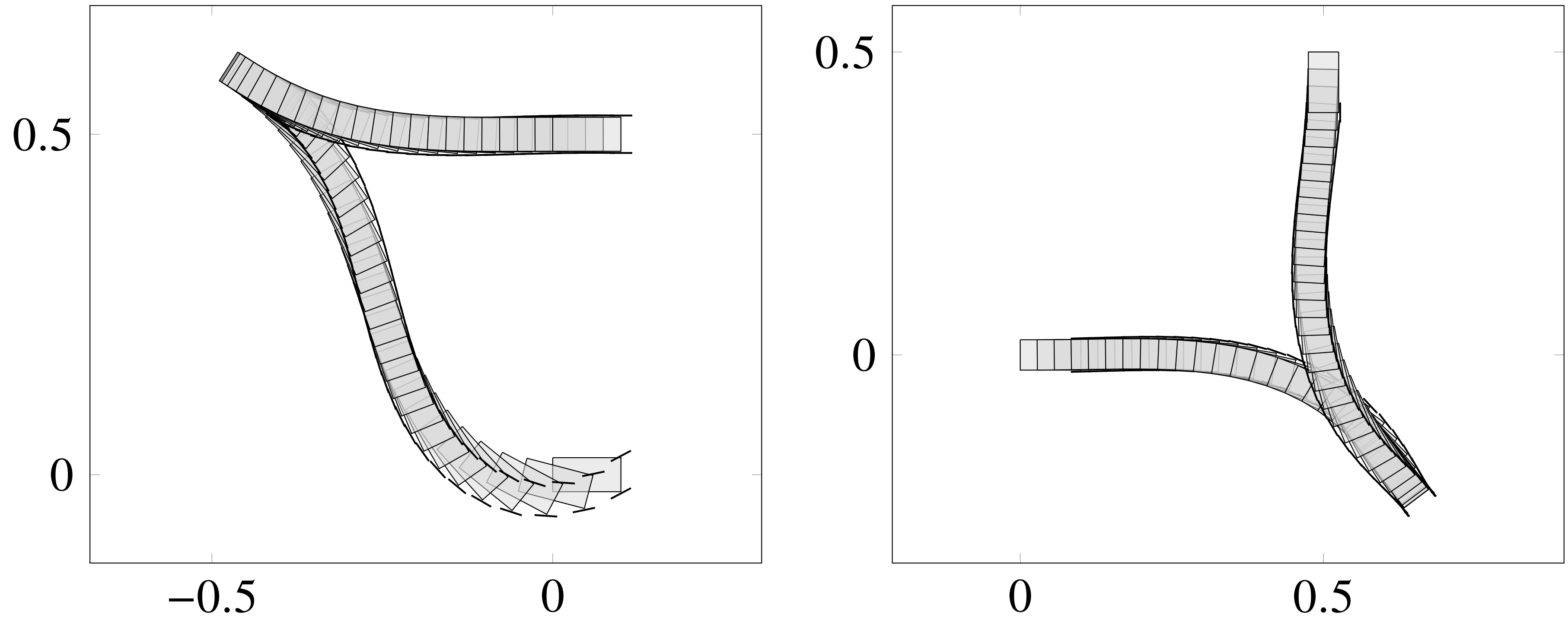
- variables are $u_1, \ldots, u_N, x_2, \ldots, x_N$

Example solution trajectories



Example solution trajectories

$x_{\text{final}} = (0, 0.5, 0)$ $\qquad$ $x_{\text{final}} = (0.5, 0.5, -\pi/2)$



Inputs for four trajectories