Lecture 12: Penalty methods for constrained optimization problems

Coralia Cartis, Mathematical Institute, University of Oxford

C6.2/B2: Continuous Optimization

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).

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.

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)

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

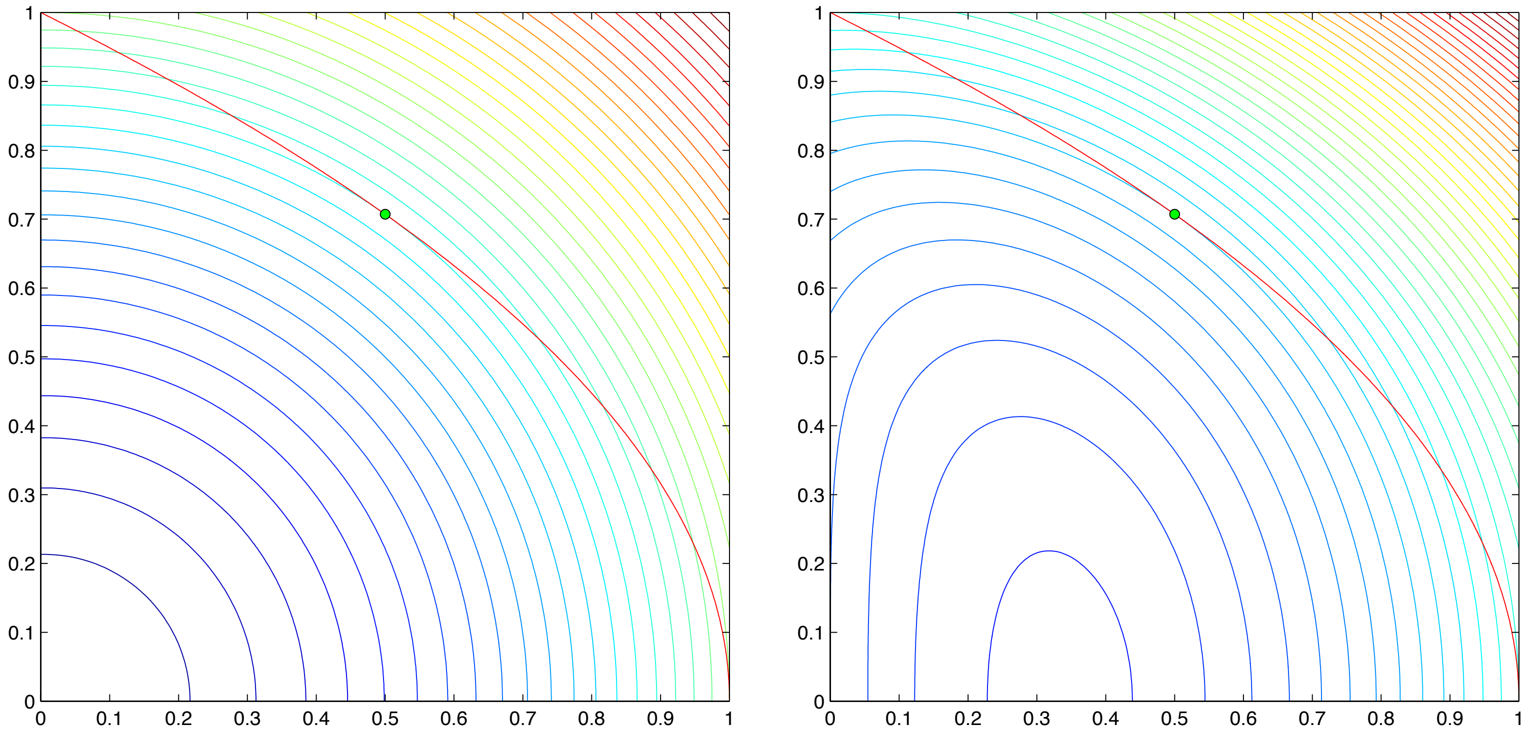
The quadratic penalty function:

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

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.

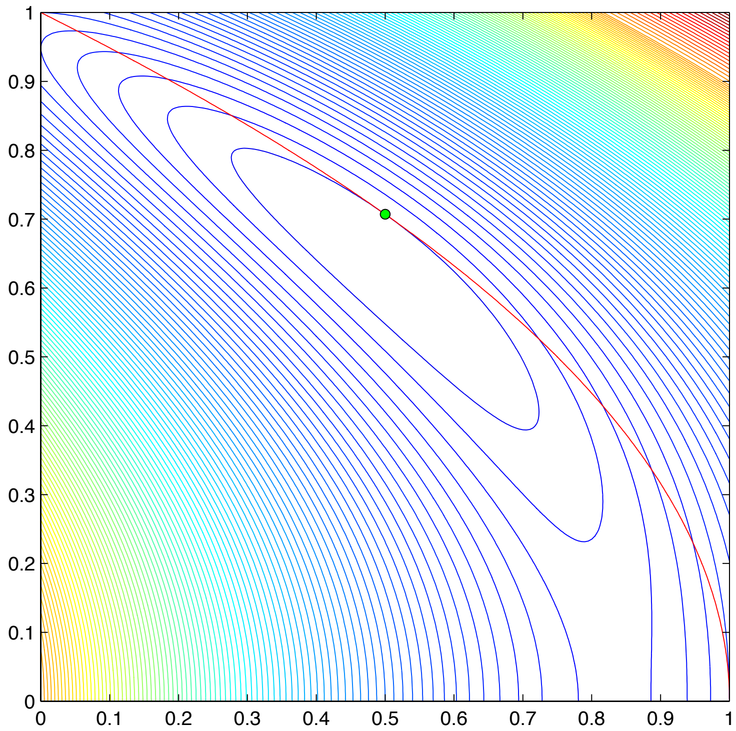
Contours of the penalty function Φ_σ - an example



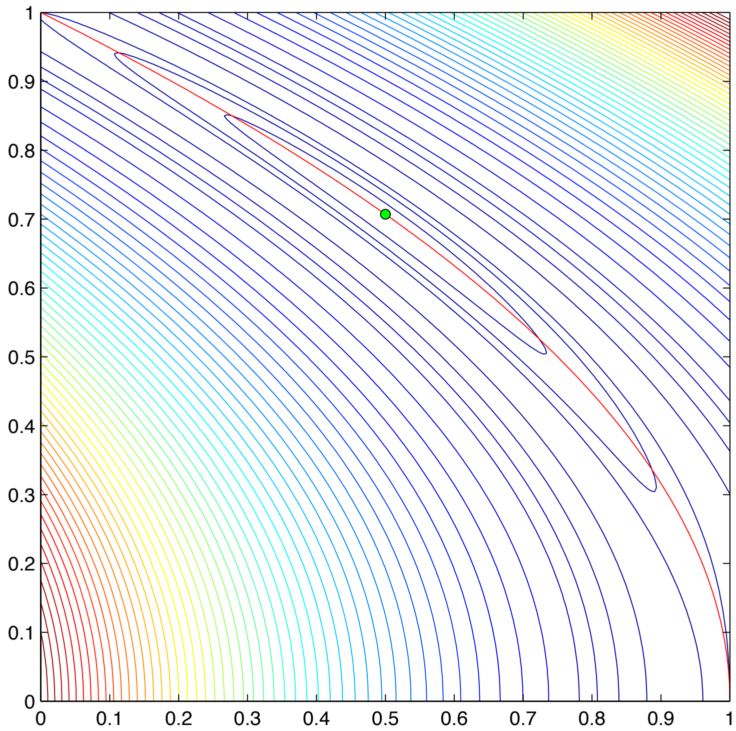
$\sigma = 100$ $\qquad\qquad$ $\sigma = 1$

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.

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. □

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 $\|c(x)\|$. $\big[$let $y^k \to \infty$ in ($\Diamond$) on the next slide$\big]$

A convergence result for the penalty method

Proof of Theorem 21.

- The KKT conditions for (eCP) are: $c(x^*) = 0$ and $\nabla f(x^*) = J(x^*)^T y^*$ for some $y \in \mathbb{R}^m$.

A convergence result for the penalty method

Proof of Theorem 21.

- The KKT conditions for (eCP) are: $c(x^*) = 0$ and $\nabla f(x^*) = J(x^*)^T y^*$ for some $y \in \mathbb{R}^m$.
- Recalling how we calculated the derivatives of a nonlinear least square function (Lecture 7); a similar calculation here gives

$$\nabla \Phi_{\sigma^k}(x^k) = \nabla f(x^k) - \frac{1}{\sigma_k} J(x^k)^T c(x^k) = \nabla f(x^k) - J(x^k)^T y^k, \quad (\Diamond)$$

where $y^k = -c(x^k)/\sigma_k \in \mathbb{R}^m$.

A convergence result for the penalty method

Proof of Theorem 21.

- The KKT conditions for (eCP) are: $c(x^*) = 0$ and $\nabla f(x^*) = J(x^*)^T y^*$ for some $y \in \mathbb{R}^m$.
- Recalling how we calculated the derivatives of a nonlinear least square function (Lecture 7); a similar calculation here gives

$$\nabla \Phi_{\sigma^k}(x^k) = \nabla f(x^k) - \frac{1}{\sigma_k} J(x^k)^T c(x^k) = \nabla f(x^k) - J(x^k)^T y^k, \quad (\Diamond)$$

where $y^k = -c(x^k)/\sigma_k \in \mathbb{R}^m$.

- Note also that if

$$\overline{y} = \text{arg} \min_{y \in \mathbb{R}^m} \|\nabla f(x) - J(x)^T y\|^2, \quad (1)$$

then $\overline{y}$ is the solution of an (overdetermined) linear least squares (Lecture 7 again).

A convergence result for the penalty method...

- Thus, from (1), recalling the normal equations, $\overline{y}$ satisfies $J(x)J(x)^T\overline{y} = -J(x)\nabla f(x)$.

A convergence result for the penalty method...

- Thus, from (1), recalling the normal equations, $\overline{y}$ satisfies $J(x)J(x)^T\overline{y} = -J(x)\nabla f(x)$. This is further equivalent to $\overline{y} = -J(x)^{+}\nabla f(x)$, where $J(x)^{+} = (J(x)J(x)^T)^{-1}J(x)$ pseudo-inverse, if $J(x)$ is full row rank (so that $J(x)J(x)^T$ is nonsingular).

A convergence result for the penalty method...

- Thus, from (1), recalling the normal equations, $\overline{y}$ satisfies $J(x)J(x)^T\overline{y} = -J(x)\nabla f(x)$. This is further equivalent to $\overline{y} = -J(x)^+\nabla f(x)$, where $J(x)^+ = (J(x)J(x)^T)^{-1}J(x)$ pseudo-inverse, if $J(x)$ is full row rank (so that $J(x)J(x)^T$ is nonsingular).
- LICQ $\Rightarrow$ $J(x^*)$ full row rank ($m \leq n$) and so $J(x^*)^+$ exists. Since we assumed that $x^k \to x^*$ and J is continuous $\Rightarrow$ $J(x^k)^+$ exists and is bounded above and cont. for all suff. large k.

A convergence result for the penalty method...

- Thus, from (1), recalling the normal equations, $\overline{y}$ satisfies $J(x)J(x)^T\overline{y} = -J(x)\nabla f(x)$. This is further equivalent to $\overline{y} = -J(x)^+\nabla f(x)$, where $J(x)^+ = (J(x)J(x)^T)^{-1}J(x)$ pseudo-inverse, if $J(x)$ is full row rank (so that $J(x)J(x)^T$ is nonsingular).
- LICQ $\Rightarrow J(x^*)$ full row rank ($m \leq n$) and so $J(x^*)^+$ exists. Since we assumed that $x^k \to x^*$ and J is continuous $\Rightarrow$ $J(x^k)^+$ exists and is bounded above and cont. for all suff. large k.

We first show that $y^k \to y^*$, as $k \to \infty$; the multiplier y^* solves (1) with $x = x^*$ and so $y^* = J(x^*)^+\nabla f(x^*)$. We evaluate

A convergence result for the penalty method...

- Thus, from (1), recalling the normal equations, $\overline{y}$ satisfies $J(x)J(x)^T\overline{y} = -J(x)\nabla f(x)$. This is further equivalent to $\overline{y} = -J(x)^+\nabla f(x)$, where $J(x)^+ = (J(x)J(x)^T)^{-1}J(x)$ pseudo-inverse, if $J(x)$ is full row rank (so that $J(x)J(x)^T$ is nonsingular).
- LICQ $\Rightarrow$ $J(x^*)$ full row rank ($m \leq n$) and so $J(x^*)^+$ exists. Since we assumed that $x^k \to x^*$ and J is continuous $\Rightarrow$ $J(x^k)^+$ exists and is bounded above and cont. for all suff. large k.

We first show that $y^k \to y^*$, as $k \to \infty$; the multiplier y^* solves (1) with $x = x^*$ and so $y^* = J(x^*)^+\nabla f(x^*)$. We evaluate

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

As $x^k \to x^*$, J^+ and ∇f continuous, then $J(x^k)^+ \to J(x^*)^+$ and $\nabla f(x^k) \to \nabla f(x^*)$. Thus $\|J(x^k)^+\nabla f(x^k) - J(x^*)^+\nabla f(x^*)\| \to 0$.

A convergence result for the penalty method...

Proof of Theorem 21. As $J(x^k)^+J(x^k) = I$ then

$$\begin{aligned}
\|J(x^k)^+\nabla f(x^k) - y^k\| &= \|J(x^k)^+\nabla f(x^k) - J(x^k)^+J(x^k)y^k\| \\
&\leq \|J(x^k)^+\| \cdot \|\nabla \Phi_{\sigma^k}(x^k)\| \leq C\epsilon_k
\end{aligned}$$

for some C; since $\epsilon_k \to 0$, we deduce $y^k \to y^*$.

A convergence result for the penalty method...

Proof of Theorem 21. As $J(x^k)^+J(x^k) = I$ then

$$\begin{aligned} \|J(x^k)^+\nabla f(x^k) - y^k\| &= \|J(x^k)^+\nabla f(x^k) - J(x^k)^+J(x^k)y^k\| \\ &\leq \|J(x^k)^+\| \cdot \|\nabla\Phi_{\sigma^k}(x^k)\| \leq C\epsilon_k \end{aligned}$$

for some C; since $\epsilon_k \to 0$, we deduce $y^k \to y^*$.

It remains to show that the limit point x^* of $\{x^k\}$ is a KKT point of (eCP). Passing to the limit in ($\Diamond$), and using $x^k \to x^*$, $y^k \to y^*$ and continuity of ∇f and J, we deduce $\nabla\Phi(x^k) \to \nabla f(x^*) - J(x^*)^T y^*$. On the other hand $\nabla\Phi(x^k) \to 0$. Thus $\nabla f(x^*) - J(x^*)^T y^* = 0$.

A convergence result for the penalty method...

Proof of Theorem 21. As $J(x^k)^+ J(x^k) = I$ then

$$\begin{aligned} \|J(x^k)^+ \nabla f(x^k) - y^k\| &= \|J(x^k)^+ \nabla f(x^k) - J(x^k)^+ J(x^k) y^k\| \\ &\leq \|J(x^k)^+\| \cdot \|\nabla \Phi_{\sigma^k}(x^k)\| \leq C\epsilon_k \end{aligned}$$

for some C; since $\epsilon_k \to 0$, we deduce $y^k \to y^*$.

It remains to show that the limit point x^* of $\{x^k\}$ is a KKT point of (eCP). Passing to the limit in ($\Diamond$), and using $x^k \to x^*$, $y^k \to y^*$ and continuity of ∇f and J, we deduce $\nabla \Phi(x^k) \to \nabla f(x^*) - J(x^*)^T y^*$. On the other hand $\nabla \Phi(x^k) \to 0$. Thus $\nabla f(x^*) - J(x^*)^T y^* = 0$.

The definition of y^k gives

$$c(x^k) = -\sigma^k y^k.$$

$\sigma^k \to 0$, $y^k \to y^*$ implies RHS $\to 0$. For the LHS, $c(x^k) \to 0$ as $x^k \to x^*$. $\Rightarrow$ $c(x^*) = 0$. Thus x^* KKT. $\Box$

Derivatives of the penalty function

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

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

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

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 $c(x)$.

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 $c(x)$.

[recall Lecture 7, nonlinear least squares derivatives]

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

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 $c(x)$.

[recall Lecture 7, nonlinear least squares derivatives]

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

- $\sigma \longrightarrow 0$: 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 $k \to \infty$.

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 $\sigma = \sigma^{k+1}$.

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

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

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

Then $\nabla^2_{xx} L(x, y(\sigma))dx + J(x)^T w = -\nabla f(x)$, and 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 $\sigma \to 0$.

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

$$w := \frac{1}{\sigma}\,(J(x)dx + c(x)) \;\leftrightarrow\; J(x)dx - \sigma w = -c(x).$$

Then $\nabla^2_{xx} L(x, y(\sigma))dx + J(x)^T w = -\nabla f(x)$, and 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.

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)