Numerical Optimisation
Constraint optimisation:
Penalty and augmented Lagrangian methods

Marta M. Betcke

m.betcke@ucl.ac.uk

Department of Computer Science,
Centre for Medical Image Computing,
Centre for Inverse Problems
University College London

Lecture 14 (based on Nocedal, Wright)

Lagrangian: primal problem

Constraint optimisation problem (general nonlinear)

$$
\begin{aligned}
\min_{x\in\mathcal{D}\subset\mathbb{R}^n} \quad & f(x) && \text{(COP)}\\
\text{subject to} \quad & f_i(x) \leq 0, \quad i = 1,\dots,m,\\
& h_i(x) = 0, \quad i = 1,\dots,p
\end{aligned}
$$

Possibly conflicting goals: minimise the function and satisfy the constraints.

Idea: Minimise a merit function $Q(x;\mu)$ with a parameter vector μ. Some minimisers of $Q(x;\mu)$ approach those of f subject to the constraints as μ approach some set $\mathcal{M}$.

Benefit: reformulation as an unconstraint problem.

Quadratic penalty

Consider a problem with equality constraints

$$\min_{x \in \mathcal{D} \subset \mathbb{R}^n} \quad f(x) \qquad \text{(COP:E)}$$
$$\text{subject to} \quad h_i(x) = 0, \quad i = 1, \dots, p.$$

The merit function (*quadratic penalty function*)

$$Q(x; \mu) := f(x) + \frac{\mu}{2} \sum_{i=1}^{p} h_i^2(x), \qquad \text{(Q)}$$

where $\mu > 0$ is the *penalty parameter*.

Framework: For a sequence $\{\mu_k\}: \ \mu_k \to \infty$ as $k \to \infty$ increasingly penalising the constraint compute the (approximate, $\|\nabla_x Q(x_k; \mu_k)\| \le \tau_k$, $\tau_k \to 0$) sequence $\{x_k\} \to x^\star$ of minimisers x_k of $Q(x; \mu_k)$.

Convergence for the quadratic penalty

Let $\{x_k\}$ be the sequence of approximate minimisers of $Q(x; \mu_k)$, such that $\|\nabla_x Q(x_k; \mu_k)\| \leq \tau_k$, $x^\star$ be the limit point of $\{x_k\}$ as the sequences of the penalty parameters $\mu_k \to \infty$ and tolerances, $\tau_k \to 0$.

- If a limit point $x^\star$ is infeasible, it is a stationary point of $\|h(x)\|^2$.
- If a limit point $x^\star$ is feasible and the constraint gradients $\nabla h_i(x^\star)$ are linearly independent, then $x^\star$ is a KKT point for (COP:E), and we have that

$$\lim_{k\to\infty} \mu_k h_i(x_k) = \nu_i^\star, \quad i = 1, \ldots, p,$$

where $\nu^\star$ is the Lagrange multiplier vector that satisfies the KKT conditions for (COP:E).

Proof:

$$\nabla_x Q(x_k; \mu_k) = \nabla f(x_k) + \sum_{i=1}^{p} \mu_k h_i(x_k) \nabla h_i(x_k) \qquad \text{(dQ)}$$

From the convergence criterium $\|\nabla_x Q(x_k; \mu_k)\| \leq \tau_k$ (using the inequality $\|a\| - \|b\| \leq \|a + b\|$) we obtain

$$\left\| \sum_{i=1}^{p} h_i(x_k) \nabla h_i(x_k) \right\| \leq \frac{1}{\mu_k} \left(\tau_k + \|\nabla f(x_k)\| \right).$$

As $k \to \infty$: $\tau_k \to 0$, $\|\nabla f(x_k)\| \to \|\nabla f(x^\star)\|$ and $\mu_k \to \infty$ thus the limit of the sequence on the l.h.s. is

$$\sum_{i=1}^{p} h_i(x^\star) \nabla h_i(x^\star) = 0.$$

i) If $\exists i \in \{1, \dots, p\} : h_i(x^\star) \neq 0$ then $\nabla h_i(x^\star)$ are linearly dependent which implies that $x^\star$ is a stationary point of $\|h(x)\|^2$.

ii) If $\nabla h_i(x^\star), i = 1, \dots, p$ are linearly independent, $h_i(x^\star) = 0, i = 1, \dots, p$ and $x^\star$ is primarily feasible i.e. satisfies the second KKT condition. It remains to show that the "dual feasibility" (the first KKT condition) is satisfied.

Case ii):

Intuition:

As $k \to \infty$, $Q(x^k)$ should approach the Lagrangian

$$\mathcal{L}(x^\star; \nu^\star) = f(x^\star) + \sum_{i=1}^{p} \nu_i^\star h_i(x^\star). \tag{L}$$

and $\nabla_x Q(x^k)$ its derivative i.e. the "dual feasibility" condition

$$\nabla_x \mathcal{L}(x^\star; \nu^\star) = \nabla f(x^\star) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x^\star). \tag{dL}$$

Rearranging (dQ) and denoting $A(x)^{\mathrm{T}} := \nabla h_i(x_k), i = 1, \dots, p$ and $\nu^k := \mu_k h(x_k)$ we obtain

$$A(x_k)^{\mathrm{T}} \nu^k = -\nabla f(x_k) + \nabla_x Q(x_k; \mu_k), \quad \|\nabla_x Q(x_k; \mu_k)\| \leq \tau_k.$$

For large enough k the matrix $A(x_k)$ has full row rank and hence the above overdetermined system has the unique solution

$$\nu^k = \left(A(x_k) A(x_k)^{\mathrm{T}}\right)^{-1} A(x_k)[-\nabla f(x_k) + \nabla_x Q(x_k; \mu_k)].$$

Taking the limit as $k \to \infty$

$$\lim_{k \to \infty} \nu^k = \nu^\star = -\left(A(x^\star) A(x^\star)^{\mathrm{T}}\right)^{-1} A(x^\star) \nabla f(x^\star)$$

and the same in (dQ) yields the "dual feasibility" condition

$$\nabla f(x^\star) + A(x^\star)^{\mathrm{T}} \nu^\star = 0.$$

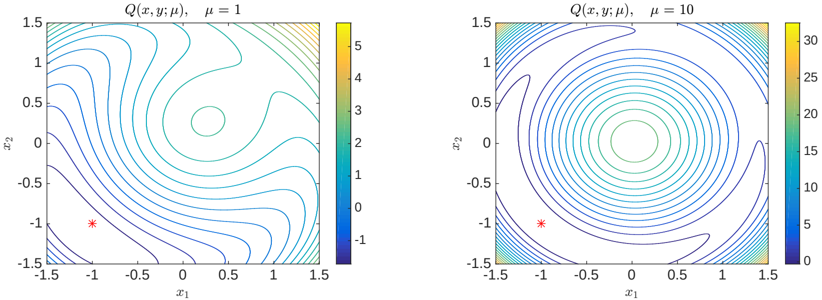
Hence, $x^\star$ is the KKT point with unique Lagrange multiplier $\nu^\star$.

Example

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 - 2 = 0. \end{aligned}$$

Solution: $(-1, -1)^{\mathrm{T}}$.

Quadratic penalty function: $Q(x; \mu) = x_1 + x_2 + \frac{\mu}{2}(x_1^2 + x_2^2 - 2)^2$.



Problems

For equality constraints, $Q(x;\mu)$ is smooth and can be solved with methods for unconstraint optimisation.

- Hessian ill-conditioning (see next slide) poses convergence problems for methods like CG or quasi Newton and affects Newton method's numerical accuracy (which however can be remedied by reformulation).
- For larger μ the quadratic model underlying most solvers is a poor approximation to $Q(x;\mu)$.
- Example:

$$\begin{aligned} \min \quad & -5x_1^2 + x_2^2 \\ \text{subject to} \quad & x_1 = 1. \end{aligned}$$

has a solution $(1,0)^{\mathrm{T}}$. The quadratic penalty function

$$Q(x;\mu) = -5x_1^2 + x_2^2 + \frac{\mu}{2}(x_1 - 1)^2$$

is unbounded for $\mu < 10$. The iterates would diverge. Unfortunately, a common problem.

Ill-conditioning of Hessian

Newton step: $\nabla^2_{xx} Q(x;\mu_k) p_n = -\nabla_x Q(x;\mu_k)$

$$\nabla^2_{xx} Q(x;\mu_k) = \nabla^2 f(x) + \sum_{i=1}^{p} \underbrace{\mu_k h_i(x)}_{\approx \nu_i^\star} \nabla^2 h_i(x) + \mu_k \underbrace{\nabla h(x)}_{=:A(x)^{\mathrm{T}}} \nabla h(x)^{\mathrm{T}}.$$

If x is sufficiently close to the minimiser of $Q(\cdot;\mu_k)$

$$\nabla^2_{xx} Q(x;\mu_k) \approx \nabla^2_{xx} \mathcal{L}(x;\nu^\star) + \mu_k A(x)^{\mathrm{T}} A(x).$$

As $\mu_k \to \infty$ the Hessian is dominated by the second term (with eigenvalues 0 and $\mathcal{O}(\mu_k)$) and hence increasingly ill-conditioned.

Alternative formulation avoids ill-conditioning, $\zeta = \mu_k A(x) p_n$

$$\begin{bmatrix} \nabla^2 f(x) + \sum_{i=1}^{p} \mu_k h_i(x) \nabla^2 h_i(x) & A(x)^{\mathrm{T}} \\ A(x) & \mu_k^{-1} I \end{bmatrix} \begin{bmatrix} p_n \\ \zeta \end{bmatrix} = \begin{bmatrix} -\nabla_x Q(x;\mu_k) \\ 0 \end{bmatrix}.$$

Still, if $\mu_k h_i(x)$ is not a good enough approximation to $\nu^\star$, inadequate quadratic model yields inadequate search direction p_n.

General constraint problem

For general constraint problems including equality and inequality constraints, the quadratic penalty function can be defined as

$$Q(x;\mu) := f(x) + \frac{\mu}{2}\sum_{i=1}^{p} h_i^2(x) + \frac{\mu}{2}\sum_{i=1}^{m}\left([f_i(x)]^+\right)^2,$$

where $[y]^+ := \max\{y, 0\}$

Note: Q may be less smooth than the objective and constraint functions e.g. $f_1(x) = x_1 \geq 0$, then $\max\{y, 0\}^2$ has discontinuous second derivate and so does Q.

Practical penalty methods

- μ_k can be chosen adaptively based on the difficulty of minimising the penalty function in each iteration i.e. when minimising $Q(x; \mu_k)$ is expensive, choose μ_{k+1} moderately larger than μ_k e.g. $\mu_{k+1} = 1.5\mu_k$, when minimising $Q(x; \mu_k)$ is cheap, choose μ_{k+1} larger e.g. $\mu_{k+1} = 10\mu_k$.

- There is no guarantee that $\|\nabla_x Q(x; \mu_k)\| \leq \tau_k$ will be satisfied. Practical implementations need safe guards to increase μ (and possibly restore the initial point) when constraint violation is not decreasing fast enough or when the iterates appear diverging.

- Choice of initial point e.g. warm start $x^s_{k+1} = x_k$ can improve performance of Newton.

Nonsmooth penalty functions

Some penalty functions are *exact* i.e. for certain choices of penalty parameters a single minimisation w.r.t. x yields the exact minimiser of f. Only nonsmooth penalty functions can be exact.

An example is ℓ_1 penalty

$$Q_1(x;\mu) := f(x) + \mu \sum_{i=1}^{p} |h_i(x)| + \mu \sum_{i=1}^{m} [f_i(x)]^+,$$

where $[y]^+ := \max\{y, 0\}$.

Framework: Adaptively estimate the threshold value μ (in the same manner as for quadratic penalty checking the feasibility $h(x_k) \leq \tau$ for a set tolerance τ). Once the threshold value μ is reached the Q_1 penalty is exact (in the sense of the next slide).

Exactness of non-smooth penalty functions

minimiser of (COP) $\Rightarrow$ **minimiser of** Q_1**:**
Let $x^\star$ be a strict local minimiser of (COP), which satisfies the 1st order necessary conditions with Lagrange multipliers $\nu^\star, \lambda^\star$. Then $x^\star$ is a local minimiser of $Q_1(x;\mu)$ for all $\mu > \mu^\star = \|(\nu^\star, \lambda^\star)^{\mathrm{T}}\|_\infty$.
If moreover, the 2nd order sufficient conditions hold at $\mu > \mu^\star$, then $x^\star$ is a strict local minimiser of $Q_1(x;\mu)$.

stationary point of $Q_1 \Rightarrow$ **KKT point of** (COP) **or infeasible stationary point:**
Let $\hat{x}$ be a stationary point of the penalty function $Q_1(x;\mu)$ for all $\mu > \hat{\mu} > 0$. Then, if $\hat{x}$ is feasible for (COP), it satisfies KKT conditions. If $\hat{x}$ is not feasible for (COP), it is an infeasible stationary point.

1d example of general constraints (threshold $\mu^\star$)

$$\min x \quad s.t. \quad x \geq 1$$

with solution $x^\star = 1$.

$$Q_1(x;\mu) = x + \mu[1-x]^+ = \begin{cases} x & x \geq 1 \\ x + \mu(1-x) & x < 1 \end{cases}$$

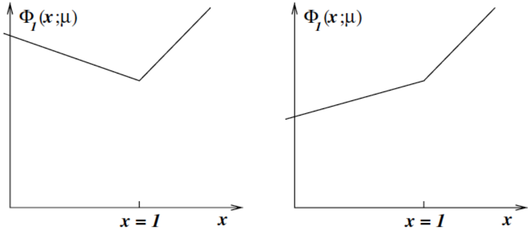


Figure: Fig. 17.3 from Nocedal, Wright: (left) $\mu > 1$, $x^\star$ minimises Q_1, (right) $\mu < 1$, Q_1 is unbounded.

Example revisited

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 - 2 = 0. \end{aligned}$$

Solution: $(-1,-1)^{\mathrm{T}}$.

ℓ_1 penalty function: $Q_1(x;\mu) = x_1 + x_2 + \mu|x_1^2 + x_2^2 - 2|$.

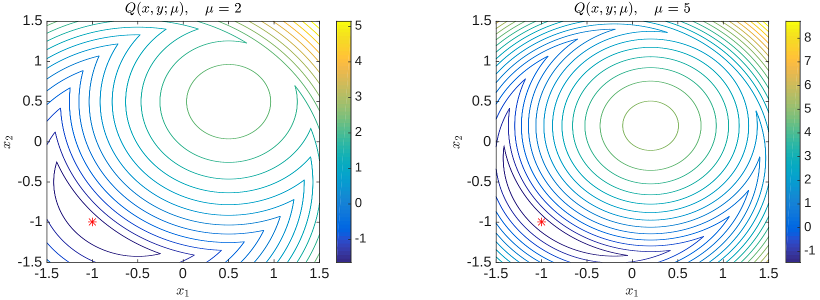


Figure: Minimiser of $Q_1(x;\mu)$ coincides with $x^\star$ for all $\mu = |\nu^\star| > 1/2$

Augmented Lagrangian

Reduces ill-conditioning by introducing explicit Lagrange multiplier estimates into the function to be minimised.
Can preserve smoothness. Can be implemented using standard unconstrained (or bound constrained) optimization.

Motivation: The minimisers x_k of $Q(x; \mu_k)$ do not quite satisfy the feasibility condition $h_i(x) = 0$

$$h_i(x_k) \approx \nu^\star / \mu_k, \quad i = 1, \dots, p.$$

Obviously, in the limit $\mu_k \to \infty$, $h_i(x) \to 0$ but can we avoid this systematic perturbation for moderate values of μ_k?

Augmented Lagrangian:

$$\mathcal{L}_A(x, \nu; \mu) := f(x) + \sum_{i=1}^{p} \nu_i h_i(x) + \frac{\mu}{2} \sum_{i=1}^{p} h_i^2(x).$$

Update of Lagrange multiplier estimate

Optimality condition for the unconstraint minimiser of $\mathcal{L}_A(x, \nu^k; \mu_k)$

$$0 \approx \nabla_x \mathcal{L}_A(x_k, \nu^k; \mu_k) = \nabla f(x_k) + \sum_{i=1}^{p} [\nu_i^k + \mu_k h_i(x_k)] \nabla h_i(x_k).$$

Optimality condition for the Lagrangian of (COP:E)

$$0 \approx \nabla_x \mathcal{L}(x_k, \nu^\star) = \nabla f(x_k) + \sum_{i=1}^{p} \nu_i^\star \nabla h_i(x_k).$$

Comparison yields (an update scheme for ν):

$$\nu_i^\star \approx \nu_i^k + \mu_k h_i(x_k), \quad i = 1, \dots, p$$

as from $h_i(x_k) = \frac{1}{\mu_k}(\nu_i^\star - \nu_i^k),\ i = 1, \dots p$ we see that if ν^k is close to $\nu^\star$ the infeasibility goes to 0 faster than $1/\mu_k$.

$$\begin{aligned} \min \quad & x_1 + x_2 \\ \text{subject to} \quad & x_1^2 + x_2^2 - 2 = 0. \end{aligned}$$

Solution: $(-1,-1)^{\mathrm{T}}$.

Augmented Lagrangian:
$\mathcal{L}(x,\nu;\mu) = x_1 + x_2 + \nu(x_1^2 + x_2^2 - 2) + \frac{\mu}{2}(x_1^2 + x_2^2 - 2)^2.$

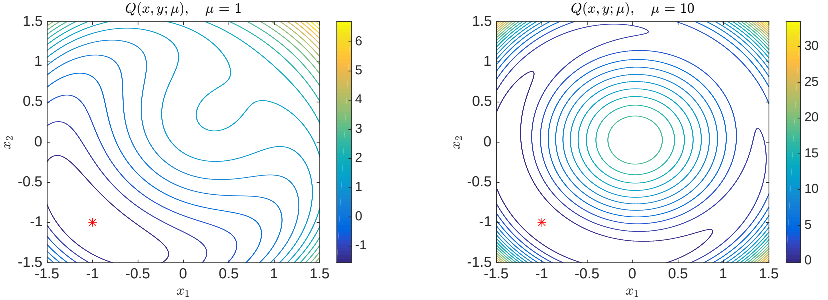


Figure: $\nu = 0.4$

Convergence

Let $x^\star$ be a local minimiser of (COP:E) at which the constraint gradients are linearly independent and which satisfies the 2nd order sufficient conditions with Lagrange multipliers $\nu^\star$. Then for all $\mu \geq \bar{\mu} > 0$, $x^\star$ is a strict local minimiser of $\mathcal{L}_A(x, \nu^\star; \mu)$. Furthermore, there exist $\delta, \epsilon, M > 0$ such that for all ν^k, μ_k satisfying

$$\|\nu^k - \nu^\star\| \leq \mu_k \delta, \quad \mu_k \geq \bar{\mu},$$

- the problem $\min \mathcal{L}_A(x, \nu^k; \mu_k)$, subject to $\|x - x^\star\| \leq \epsilon$, has a unique solution x_k and it holds

$$\|x_k - x^\star\| \leq M\|\nu^k - \nu^\star\|/\mu_k$$

- it holds

$$\|\nu^{k+1} - \nu^\star\| \leq M\|\nu^k - \nu^\star\|/\mu_k,$$

  where $\nu^{k+1} = \nu^k + \mu_k h(x_k)$.
- the matrix $\nabla^2_{xx}\mathcal{L}_A(x_k, \nu^k; \mu_k)$ is positive definite and the constraint gradients $\nabla h_i(x_k), i = 1, \ldots, p$ are linearly independent.

- **Bound constraint formulation:** convert inequality constraints into equality constraints using slack variables

$$f_i(x) - s_i = 0, \quad s_i \leq 0, \quad i \in \{1, \ldots .m\}.$$

Bound constraints are not transformed. Solve by projected gradient algorithm

$$x_{k+1} = P(x_k - \nabla_x \mathcal{L}_A(x, \nu; \mu)|_{x_k}; l, u) = 0,$$

where $P(\cdot; l, u)$ projects on the box $[l, u]$.

See Algorithm 17.4 in Nocedal Wright for an implementation.

- **Linearly constraint formulation:** transform into equality constraint problem with linearised constraints

  min $F_k(x)$, subject to $f_i(x_k)+\nabla f_i^{\mathrm{T}}(x_k)(x-x_k) = 0$, $l \leq x \leq u$.

  At each iteration k, choose F_k as

  $$F_k(x) = f(x) + \sum_{i=1}^{m} \nu_i^k \bar{f}_i^k(x),$$

  explicitly including the higher order constraint violations

  $$\bar{f}_i^k(x) = f_i(x) - f_i(x_k) - \nabla f_i(x_k)^{\mathrm{T}}(x - x_k).$$

  Preferred choice (larger convergence radius in practise)

  $$F_k(x) = f(x) + \sum_{i=1}^{m} \nu_i^k \bar{f}_i^k(x) + \frac{\mu}{2} \sum_{i=1}^{m} (\bar{f}_i^k(x))^2$$

- **Unconstraint formulation:** essentially extension of augmented Lagrangian to inequality constraints (proximal term instead of quadratic penalty).