The Augmented Lagrangian Method

Lecture 14, Continuous Optimisation
Oxford University Computing Laboratory, HT 2006
Notes by Dr Raphael Hauser (hauser@comlab.ox.ac.uk)

In Lecture 13 we saw that the quadratic penalty method has the disadvantage that the penalty parameter μ has to be reduced to very small values before x_k becomes feasible to high accuracy.

Moreover, we pointed out that reducing μ to very small values can lead to numerical instabilities if the method is not implemented very carefully.

We will now see a related method that does not require μ_k to converge to zero, and yet in a neighbourhood of a KKT point x^* of the nonlinear optimisation problem

$$\text{(NLP)} \qquad \begin{aligned} \min_{x\in\mathbb{R}^n} \ & f(x) \\ \text{s.t.} \ & g_{\mathcal{E}}(x) = 0 \\ & g_{\mathcal{I}}(x) \geq 0, \end{aligned}$$

the iterates x_k still converge to x^* if the LICQ and the second order sufficient optimality conditions hold at this point. In fact, μ can even be held constant after a while and the convergence of x_k continues!

Motivation:

The method is motivated by the observation that if we knew the Lagrange multipliers λ^* such that (x^*, λ^*) is a KKT point for (NLP), then we could find x^* by solving the unconstrained problem

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda^*). \tag{1}$$

Indeed, as already remarked in Lemma 1.2 i) of Lecture 12, the first set of KKT conditions $\nabla_x \mathcal{L}(x^*, \lambda^*) = 0$ amount to the first order necessary optimality conditions for (1).

Of course, λ^* is not known, but we know from Lecture 13 that one can obtain estimates $\lambda^{[k]}$ which can be used to set up the problem

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x,\lambda^{[k]}).$$

as an approximation of (1).

If the estimates $\lambda^{[k]}$ can be iteratively improved and made to converge to λ^*, then this can form the basis of an algorithmic framework for solving (NLP).

The Merit Function:

The merit function used by this algorithm is the *augmented Lagrangian* of (NLP), defined as follows,

$$\begin{aligned}\mathcal{L}_A(x,\lambda,\mu) &= \mathcal{L}(x,\lambda) + \frac{1}{2\mu}\sum_{i\in\mathcal{I}\cup\mathcal{E}} \tilde{g}_i^2(x)\\ &= f(x) - \sum_{i\in\mathcal{I}\cup\mathcal{E}} \lambda_i g_i(x) + \sum_{i\in\mathcal{I}\cup\mathcal{E}} \frac{\tilde{g}_i(x)}{2\mu} g_i(x)\\ &= f(x) + \sum_{i\in\mathcal{I}\cup\mathcal{E}} \left(\frac{\tilde{g}_i(x)}{2\mu} - \lambda_i\right) g_i(x),\end{aligned}$$

where $\tilde{g}_i$ is defined as in Lecture 13,

$$\tilde{g}_i(x) = \begin{cases} g_i(x) & (i\in\mathcal{E})\\ \min(g_i(x),0) & (i\in\mathcal{I}).\end{cases}$$

Algorithm: Augmented Lagrangian Method (AL)

S0 Initialisation: choose the following,

$x_0 \in \mathbb{R}^n$ (starting point, not necessarily feasible)

$\lambda^{[0]} \in \mathbb{R}^{|\mathcal{E}\cup\mathcal{I}|}$ (initial "guestimate" of Lagrange multiplier vector)

$\mu_0 > 0$ (initial value of homotopy parameter)

$(\tau_k)_{\mathbb{N}_0} \searrow 0$ (error tolerance)

S1 For $k = 0,1,2,\dots$ repeat

$y^{[0]} := x_k$, $l := 0$

until $\|\nabla_x \mathcal{L}_A(y^{[l]},\lambda^{[k]},\mu_k)\| \le \tau_k$ repeat

compute $y^{[l+1]}$ such that $\mathcal{L}_A(y^{[l+1]},\lambda^{[k]},\mu_k) < \mathcal{L}_A(y^{[l]},\lambda^{[k]},\mu$

(using unconstrained minimisation method)

$l \leftarrow l+1$

end

$x_{k+1} := y^{[l]}$

$\lambda_i^{[k+1]} := \lambda_i^{[k]} - \frac{\tilde{g}_i(x_{k+1})}{\mu_k}, \qquad (i \in \mathcal{E} \cup \mathcal{I}),$

$\lambda_i^{[k+1]} \leftarrow \max(0, \lambda_i^{[k+1]}), \qquad (i \in \mathcal{I})$

choose $\mu_{k+1} \in (0, \mu_k)$

end

A quick argument gives insight into why this method can be expected to converge before μ_k reaches very small values:

- We have

$$\nabla_x \mathcal{L}_A(x_{k+1}, \lambda^{[k]}, \mu_k) = \nabla f(x_{k+1}) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \Big(\lambda_i^{[k]} - \frac{\tilde{g}_i(x_{k+1})}{\mu_k}\Big) \nabla g_i(x_{k+1}).$$

- Using $\|\nabla_x \mathcal{L}_A(x_{k+1}, \lambda^{[k]}, \mu_k)\| \leq \tau_k$, we find

$$\sum_i \Big(\lambda_i^{[k]} - \frac{\tilde{g}_i(x_{k+1})}{\mu_k}\Big) \nabla g_i(x_{k+1}) = \nabla f(x_{k+1}) + O(\tau_k).$$

- By arguments similar to those in Theorem 2.2 in Lecture 13,

$$\lambda_i^{[k]} - \frac{\tilde{g}_i(x_{k+1})}{\mu_k} \simeq \lambda_i^*, \qquad (i \in \mathcal{E} \cup \mathcal{I}).$$

- Therefore, we have

$$\tilde{g}_i(x_{k+1}) \simeq \mu_k \Big(\lambda_i^{[k]} - \lambda_i^*\Big), \qquad (i \in \mathcal{E} \cup \mathcal{I}),$$

  which suggests that if $\lambda^{[k]} \to \lambda^*$ then all constraint residuals converge to zero like a function $o(\mu_k)$, where

$$\lim_{\mu \to 0} \frac{o(\mu)}{\mu} = 0.$$

  That is, the convergence is much faster than the $O(\mu_k)$ convergence obtained in the quadratic penalty function method.

Theorem 1: Let x^* be a local minimiser of (NLP) where the LICQ and the first and second order sufficient optimality conditions are satisfied for some Lagrange multiplier vector λ^*. Then there exists a constant $\bar{\mu} > 0$ such that x^* is a strict local minimiser of

$$\min_{x \in \mathbb{R}^n} \mathcal{L}_A(x, \lambda^*, \mu)$$

for all $\mu \in (0, \bar{\mu}]$.

Theorem 2: For (x^*, λ^*) and $\bar{\mu}$ as in Theorem 1 there exist constants $M, \varepsilon, \delta > 0$ such that the following is true:

i) If $\mu_k \leq \bar{\mu}$ and

$$\|\lambda^{[k]} - \lambda^*\| \leq \frac{\delta}{\mu_k}, \tag{2}$$

then the constrained minimisation problem

$$\begin{aligned} &\min_x \mathcal{L}_A(x, \lambda^{[k]}, \mu_k) \\ &\text{s.t. } \|x^* - x\| \leq \varepsilon \end{aligned} \tag{3}$$

has a unique minimiser x_{k+1},

and furthermore,

$$\|x^* - x_{k+1}\| \leq M\mu_k \|\lambda^{[k]} - \lambda^*\|, \tag{4}$$

ii) if μ_k and $\lambda^{[k]}$ are as in part i) and if $\lambda^{[k+1]}$ is chosen as in Algorithm (AL), then

$$\|\lambda^{[k+1]} - \lambda^*\| \leq M\mu_k \|\lambda^{[k]} - \lambda^*\|. \tag{5}$$

Some remarks about this result:

- (3) suggests the use of a trust-region method in the inner loop of Algorithm (AL).

- Without loss of generality, we may assume that $\bar{\mu} \leq (2M)^{-1}$. Note that if $(\lambda^{[k]}, \mu_k)$ satisfy the conditions of part i) of the theorem,

$$\begin{aligned} &\text{I)} \qquad \mu_k \leq \bar{\mu}, \\ &\text{II)} \qquad \|\lambda^{[k]} - \lambda^*\| \leq \frac{\delta}{\mu_k}, \end{aligned}$$

  and if it is also the case that

$$\text{III)} \qquad x_k \in B_\varepsilon(x^*),$$

  then x_k is a feasible starting point for the constrained problem

$$\begin{aligned} &\min_x \mathcal{L}_A(x, \lambda^{[k]}, \mu_k) \\ &\text{s.t. } \|x^* - x\| \leq \varepsilon. \end{aligned}$$

  Furthermore, we have

$$\begin{aligned} &\text{I')} \qquad \mu_{k+1} \leq \mu_k \overset{\text{I)}}{\leq} \bar{\mu}, \\ &\text{II')} \qquad \|\lambda^{[k+1]} - \lambda^*\| \overset{\text{II)},(5)}{\leq} M\mu_k \frac{\delta}{\mu_k} = \delta M < \frac{\delta}{\bar{\mu}} \overset{\text{I')}}{\leq} \frac{\delta}{\mu_{k+1}}, \\ &\text{III')} \qquad x_{k+1} \in B_\varepsilon(x^*). \end{aligned}$$

  Hence, by induction the relations I), II) and III) hold at every subsequent iteration j and the assumptions of part i) remain valid.

- Let k_0 be the iteration where (4) and (5) first hold,

$$\|x^* - x_{k+1}\| \leq M\mu_k \|\lambda^{[k]} - \lambda^*\|,$$
$$\|\lambda^{[k+1]} - \lambda^*\| \leq M\mu_k \|\lambda^{[k]} - \lambda^*\|.$$

  Then induction on k shows that

$$\|\lambda^{[k]} - \lambda^*\|, \|x_k - x^*\| \leq (M\bar{\mu})^{k-k_0} \|\lambda^{[k_0]} - \lambda^*\| \leq \frac{1}{2^{k-k_0}} \|\lambda^{[k_0]} - \lambda^*\|.$$

  Therefore, $x_k \to x^*$ and $\lambda^{[k]} \to \lambda^*$ at a Q-linear rate if $\mu \leq \bar{\mu}$ is held fixed.

Reading Assignment: Lecture-Note 14.

Recommended Additional Reading: Section 17.4, Nocedal–Wright.