SECTION C: CONTINUOUS OPTIMISATION
LECTURE 14: THE AUGMENTED LAGRANGIAN METHOD

HONOUR SCHOOL OF MATHEMATICS, OXFORD UNIVERSITY
HILARY TERM 2006, DR RAPHAEL HAUSER
WITH A FEW ADDITIONS FROM DR. NICK GOULD

1. The Augmented Lagrangian Method. 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

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

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!

1.1. 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

$$\min_{x\in\mathbb{R}^n} \mathcal{L}(x, \lambda^*). \tag{1.1}$$

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 (1.1).

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 (1.1).

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).

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

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

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

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

$\mathcal{L}_A$ is thus nothing else but the Lagrangian "augmented" by the quadratic penalty term introduced in Lecture 13, ensuring that x becomes gradually more feasible as the homotopy parameter μ is reduced.

1.3. The Algorithm.

ALGORITHM 1.1 (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 $\|\nabla_x \mathcal{L}_A(y^{[l]}, \lambda^{[k]}, \mu_k)\| \leq \tau_k$ *repeat*
  compute $y^{[l+1]}$ *such that* $\mathcal{L}_A(y^{[l+1]}, \lambda^{[k]}, \mu_k) < \mathcal{L}_A(y^{[l]}, \lambda^{[k]}, \mu_k)$
   (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

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

Arguments similar to those given in the proof of Theorem 2.2 in Lecture 13 show that

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

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.

This argument can be made precise in a neighbourhood of a point at which the sufficient second order optimality conditions hold. In fact, the following theorem indicates that μ does not have to be reduced to zero at all.

Theorem 1.2. *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 $\mu \in (0, \bar{\mu}]$.

For a proof see e.g. Nocedal–Wright, Theorem 17.5. Furthermore, this theorem can be strengthened to show that if $(x_k, \lambda^{[k]})$ ever enters a sufficiently small neighbourhood of (x^*, λ^*) and $\mu_k \leq \bar{\mu}$, then it is the case that $(x_k, \lambda^{[k]}) \to (x^*, \lambda^*)$ irrespective of whether μ_k is further decreased or not.

Theorem 1.3. *For (x^*, λ^*) and $\bar{\mu}$ as in Theorem 1.2 there exist constants $M, \varepsilon, \delta > 0$ such that the following is true:*

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

$$\|\lambda^{[k]} - \lambda^*\| \leq \frac{\delta}{\mu_k}, \tag{1.2}$$

then the constrained minimisation problem

$$\begin{aligned} &\min_x \mathcal{L}_A(x, \lambda^{[k]}, \mu_k) \\ &\text{s.t. } \|x^* - x\| \leq \varepsilon \end{aligned} \tag{1.3}$$

has a unique minimiser x_{k+1} and

$$\|x^* - x_{k+1}\| \leq M \mu_k \|\lambda^{[k]} - \lambda^*\|, \tag{1.4}$$

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

$$\|\lambda^{[k+1]} - \lambda^*\| \leq M \mu_k \|\lambda^{[k]} - \lambda^*\|. \tag{1.5}$$

We conclude with a few comments on why this result is interesting.

- 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 and if $x_k \in B_\varepsilon(x^*)$, then x_k is a good starting point for solving the problem (1.3) and we have

$$\begin{aligned} &x_{k+1} \in B_\varepsilon(x^*) \\ &\|\lambda^{[k+1]} - \lambda^*\| \overset{(1.2),(1.5)}{\leq} M\mu_k \frac{\delta}{\mu_k} = \delta M < \frac{\delta}{\bar{\mu}} \leq \frac{\delta}{\mu_{k+1}}, \end{aligned}$$

  where the last inequality follows from $\mu_{k+1} \leq \mu_k$. Thus, the same conditions hold again, and by induction they hold for all subsequent iterations.
- Let k_0 be the iteration where (1.4) and (1.5) first hold. Induction on k shows that

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

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

Additional Recommended Reading: Section 17.4, Nocedal–Wright.