

The augmented Lagrangian method for equality-constrained optimization

Mark S. Gockenbach

1 Introduction

The augmented Lagrangian method for solving

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0 \end{aligned} \tag{1}$$

is similar to the quadratic penalty approach. However, instead of adding the penalty term to the objective function f , it is added to the Lagrangian ℓ , resulting in the *augmented Lagrangian*

$$L(x; \lambda; \mu) = f(x) - \lambda \cdot g(x) + \frac{1}{2\mu} \|g(x)\|^2. \tag{2}$$

Since f and ℓ (for any λ) agree on the feasible set $g(x) = 0$, the basic idea is the same as before: a small value of μ forces the minimizer(s) of L to lie close to the feasible set, while, at the same time, values of x that reduce f are preferred. The advantage of the augmented Lagrangian approach is that by including an explicit estimate of the Lagrange multiplier, it is not necessary to decrease μ to zero in order to obtain convergence, and so various numerical problems are avoided.

I will assume that x^* is a local minimizer of (1) and λ^* is the corresponding Lagrange multiplier. If x_μ^* is a minimizer of L for a given μ and λ , then

$$\nabla L(x_\mu^*; \lambda; \mu) = \nabla f(x_\mu^*) - \nabla g(x_\mu^*)\lambda + \frac{1}{\mu} \nabla g(x_\mu^*)g(x_\mu^*) = 0,$$

that is,

$$\nabla f(x_\mu^*) = \nabla g(x_\mu^*) \left(\lambda - \frac{1}{\mu} g(x_\mu^*) \right).$$

Since

$$\nabla f(x^*) = \nabla g(x^*)\lambda^*,$$

it follows that

$$\lambda - \frac{1}{\mu} g(x_\mu^*)$$

should be a better estimate of λ^* than is λ . This suggests the following strategy:

Choose an initial estimate $\lambda^{(0)}$ of λ^* and some $\mu > 0$.

For $k = 1, 2, 3, \dots$

Define $x^{(k)}$ to be a minimizer of $L(\cdot; \lambda^{(k-1)}; \mu)$.

Define $\lambda^{(k)} = \lambda^{(k-1)} - (1/\mu)g(x^{(k)})$.

Under certain conditions that I specify below, it is possible to prove the following:

k	$x^{(k)}$	$ g(x_\mu^*) $	$\lambda^{(k)}$
1	(0.34798, 1.0326)	$1.8737 \cdot 10^{-1}$	-1.8737
2	(0.31710, 0.96303)	$2.7985 \cdot 10^{-2}$	-2.1535
3	(0.31249, 0.95237)	$4.6521 \cdot 10^{-3}$	-2.2000
4	(0.31173, 0.95059)	$7.8649 \cdot 10^{-4}$	-2.2079
5	(0.31160, 0.95028)	$1.3335 \cdot 10^{-4}$	-2.2093
6	(0.31158, 0.95023)	$2.2616 \cdot 10^{-5}$	-2.2095

Table 1: The results from Example 1.1.

1. If μ is sufficiently small, then $L(\cdot; \lambda^*; \mu)$ has a local minimizer at x^* and, for all λ sufficiently close to λ^* , $L(\cdot; \lambda; \mu)$ has a unique local minimizer in a neighborhood of x^* .
2. The sequence $(x^{(k)}, \lambda^{(k)})$ converges to (x^*, λ^*) as $k \rightarrow \infty$.

Before I develop the theory, I apply the augmented Lagrangian method to the following example, which I earlier treated by the quadratic penalty method.

Example 1.1 I define $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ by $f(x) = (x_1 - 1)^2 + 2(x_2 - 2)^2$ and $g(x) = x_1^2 + x_2^2 - 1$, respectively. The global minimizer of f , subject to the constraint $g(x) = 0$, and the corresponding Lagrange multiplier are

$$x^* \doteq \begin{bmatrix} 0.31157 \\ 0.95022 \end{bmatrix}, \quad \lambda^* \doteq -2.2095.$$

Choosing $\mu = 10^{-1}$ and the initial Lagrange multiplier estimate to be $\lambda^{(0)} = 0$, I obtained the results given in Table 1. The reader should notice that $x^{(k)} \rightarrow x^*$ and $\lambda^{(k)} \rightarrow \lambda^*$, in spite of the fact that μ is held constant at 10^{-1} .

2 Convergence analysis

The convergence analysis of the augmented Lagrangian method is similar to that of the quadratic penalty method, but significantly more complicated because there are two parameters λ, μ instead of just one. As a straightforward generalization of the previous method, I can define

$$F(x, \lambda_+; \lambda; \mu) = \begin{bmatrix} \nabla f(x) - \nabla g(x)\lambda_+ \\ -g(x) - \mu(\lambda_+ - \lambda) \end{bmatrix}$$

and solve for (x, λ_+) , regarding λ and μ as parameters. First of all, assuming as usual that x^*, λ^* is a local minimizer-Lagrange multiplier pair,

$$F(x^*, \lambda^*; \lambda^*; \mu) = \begin{bmatrix} \nabla f(x^*) - \nabla g(x^*)\lambda^* \\ -g(x^*) - \mu(\lambda^* - \lambda^*) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

for all $\mu > 0$. Moreover, the Jacobian of F (with respect to the variables x, λ_+), is

$$J(x, \lambda_+; \lambda; \mu) = \begin{bmatrix} \nabla^2 \ell(x; \lambda_+) & -\nabla g(x) \\ -\nabla g(x)^T & -\mu I \end{bmatrix}.$$

Assuming x^* is a nonsingular point of the NLP, the matrix

$$\begin{bmatrix} \nabla^2 \ell(x^*; \lambda^*) & -\nabla g(x^*) \\ -\nabla g(x^*)^T & 0 \end{bmatrix}$$

is nonsingular (as I showed previously when discussing the quadratic penalty function), and

$$J(x^*, \lambda^*; \lambda^*; \mu) = \begin{bmatrix} \nabla^2 \ell(x^*; \lambda^*) & -\nabla g(x^*) \\ -\nabla g(x^*)^T & -\mu I \end{bmatrix} \rightarrow \begin{bmatrix} \nabla^2 \ell(x^*; \lambda^*) & -\nabla g(x^*) \\ -\nabla g(x^*)^T & 0 \end{bmatrix}$$

as $\mu \rightarrow 0$. Therefore, there exists $\hat{\mu} > 0$ such that $J(x^*, \lambda^*; \lambda^*; \mu)$ is nonsingular for all $\mu \in [0, \hat{\mu}]$. The implicit function theorem¹ then implies that there exists a neighborhood N of λ^* such that there exist functions x, λ_+ , defined on $N \times [0, \hat{\mu}]$, such that

- $x(\lambda^*; \mu) = x^*$, $\lambda_+(\lambda^*; \mu) = \lambda^*$ for all $\mu \in [0, \hat{\mu}]$;
- for all $\lambda \in N, \mu \in [0, \hat{\mu}]$,

$$F(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) = 0.$$

The functions x, λ_+ satisfy

$$\nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu))\lambda_+(\lambda; \mu) = 0, \quad (3)$$

$$g(x(\lambda; \mu)) - \mu(\lambda_+(\lambda; \mu) - \lambda) = 0. \quad (4)$$

Solving (4) for $\lambda_+(\lambda; \mu)$ yields

$$\lambda_+(\lambda; \mu) = \lambda - \frac{1}{\mu}g(x(\lambda; \mu));$$

substituting this into (3) then produces

$$\nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu)) \left(\lambda - \frac{1}{\mu}g(x(\lambda; \mu)) \right) = 0.$$

Rearranging this last equation shows that

$$\nabla L(x(\lambda; \mu); \lambda; \mu) = 0;$$

in other words, $x(\lambda; \mu)$ is a stationary point of $L(\cdot; \lambda; \mu)$ for each $\lambda \in N$ and each $\mu \in [0, \hat{\mu}]$.

Since

$$\nabla^2 L(x(\lambda; \mu); \lambda; \mu) = \nabla^2 \ell(x(\lambda; \mu); \lambda_+(\lambda; \mu)) + \frac{1}{\mu} \nabla g(x) \nabla g(x)^T,$$

and $x(\lambda; \mu) \rightarrow x^*, \lambda_+(\lambda; \mu) \rightarrow \lambda^*$ as $\lambda \rightarrow \lambda^*$, it is straightforward to show that

$$\nabla^2 L(x(\lambda; \mu); \lambda; \mu)$$

is positive definite for λ sufficiently close to λ^* and for μ sufficiently small. (The proof is similar to the case of the quadratic penalty function.) I have therefore proved the following theorem.

Theorem 2.1 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are twice continuously differentiable and x^* is a local minimizer of the NLP*

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0. \end{aligned}$$

If x^ is a nonsingular point and λ^* is the corresponding Lagrange multiplier, then there exists $\hat{\mu} > 0$ and $\epsilon > 0$ and a function $x : N \times [0, \hat{\mu}] \rightarrow \mathbb{R}^n$, $N = B_\epsilon(\lambda^*)$, with the following properties:*

¹Actually, a different version of the implicit function theorem is needed here. The parameter μ varies over a compact set and it is possible to show that there is a neighborhood N of λ^* such that for all $\lambda \in N$ and for all $\mu \in [0, \hat{\mu}]$, there exist unique $x(\lambda; \mu), \lambda_+(\lambda; \mu)$ satisfying

$$F(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) = 0.$$

The ordinary implicit function theorem, as presented earlier, would give a possibly different neighborhood for each value of μ ; as stated, it would not guarantee that a single neighborhood works for all $\mu \in [0, \hat{\mu}]$. The improved version of the implicit function theorem can be found in Bertsekas [1], page 12.

1. x is continuously differentiable;
2. $x(\lambda^*; \mu) = x^*$ for all $\mu \in [0, \hat{\mu}]$;
3. $x(\lambda^*; \mu)$ is the unique local minimizer of $L(\cdot; \lambda; \mu)$ in N ;

According to the previous theorem, if μ is sufficiently small and $\lambda \rightarrow \lambda^*$, then $x(\lambda; \mu) \rightarrow x^*$. However, since λ^* is unknown, the condition $\lambda \rightarrow \lambda^*$ cannot be enforced directly. Instead, the augmented Lagrangian method updates λ using the results of the unconstrained minimization: $\lambda \leftarrow \lambda_+(\lambda; \mu)$. It is necessary to prove, then, that updating λ in this manner produces a sequence of Lagrange multiplier estimates converging to λ^* .

Since λ_+ is a continuously differentiable function of λ and $\lambda_+(\lambda^*; \mu) = \lambda^*$, I can write

$$\lambda_+(\lambda; \mu) = \lambda^* + \int_0^1 \nabla \lambda_+(\lambda^* + t(\lambda - \lambda^*); \mu)^T (\lambda - \lambda^*) dt$$

Using the triangle inequality for integrals, it follows that

$$\|\lambda_+(\lambda; \mu) - \lambda^*\| \leq \int_0^1 \|\nabla \lambda_+(\lambda^* + t(\lambda - \lambda^*); \mu)^T\| \|\lambda - \lambda^*\| dt \leq C(\mu) \|\lambda - \lambda^*\|,$$

where $C(\mu)$ is an upper bound for $\|\nabla \lambda_+(\cdot; \mu)^T\|$. Similarly,

$$x(\lambda; \mu) = x^* + \int_0^1 \nabla x(\lambda^* + t(\lambda - \lambda^*); \mu)^T (\lambda - \lambda^*) dt$$

and so

$$\|x(\lambda; \mu) - x^*\| \leq \int_0^1 \|\nabla x(\lambda^* + t(\lambda - \lambda^*); \mu)^T\| \|\lambda - \lambda^*\| dt \leq D(\mu) \|\lambda - \lambda^*\|,$$

where $D(\mu)$ is an upper bound for $\|\nabla x(\cdot; \mu)^T\|$.

The functions x, λ_+ are defined by the equations

$$\begin{aligned} \nabla f(x(\lambda; \mu)) - \nabla g(x(\lambda; \mu)) \lambda_+(\lambda; \mu) &= 0, \\ g(x(\lambda; \mu)) - \mu (\lambda_+(\lambda; \mu) - \lambda) &= 0. \end{aligned}$$

Differentiating these equations with respect to λ and simplifying the results yields

$$\nabla^2 \ell(x(\lambda; \mu); \lambda_+(\lambda; \mu)) \nabla x(\lambda; \mu)^T - \nabla g(x(\lambda; \mu)) \nabla \lambda_+(\lambda; \mu)^T = 0, \quad (5)$$

$$-\nabla g(x(\lambda; \mu))^T \nabla x(\lambda; \mu)^T - \mu \nabla \lambda_+(\lambda; \mu)^T + \mu I = 0, \quad (6)$$

or

$$J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) \begin{bmatrix} \nabla x(\lambda; \mu)^T \\ \nabla \lambda_+(\lambda; \mu)^T \end{bmatrix} = \begin{bmatrix} 0 \\ -\mu I \end{bmatrix}.$$

Since $J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu) \rightarrow J(x^*, \lambda^*; \lambda^*; \mu)$ as $\lambda \rightarrow \lambda^*$, it follows that

$$\|J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu)^{-1}\|$$

is bounded above for all λ sufficiently close to λ^* . Therefore, from

$$\begin{bmatrix} \nabla x(\lambda; \mu)^T \\ \nabla \lambda_+(\lambda; \mu)^T \end{bmatrix} = \mu J(x(\lambda; \mu), \lambda_+(\lambda; \mu); \lambda; \mu)^{-1} \begin{bmatrix} 0 \\ -I \end{bmatrix},$$

I can deduce that there exist $\hat{\mu} > 0$ and $M > 0$ such that, for all $\mu \in (0, \hat{\mu})$,

$$\|\nabla x(\lambda; \mu)^T\| \leq \mu M, \quad \|\nabla \lambda_+(\lambda; \mu)^T\| \leq \mu M.$$

Using μM in place of $C(\mu)$ and $D(\mu)$ above, I obtain

$$\begin{aligned}\|\lambda_+(\lambda; \mu) - \lambda^*\| &\leq \mu M \|\lambda - \lambda^*\|, \\ \|x(\lambda; \mu) - x^*\| &\leq \mu M \|\lambda - \lambda^*\|\end{aligned}$$

for all $\mu \in (0, \hat{\mu})$. I have therefore proved the following theorem:

Theorem 2.2 *Suppose $f : \mathbb{R}^n \rightarrow \mathbb{R}$ and $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ are twice continuously differentiable and x^* is a local minimizer of the NLP*

$$\begin{aligned}\min \quad & f(x) \\ \text{s.t.} \quad & g(x) = 0.\end{aligned}$$

If x^ is a nonsingular point and λ^* is the corresponding Lagrange multiplier, then there exist $\hat{\mu} > 0$ and $\epsilon > 0$ and a function $x : N \times [0, \hat{\mu}] \rightarrow \mathbb{R}^n$, $N = B_\epsilon(\lambda^*)$, with the following properties:*

1. *x is continuously differentiable;*
2. *$x(\lambda^*; \mu) = x^*$ for all $\mu \in [0, \hat{\mu}]$;*
3. *$x(\lambda^*; \mu)$ is the unique local minimizer of $L(\cdot; \lambda; \mu)$ in N ;*

Moreover, defining $\lambda_+(\lambda; \mu) = \lambda - g(x(\lambda; \mu))/\mu$ for $\mu \in (0, \hat{\mu})$, there exists a constant $M > 0$ such that

$$\|\lambda_+(\lambda; \mu) - \lambda^*\| \leq \mu M \|\lambda - \lambda^*\|, \tag{7}$$

$$\|x(\lambda; \mu) - x^*\| \leq \mu M \|\lambda - \lambda^*\| \tag{8}$$

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

By decreasing $\hat{\mu}$ if necessary, I can assume that $\mu M < 1$ for all $\mu \in [0, \hat{\mu}]$. Therefore, inequalities (7) and (8) that the augmented Lagrangian method produces sequences $\{x^{(k)}\}$ and $\{\lambda^{(k)}\}$ converging to x^* and λ^* , provided μ is sufficiently small (but fixed) and $\lambda^{(0)}$ is sufficiently close to λ^* .

References

- [1] Dimitri P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, San Diego, 1982.