The Augmented Lagrangian Method for Equality-Constrained Optimization

One of the most powerful general ideas for solving mathematics problems is to reduce a complicated problem to a problem that you already know how to solve. If we can recast a constrained optimization problem as an unconstrained problem, then we can use the BFGS method that we already have. We will employ this strategy for equality-constrained problems. Problems with inequality constraints can be recast so that all inequalities are merely bounds on variables, and then we will need to modify the method for equality-constrained problems. For now, we consider only problems of minimizing $f(\mathbf{x})$ subject to $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{g} \in \mathbb{R}^m$ with $m < n$.

One strategy for recasting a constrained problem as an unconstrained problem is to construct the Lagrangian function $\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda^T \mathbf{g}(\mathbf{x})$. We can then use the first-order necessary condition $\nabla(\mathcal{L}(\mathbf{x}_*, \lambda_*)) = \mathbf{0}$. This works if the problem is known to have only local minimizers. It does not work in general because the local minimizer of f is a saddle of $\mathcal{L}$.

A second strategy is to augment the objective function with a quadratic penalty term:

$$F_\rho(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\rho \sum_{i=1}^{m} g_i^2(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2}\rho\, \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x}).$$

The idea here is that the minimizer of F_ρ should converge to the desired solution as $\rho \to \infty$. This method has been used successfully, but it is problematic. Sometimes there is no minimizer if ρ is not large enough. More importantly, the function becomes ill-conditioned as $\rho \to \infty$, which poses a problem for the numerical solution. In practice, we should start with a modest value of ρ and then use the result as a starting iterate for a larger value of ρ.

The idea of the augmented Lagrangian method is to combine the Lagrangian formulation with a penalty function while considering only derivatives with respect to $\mathbf{x}$. This means that λ will be estimated and updated at each iteration. What makes the method work well is that the convergence of λ_k eliminates the need for $\rho \to \infty$.

In developing the augmented Lagrangian method, we need to do the following:

1. Identify the correct updating formula for λ_k;
2. Show that the iteration scheme converges without requiring $\rho \to \infty$ when $\lambda = \lambda_*$;
3. Show that the iteration scheme converges without requiring $\rho \to \infty$ when λ_k is updated using the formula of item 1.

Of these tasks, item 3 requires some technical analysis arguments and is not particularly instructive. Item 2 is instructive, particularly because it is not at all obvious that the minimizer of the unconstrained problem could be correct when the penalty parameter is finite.

The General Augmented Lagrangian Scheme

The augmented Lagrangian function for an equality-constrained problem is

$$F_\rho(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda^T \mathbf{g}(\mathbf{x}) + \frac{1}{2}\rho\, \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x}).$$

We fix λ) so that our unconstrained problem will have a local minimizer rather than a saddle, so the function we optimize in iteration k will be

$$\phi_k(\mathbf{x}) = F_{\rho_k}(\mathbf{x}, \lambda_\mathbf{k}) = f(\mathbf{x}) - \lambda_\mathbf{k}{}^T \mathbf{g}(\mathbf{x}) + \frac{1}{2}\rho_k\, \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x}). \tag{1}$$

Each iteration will use the BFGS algorithm to identify the approximate minimizer of ϕ_k, which will then become $\mathbf{x_{k+1}}$. We'll choose some sequence ρ_k and we'll need to update $\lambda_\mathbf{k}$. In the BFGS scheme, we'll use the final iterate $\mathbf{x_{k-1}}$ and Broyden matrix B_{k-1} from the previous iteration as the initial choices for $\mathbf{x_k}$ and B_k, with the identity matrix for the initial iteration.

From (1), we have the first-order necessary condition

$$0 = \nabla\phi(k)(\mathbf{x_{k+1}}) = \nabla f(\mathbf{x_{k+1}}) - (\nabla\mathbf{g}(\mathbf{x_{k+1}}))\lambda_\mathbf{k} + \rho_k(\nabla\mathbf{g}(\mathbf{x_{k+1}}))\mathbf{g}(\mathbf{x_{k+1}});$$

thus,

$$\nabla f(\mathbf{x_{k+1}}) = \nabla\mathbf{g}(\mathbf{x_{k+1}})[\lambda_\mathbf{k} - \rho_k\, \mathbf{g}(\mathbf{x_{k+1}})]. \tag{2}$$

We would like $\mathbf{x_{k+1}}$ to satisfy the Lagrange multiplier rule, which requires

$$\nabla f(\mathbf{x_{k+1}}) = \nabla\mathbf{g}(\mathbf{x_{k+1}})\lambda_\mathbf{k+1}. \tag{3}$$

Comparison of (2) and (3) indicates the correct update formula for $\lambda_\mathbf{k}$:

$$\lambda_\mathbf{k+1} = \lambda_\mathbf{k} - \rho_k\, \mathbf{g}(\mathbf{x_{k+1}}). \tag{4}$$

We will use this formula after we have determined $\mathbf{x_{k+1}}$.

An Example

Consider the problem of minimizing $f(x_1, x_2) = x_1 + x_2$ subject to the constraint $g(x_1, x_2) = x_1^2 + x_2^2 - 2 = 0$. This problem is easily solved by hand, with the result $\mathbf{x}_*{}^T = (1\ 1)$ and $\lambda_* = -1/2$. When we apply the augmented Lagrangian formulation, the first order necessary condition yields the equations

$$1 = 2\lambda_k x_1 - 2\rho_k x_1(x_1^2 + x_2^2 - 2), \qquad 1 = 2\lambda_k x_2 - 2\rho_k x_2(x_1^2 + x_2^2 - 2).$$

Multiplying the first by x_2 and the second by x_1 and subtracting yields the result $x_2 = x_1$, from which we obtain the equation

$$1 = 2\lambda_k x_1 - 4\rho_k x_1(x_1^2 - 1).$$

We cannot solve this equation exactly, but we can obtain an asymptotic approximation in the limit $\rho \to \infty$. This is a messy calculation, and the details are not crucial. What really matters is the value of the constraint function as the next iterate, which is

$$g(\mathbf{x_{k+1}}) \approx \frac{\lambda_k + \frac{1}{2}}{\rho_k}.$$

This means that we have two ways to work toward achieving the desired result of $g(\mathbf{x_{k+1}}) = 0$: by increasing ρ_k and by getting λ_k to converge to the correct value -1/2. More asymptotic work eventually yields the result

$$\lambda_{k+1} + \frac{1}{2} \approx \frac{\lambda_k + \frac{1}{2}}{8\rho_k}.$$

This is an excellent result. As we increase ρ, we get better approximations for λ. Both of these changes move the iterates toward feasibility. In contrast, the penalty method corresponds to taking $\lambda_k = 0$ for all k. In this case, the numerator of the approximation for g makes no contribution to the convergence.

This example helps show why the augmented Lagrangian method can be expected to converge without making ρ as large as is necessary for the penalty function method. The actual result is much better than this. Not only does the solution converge faster with $\lambda_k \to \lambda_*$ and $\rho \to \infty$, but $\rho \to \infty$ is not even necessary. Notice in the example that keeping ρ_k fixed still means that $\lambda_k + 1/2 \to 0$, which is enough to satisfy the constraint equation as $k \to \infty$. We'll need to prove that this is always the case; otherwise we can't have confidence in the method.

Convergence of the Solution if $\lambda_\mathbf{k} = \lambda_*$

We will now prove the following theorem:

Theorem 1 *Suppose $\mathbf{x}_*$ is a local minimizer for $f(\mathbf{x})$ subject to $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{g} \in R^m$ with $m < n$. Then $\mathbf{x}_*$ is a local minimizer for*

$$\psi_\rho(\mathbf{x}) = F_\rho(\mathbf{x}, \lambda_*) = f(\mathbf{x}) - {\lambda_*}^T \mathbf{g}(\mathbf{x}) + \frac{1}{2}\rho\, \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x})$$

for ρ sufficiently large.

To prove the theorem, we need to show that $\mathbf{x}_*$ satisfies the first-order necessary condition and the second-order sufficient condition. The first-order condition is easy. We can rewrite ψ_ρ in terms of the Lagrangian as

$$\psi_\rho(\mathbf{x}) = F_\rho(\mathbf{x}, \lambda_*) = \mathcal{L}(\mathbf{x}, \lambda_*) + \frac{1}{2}\rho\, \mathbf{g}^T(\mathbf{x})\mathbf{g}(\mathbf{x}). \tag{5}$$

Thus,

$$\nabla\psi_\rho(\mathbf{x}) = \nabla_x \mathcal{L}(\mathbf{x}, \lambda_*) + \rho\,(\nabla\mathbf{g}(\mathbf{x}))\mathbf{g}(\mathbf{x}). \tag{6}$$

The first term of $\nabla\psi_\rho(\mathbf{x}_*)$ vanishes because $\mathbf{x}_*$ is a local minimizer of the original problem, and the second term vanishes because $\mathbf{x}_*$ satisfies the constraints.

The second-order condition is much more difficult. It takes a fair bit of work to obtain the identity

$$H_\psi(\mathbf{x}_*) = H(\mathbf{x}_*) + \rho\, J^T(\mathbf{x}_*) J(\mathbf{x}_*), \tag{7}$$

where H is the Hessian of $\mathcal{L}$ and J is the Jacobian of $\frac{1}{2}\mathbf{g}^T\mathbf{g}$.