The Augmented Lagrangian Function and General Discussion on Lagrange Multiplier Methods

Toby Sanders

January 23, 2022

Contents

1 Part I: loosely tied together ideas

First, reviewing some concepts from the 1969 seminal paper "Multiplier and Gradient Methods," by Magnus Hestenes, which is essentially the first formal introduction of the augmented Lagrangian function. Second, discussed are some general concepts from Lagrange multiplier theory.

1.1 "Multiplier and Gradient Methods," by Magnus Hestenes

Problem: Find

$$\min_{x\in\mathbb{R}^N} f(x) \quad \text{subject to} \quad g(x) = 0.$$

Suppose x^* is the true minimizer and that $\nabla g(x^*) \neq 0$. Also suppose $f, g \in C^2$.

We may consider the method of Lagrange multipliers, which says there exist a constant λ such that

$$G = f + \lambda g$$

satisfies $\nabla G(x^*) = 0$, i.e. the two gradients are parallel at the minimizer. (To argue this, imagine walking along a the contour of $g = 0$, in which the gradient of g is always perpendicular to this contour. Furthermore, we have that $\nabla f(x^*) \cdot h = 0$, where h is a vector that points in the direction of the contour of $g = 0$, then ∇f is also perpendicular to the contour at x^*.). Moreover, (by the "second derivative rule"), for all $h \neq 0$ such that

$$\nabla g(x^*) \cdot h = 0,$$

we have

$$h^T H(G(x^*))h > 0,$$

where H denotes the Hessian. This implies that there exist a positive c such that

$$h^T H(G(x^*))h + c[\nabla g(x^*) \cdot h]^2 > 0,$$

for ALL $h \neq 0$. Setting

$$F = f + \lambda g + \frac{1}{2}cg^2 \tag{1}$$

we see that

$$\nabla F(x^*) = \nabla G(x^*) = 0,$$

$$h^T H(F(x^*))h = h^T H(G(x^*))h + c[\nabla g(x^*) \cdot h]^2 > 0, \quad h \neq 0.$$

In light of this, we see x^* is an unconstrained local minimum to F.

Therefore, minimizing (1), is the minimization of the so called *augmented Lagrangian Function*.

Finding λ

Consider minimization of the penalty function

$$f_n(x) = f(x) + \frac{1}{2}ng^2(x),$$

whose minimizers are given by the sequence $\{x_n\}$. A limit point, if it exists, is given by x^*. Moreover

$$0 = \nabla f_n(x_n) + ng(x_n)\nabla g(x_n),$$

and so if $\nabla g(x^*) \neq 0$, then $\lambda_n = ng(x_n)$ converges to λ.

Observe that

$$f_n(x_n) = f(x_n) + \frac{1}{2}ng^2(x_n) \leq f_n(x^*) = f(x^*).$$

and if x_n is sufficiently close to x^*, then (by the result in the previous section)

$$f(x^*) \leq f(x_n) + \lambda g(x_n) + \frac{1}{2}cg^2(x_n)$$

Combining the above two inequalities leads to

$$(n - c)g^2(x_n) \leq 2\lambda g(x_n)$$

If $\nabla f(x^*) \approx 0$, then λ is fairly small, and in general $x_n = x^*$ whenever $n > c$. In general this is not the case, and for large values of n the method becomes sensitive to round-off errors in the term ng^2. Therefore we use the augmented Lagrangian function (1).

Observe that for some n iteration guess of λ given by λ_n the minimizer x_n to $F(x, \lambda_n)$ satisfies

$$\nabla F(x_n, \lambda_n) = \nabla f(x_n) + (\lambda_n + c_n g(x_n))\nabla g(x_n).$$

By the ordinary Lagrange multiplier method, this suggests

$$\lambda_{n+1} = \lambda_n + c_n g(x_n).$$

Next we show how this is really just a minmax problem.

1.2 More discussion on LM methodology

Consider again the general problem of finding

$$\min_{x\in\mathbb{R}^N} f(x) \quad \text{subject to} \quad g(x) = 0,$$

for which there exists a constant λ so that x^* is a saddle point to the Lagrangian function

$$L(x, \lambda) = f(x) + \lambda g(x).$$

In other words, $\nabla\mathcal{L}(x^*, \lambda^*) = 0$. As it turns out, under certain conditions the optimal point x^* may be obtained through the dual problem

$$\max_{\lambda} \min_{x} \mathcal{L}(x, \lambda).$$

Define the dual function to be

$$v(\lambda) = \min_{x} \mathcal{L}(x, \lambda).$$

Observe that the dual function satisfies $v(\lambda) \leq f(x^*)$ for all λ, hence

$$f(x^*) = \max_{\lambda} v(\lambda).$$

Consider solving the minmax problem using an alternation gradient decent/ascent method. This iterative scheme looks like

$$\begin{aligned} x^{k+1} &= x^k - \tau\nabla_x\mathcal{L}(x^k, \lambda^k) \\ \lambda^{k+1} &= \lambda^k + \gamma g(x^{k+1}) \end{aligned}.$$

Notice the update on λ is the usual one...

A simple example

As a simple sanity check, I consider minimizing $f(x, y) = 2x + y$ over the constraint $h(x, y) = x^2 + y^2 - 1 = 0$. The minimizer is easily found to be $(x, y) = (-2/\sqrt{5}, -1/\sqrt{5})$. One may also find that the right multiplier in this case is $\lambda = \sqrt{5/4}$. As a numerical test, one may then finally check that these values satisfy the minmax problem.

1.3 Rigorous proof for first order condition for linear equality constraints

Consider the problem

$$\min_{x} f(x) \quad s.t. \quad Ax = b,$$

where $A \in \mathbb{R}^{m\times n}$, which has a local minimum at x^*. Let $\bar{x}$ be any particular solution to $Ax = b$. Then $x = \bar{x} + p$ also satisfies the equality constraint for $p \in Null(A)$. Let $Z \in \mathbb{R}^{n\times r}$ be a basis for the null space of A, so that $p = Zv$. Then we have the equivalent reduced minimization problem

$$\min_{v} \left(f(\bar{x} + Zv) = \phi(v)\right).$$

Then it can be shown that the first order necessary condition is given by

$$0 = \nabla \phi(v^*) = Z^T \nabla f(\bar{x} + Zv^*).$$

This is apparently call the reduced gradient or projected gradient.

Next consider decomposing the gradient at the optimizer as

$$\nabla f(x^*) = Zv^* + A^T \lambda_*.$$

Multiplying through by Z^T and using the projected gradient condition we observe that

$$0 = Z^T \nabla f(x^*) = Z^T Z v^* + Z^T A^T \lambda_* = Z^T Z v^* + 0 * \lambda_*.$$

This implies that $Zv^* = 0$, and hence returning back to decomposed gradient above we see that

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

which is the Lagrange multiplier condition in the linear equality case, i.e.

$$\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i \nabla g_i(x^*)$$

2 Part II: Proximal form of ADMM

Consider the following minimization problem:

$$\min_u f(u) + g(u), \tag{2}$$

which is equivalent to

$$\min_{u,v} f(u) + g(v), \;\; s.t. \; u = v. \tag{3}$$

The second form has the following augmented Lagrangian function

$$\mathcal{L}(u, v, \lambda) = f(u) + g(v) + \lambda^T (u - v) + \frac{\beta}{2} \|u - v\|_2^2, \tag{4}$$

which we find the minmax solution to as was shown earlier. To solve this we alternate minimization over u and v. Interestingly, this algorithm may be interpreted as an alternating proximal gradient algorithm. The proximal operator for an arbitrary function h and scalar $\gamma > 0$ is defined as

$$\text{prox}_{\gamma h}(x) = \arg\min_u \gamma h(u) + \frac{1}{2} \|u - x\|_2^2.$$

Define $\rho = \lambda / \beta$ and rewrite $\mathcal{L}$ as

$$\begin{aligned} \mathcal{L}(u, v, \lambda) &= f(u) + g(v) + \beta \rho^T (u - v) + \frac{\beta}{2} \|u - v\|_2^2 + \frac{\beta}{2} \|\rho\|_2^2 - \frac{\beta}{2} \|\rho\|_2^2 \\ &= f(u) + g(v) + \frac{\beta}{2} \|\rho + u - v\|_2^2 - \frac{\beta}{2} \|\rho\|_2^2 \end{aligned} \tag{5}$$

Then the minimization steps over u and v take the form

$$\begin{aligned}
&\arg\min_u \mathcal{L}(u, v, \lambda) \\
&= \arg\min_u f(u) + \frac{\beta}{2}\|u - (v - \rho)\|_2^2 \\
&= \operatorname{prox}_{f/\beta}(v - \rho) \\
&\arg\min_v \mathcal{L}(u, v, \lambda) \\
&= \arg\min_v g(v) + \frac{\beta}{2}\|v - (u + \rho)\|_2^2 \\
&= \operatorname{prox}_{g/\beta}(u + \rho)
\end{aligned} \tag{6}$$

This is sometimes called the scaled form of ADMM. The update on λ (and hence ρ) is the usual one:

$$\begin{aligned}
&\lambda \leftarrow \lambda + \beta(u - v) \\
&\rho \leftarrow \rho + (u - v)
\end{aligned} \tag{7}$$

3 Part III: More rigorous theory

Below are mostly concepts from Bertsekas book title "Nonlinear programming."

Problem 1:

$$\min_x f(x) \ s.t. \ h_i(x) = 0, \ i = 1, \dots, m, \tag{8}$$

whose local minimizer is x^*, and let $h = (h_1, \dots, h_m)$.

Let $\epsilon > 0$ be such that $f(x^*) \le f(x)$ for all feasible x with the ϵ neighborhood of x^*, i.e. if $h(x) = 0$ and $\|x - x^*\| \le \epsilon$, then $f(x^*) \le f(x)$.

Define $S := \{x \,|\, \|x - x^*\| \le \epsilon\}$,

$$F^k(x) = f(x) + \frac{k}{2}\|h(x)\|_2^2 + \frac{\alpha}{2}\|x - x^*\|_2^2,$$

and

$$x^k = \arg\min_{x \in S} F^k(x),$$

It is easy to see that $F^k(x^k)$ is bounded, since

$$F^k(x^k) = f(x^k) + \frac{k}{2}\|h(x^k)\|_2^2 + \frac{\alpha}{2}\|x^k - x^*\|_2^2 \le f(x^*), \tag{9}$$

and hence

$$\lim_{k \to \infty} \|h(x^k)\|_2^2 \to 0.$$

Therefore any limit point $\overline{x}$ of x^k satisfies $h(\overline{x}) = 0$. Furthermore, from (9) we have

$$f(x^k) + \frac{\alpha}{2}\|x^k - x^*\|_2^2 \le f(x^*),$$

and taking the limit as $k \to \infty$ we obtain

$$f(\overline{x}) + \frac{\alpha}{2}\|\overline{x} - x^*\|_2^2 \leq f(x^*).$$

Since $h(\overline{x}) = 0$ and $\overline{x} \in S$, then

$$f(x^*) \leq f(\overline{x}).$$

Combining the above two equations we see that $\overline{x} = x^*$, hence $x^k \to x^*$.

Proposition 1. *This is prop. 4.1.1. in Bertsekas book, which I abbreviate, and also exclude the second order conditions.*

There exist unique Lagrange multiplier vector λ^* such that

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0.$$

Proof. From the work above we have

$$0 = \nabla F^k(x^k) = \nabla f(x^k) + k\nabla h(x^k)h(x^k) + \alpha(x^k - x^*) \tag{10}$$

Solving for $kh(x^k)$ we obtain

$$kh(x^k) = -(\nabla h(x^k)^T \nabla h(x^k))^{-1} \nabla h(x^k)^T (\nabla f(x^k) + \alpha(x^k - x^*))$$

Taking the limit as $k \to \infty$ we obtain

$$\lambda^* = -(\nabla h(x^*)^T \nabla h(x^*))^{-1} \nabla h(x^*)^T \nabla f(x^*)$$

Taking the limit in (10) we obtain

$$\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0.$$

□

Problem 2 (ICP):

$$\begin{aligned} &\min_x f(x) \\ &s.t.\ h_i(x) = 0,\ i = 1, \ldots, m \\ &\quad\ g_j(x) \leq 0,\ j = 1, \ldots, r. \end{aligned} \tag{11}$$

Define the Lagrangian function by

$$\begin{aligned} \mathcal{L}(x, \lambda, \mu) &= f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) \\ &= f(x) + \lambda^T h(x) + \mu^T g(x) \end{aligned} \tag{12}$$

Proposition 2 (First order KKT conditions)**.** *Let x^* be a local minimum to ICP. Then there exists unique Lagrange multipliers λ^* and μ^* such that*

$$\begin{aligned} &\nabla_x \mathcal{L}(x^*, \lambda^*, \mu^*) = 0 \\ &\mu_j \geq 0, \quad \forall j \\ &\mu_j^* = 0, \quad \forall j \notin A(x^*), \end{aligned} \tag{13}$$

where $A(x^*)$ is the set of active constraints at x^*, i.e. $j \in A(x)$ if $g_j(x) = 0$.

Proof. All assertions follow from the previous proposition, except for $\mu_j^* \geq 0$ for $j \in A(x^*)$. Define $g_j^+(x) = max\{0, g_j(x)\}$. Define

$$F^k(x) = f(x) + \frac{k}{2}\|h(x)\|_2^2 + \frac{k}{2}\sum_{j=1}^{r}(g^+(x))^2 + \frac{\alpha}{2}\|x - x^*\|_2^2,$$

and consider solving

$$x^k = \arg\min_{x \in S} F^k(x),$$

where $S = \{x \mid \|x - x^*\| \leq \epsilon\}$, where ϵ is such that $f(x^*) \leq f(x)$ for all feasible $x \in S$. Note that the gradient of $(g_j^+(x))^2$ is $2g_j^+(x)\nabla g_j(x)$. A similar argument to the equality constrained case shows that $x^k \to x^*$ and the Lagrange multipliers are given by

$$\begin{aligned} \lambda_i^* &= \lim_{k \to \infty} kh_i(x^k), \quad \forall i \\ \mu_j^* &= \lim_{k \to \infty} kg_j^+(x^k), \quad \forall j. \end{aligned} \tag{14}$$

Since $g_j^+(x^k) \geq 0$ we obtain $\mu_j^* \geq 0$ for all j. □

Remark: The condition $\mu_j^* = 0$ for all $j \notin A(x^*)$ can be written compactly as

$$\mu_j^* g_j(x^*) = 0 \quad \forall j.$$

Duality: smooth convex f with linear constraints

Problem 3 (ICP) with linear constraints:

$$\begin{aligned} &\min_x f(x) \\ &s.t.\ Ex = d \\ &\qquad Ax \leq b, \end{aligned} \tag{15}$$

with f smooth and convex. The Lagrangian for this problem is of course

$$\mathcal{L}(x, \lambda, \mu) = f(x) + \lambda^T(Ex - d) + \mu^T(Ax - b).$$

Define

$$q(\lambda, \mu) = \inf_x \mathcal{L}(x, \lambda, \mu).$$

Then problem 3 has the equivalent dual problem:

$$\max_{\lambda, \mu} q(\lambda, \mu)\ s.t.\ \mu \geq 0.$$

Proof. For all feasible x and $\mu \geq 0$, we have $\lambda^T(Ex - d) = 0$ and $\mu^T(Ax - b) \leq 0$, so that

$$q(\lambda, \mu) \leq f(x) + \lambda^T(Ex - d) + \mu^T(Ax - b) \leq f(x).$$

Taking the minimum over all feasible x on the right hand side we obtain

$$q(\lambda, \mu) \leq f(x^*),$$

hence it suffices to show that for certain $\mu \geq 0, \lambda$, that $q(\lambda, \mu) = f(x^*)$. Using the KKT conditions/theorem, selecting the unique multipliers μ^* and λ^*, we see this is achieved. □