Numerical Optimization with applications
CHAPTER 17: Penalty and Augmented Lagrangian Methods

104021601 林俊傑
104021602 吳彥儒
104021615 黃翊軒

January 15, 2017

17.1 THE QUADRATIC PENALTY METHOD

Given the original constrained optimization problem

$$\min_x f(x) \quad \text{subject to } c_i(x) = 0, \quad i \in \mathcal{E}, \quad c_i(x) \geq 0, \quad i \in \mathcal{I}, \tag{1}$$

we could define the corresponding quadratic penalty function as

$$Q(x;\mu) := f(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x) + \frac{\mu}{2}\sum_{i\in\mathcal{I}} \left([c_i(x)]^-\right)^2, \tag{2}$$

where $\mu > 0$ and $[y]^-$ is the abbreviated symbol of $\max(-y, 0)$.

If we take a sequence $\mu_k \nearrow \infty$ into the quadratic penalty function, we could find that $Q(x;\mu_k)$ deverges if x is infeasible. The larger μ_k is, the severer constraint violations we panalize. As a result, the minimizer of the quadratic penalty function $Q(x;\mu_k)$ is closer to the feasible region as k increases.

Framework 17.1 (Quadratic Penalty Method).

Given $\mu_0 > 0$, a nonnegative sequence $\{\tau_k\}$ with $\tau_k \to 0$, and a starting point x_0^s;
for $k = 0, 1, 2, \ldots$
  Find an approximate minimizer x_k of $Q(\cdot;\mu_k)$, starting at x_k^s,
    and terminating when $\|\nabla_x Q(x;\mu_k)\| \leq \tau_k$;
  if final convergence test satisfied
    stop with approximate solution x_k;
  end (if)
  Choose new penalty parameter $\mu_{k+1} > \mu_k$;
  Choose new starting point x_{k+1}^s;
end (for)

We have two theorems to support the convergence of Framework 17.1.

Theorm 17.1 states that the global minimizer x_k of quadratic penalty function $Q(x;\mu_k)$ converges to the constrained optimization problem x, i.e. $x_k \to x$.

Theorm 17.2 states that if $\tau_k \to 0$ and x_k only satisfies

$$\|\nabla_x Q(x;\mu_k)\| \leq \tau_k,$$

then

$$x_k \to x^*,$$

where x^* is a stationary point of $\|c(x)\|^2$. Besides, if $\nabla c_i(x^*)$ is linearly independent, then

$$\lim_{k\to\infty} -\mu_k c_i(x_k) = \lambda_i^* \quad \forall i \in \mathcal{E}$$

and (X^*, λ^*) satisfy the KKT conditions.

Practical problems

Even if $\nabla^2 f(x^*)$ is well-conditioned, the Hessian $\nabla^2_{xx} Q(x;\mu_k)$ might become ill-conditioned as $\mu_k \to \infty$.

By defining

$$A(x)^T = (\nabla c_i(x))_{i\in\mathcal{E}}$$

and considering equality constraints only, we have

$$\nabla^2_{xx} Q(x;\mu_k) = \nabla^2 f(x) + \sum_{i\in\mathcal{E}} \mu_k c_i(x) \nabla^2 c_i(x) + \mu_k A(X)^T A(X).$$

From **Theorem 17.2**, we have

$$\mu_k c_i(x) \approx -\lambda_i^*$$

for x near a minimizer. Hence, we obtain

$$\nabla^2_{xx} Q(x;\mu_k) \approx \nabla^2_{xx} \mathcal{L}(x,\lambda^*) + \mu_k A(X)^T A(X).$$

We find that $\nabla^2_{xx} Q(x;\mu_k)$ have problems with ill-conditioning since the second term diverges as $\mu_k \to \infty$.

For Newton's method step

$$\nabla^2_{xx} Q(x;\mu_k) p = \nabla_x Q(x;\mu),$$

we can apply a reformulation

$$\begin{pmatrix} \nabla^2 f(x) + \sum_{i\in\mathcal{E}} \mu_k c_i(x) \nabla^2 c_i(x) & A(x)^T \\ A(x) & -(1/\mu_k) I \end{pmatrix} \begin{pmatrix} p \\ \mu A(x) p \end{pmatrix} = \begin{pmatrix} -\nabla_x Q(x;\mu_k) \\ 0 \end{pmatrix}$$

to avoid the ill-conditioning since p solves both systems. Note that this system has dimension $n + |\mathcal{E}|$ rather than n.

17.2 NONSMOOTH PENALTY FUNCTIONS

A penalty function ia called *exact* if, for certains coice of penalty parameters, the minimizer $x^\star$ is the exact solution of the original constrained optimization problem. Nevertheless, the quadratical penalty function is not exact. In this section, we introduce the *nonsmooth* penalty functions.
A popular nonsmooth penalty function is the l_1 *penalty function* defined by

$$\phi_1(x;\mu) = f(x) + \mu \sum_{i\in\mathcal{E}} |c_i(x)| + \mu \sum_{i\in\mathcal{I}} [c_i(x)]^-. \tag{3}$$

The next two theorems establish the *exactness* of (3).
Theorm 17.3 states that if $x^\star$ is a strictly local minimizer of (1), with Lagrange miltipliers $\lambda^\star$. Then $x^\star$ is a local minimizer of (3) $\quad \forall\ \mu > \mu^\star$, where

$$\mu^\star = \|\lambda^\star\|_\infty \tag{4}$$

Theorm 17.4 states that if $\hat{x}$ is a stationary points of $\phi_1(x;\mu)$ for all μ large enouth. Then, $\hat{x}$ is either satisfying KKT conditions for (1) or it is an infeasible stationary points.
Define the measure of infeasibility

$$h(x) = sum_{i\in\mathcal{E}} |c_i(x)| + sum_{i\in\mathcal{I}} [c_i(x)]^- \tag{5}$$

Then, we can develope an algorithm framwork via the l_1 penalty funtion.
Framework 17.2 (Classical ℓ_1 Penalty Method).

Given $\mu_0 > 0$, tolerance $\tau > 0$, starting point x_0^s;
for $k = 0, 1, 2, \ldots$
  Find an approximate minimizer x_k of $\phi_1(x;\mu_k)$, starting at x_k^s;
  if $h(x_k) \le \tau$
    stop with approximate solution x_k;
  end (if)
  Choose new penalty parameter $\mu_{k+1} > \mu_k$;
  Choose new starting point x_{k+1}^s;
end (for)

Since $\phi_1(x;\mu)$ is nonsmooth, the minimization will be difficule. However, we can transform $\phi_1(x;\mu)$ into a smooth model.

A PRATICAL l_1 PENALTY METHOD

As we did for the unconstrained optimization problem, we can transform (3) into a smooth model by replacing f by its Taylor expension and c_i by its linearization,as follows:

$$q(p;\mu) = f(x) + \triangledown f(x)^T p + \frac{1}{2}p^T W p + \mu \sum_{i\in\mathcal{E}} |c_i(x) + \triangledown c_i(x)^T p| + \mu \sum_{i\in\mathcal{I}} [c_i(x) + \triangledown c_i(x)^T p]^-$$

where W is an approximation of Hessian about f and c_i. The function $q(p;\mu)$ is still not smooth, but we can reformulate it into a smooth quadratic optimization problem by introducing some new variables, as follows:

$$\begin{aligned} \min_{p,r,s,t} \quad & f(x) + \frac{1}{2}p^T W p + \triangledown f(x)^T p + \mu \sum_{i\in\mathcal{E}} |r_i + s_i| + \mu \sum_{i\in\mathcal{I}} t_i \\ \textit{subject to} \quad & \triangledown c_i(x)^T p + c_i(x) = r_i - s_i, \quad i \in \mathcal{E} \\ & \triangledown c_i(x)^T p + c_i(x) \geq -t_i, \quad i \in \mathcal{I} \\ & r, s, t \geq 0 \end{aligned} \tag{6}$$

Even after adding a trust region constraint $\|p\|_\infty \leq \triangle$, (6) is still a quadratic problem. It can be solved by a quadratic programming solver.

A GENERAL CLASS OF NONSMOOTH PENALTY METHODS

Exact nonsmooth penalty funtions can use other norms.

$$\phi(x;\mu) = f(x) + \mu \|c_\mathcal{E}(x)\| + \mu \left\|[c_\mathcal{I}(x)]^-\right\| \tag{7}$$

Framework 17.2 can work on these penalty functions by simply redefinind the measure of infeasibility (5) as $h(x) = \|c_\mathcal{E}(x)\| + \|[c_\mathcal{I}(x)]^-\|$.
The properties garguaranteed by Theorem 17.3 and Theorem 17.4 can be extended to the general class (7).In Theorem 17.3, we replace $\mu^\star$ in (4) by

$$\mu^\star = \|\lambda^\star\|_D,$$

where $\|\bullet\|_D$ is the dual norm of $\|\bullet\|$. Theorem 17.4 applies without modification.

17.3 AUGMENTED LAGRANGIAN METHOD: EQUALITY CONSTRAINTS

In section 17.1, we know that even μ_k is large, the approximate minimizer x_k of the quadratic penalty function $Q(x;\mu_k)$ may be infeasible, the violation of $c_i(x) \approx -\lambda_i^*/\mu_k$. To make the approximate solution x_k closer to the feasible region, we introduce the Augmented Lagrangian function:

$$\mathcal{L}_A(x,\lambda;\mu) := f(x) - \sum_{i\in\mathcal{E}} \lambda_i c_i(x) + \frac{\mu}{2}\sum_{i\in\mathcal{E}} c_i^2(x) \tag{8}$$

Use the fact of Theorem 2.2 and (17.17), and rearranging the expression, we have $c_i(x_k) \approx -\frac{1}{\mu_k}(\lambda_i^* - \lambda_i^k)$, the violent of x_k is much smaller than $\frac{1}{\mu_k}$. We can set the Lagrangian multiplier vector of the next step $\lambda_i^{k+1} = \lambda_i^k - \mu_k c_i(x_k)$, for all $i \in \mathcal{E}$.

Framework 17.3 (Augmented Lagrangian Method-Equality Constraints).

Given $\mu_0 > 0$, tolerance $\tau_0 > 0$, starting points x_0^s and λ^0;
for $k = 0, 1, 2, \ldots$
  Find an approximate minimizer x_k of $\mathcal{L}_A(\cdot, \lambda^k; \mu_k)$, starting at x_k^s,
    and terminating when $\|\nabla_x \mathcal{L}_A(x_k, \lambda^k; \mu_k)\| \leq \tau_k$;
  if a convergence test for (17.1) is satisfied
    stop with approximate solution x_k;
  end (if)
  Update Lagrange multipliers using (17.39) to obtain λ^{k+1};
  Choose new penalty parameter $\mu_{k+1} \geq \mu_k$;
  Set starting point for the next iteration to $x_{k+1}^s = x_k$;
  Select tolerance τ_{k+1};
end (for)

Theorm 17.5 states that if we know the exact Lagrangian multiplier λ^*, then the solution of (1) is a strict minimizer of $\mathcal{L}_A(x, \lambda; \mu)$ for μ large enough. Even though we only have a "good" estimate of λ^*, we can still get a good estimate of x^* by minimizing $\mathcal{L}_A(x, \lambda; \mu)$ with large μ.
Theorm 17.6 states the advantage of the augmented Lagrangian method. Different from the quadratic penalty method, we can get a good approximation of x^* if λ_k is close to λ^* or if the penalty parameter μ_k is large. On the other hand, by (17.46), we can improve the accuracy of λ^* by choosing a large μ_k.

17.4 PRACTICAL AUGMENTED LAGRANGIAN METHOD

In section 17.3, we only discuss the problem with equality constrains. Now for the general case, there are three useful formulations.
Bound-Constrained Formulation
Use the slack variable s_i to turn inequalities into equalities. That is

$$c_i(x) - s_i = 0, \quad s_i \geq 0, \quad \forall i \in \mathcal{I}$$

We can reformulate the problem into

$$\min_{x \in \mathbb{R}^n} f(x) \quad s.t. \quad c_i(x) = 0, \quad i = 1, 2, ..., m, \quad l \leq x \leq u$$

The Bounded-constrained Lagrangian will be:

$$\min_x \mathcal{L}_A(x, \lambda; \mu) = f(x) - \sum_{i=1}^{m} \lambda_i c_i(x) + \frac{\mu}{2} \sum_{i=1}^{m} c_i^2(x) \quad s.t. \quad l \leq x \leq u$$

Solve this problem and update λ and μ repeatedly.
Linearly Constrained Formulation
LCL method is to solve the subproblem of minimizing the augmented Lagrangian function subject to linearization of the constrains.

$$\begin{aligned} \min_x \quad & F_k(x) \\ s.t. \quad & c(x_k) + A_k(x - x_k) = 0, \quad l \leq x \leq u. \end{aligned}$$

where

$$c_i^{-k}(x) = c_i(x) - c_i(x_k) - \triangledown c_i(x_k)^T (x - x_k).$$

Current Augmented Lagrangian funciton

$$F_k(x) = f(x) - \sum_{i=1}^{m} \lambda_i^k c_i^{-k}(x) + \frac{\mu}{2} \sum_{i=1}^{m} [c_i^{-k}(x)]^2$$

Unconstrained Formulation

Suppose the problem has no equality constrain, i.e. $\mathcal{E} = \emptyset$, then we can rewrite the problem as

$$\min_{x \text{ feasible}} f(x) = \min_{x \in \mathbb{R}^n} F(x)$$

where

$$F(x) = \max_{\lambda \geq 0} \{f(x) - \sum_{i \in \mathcal{I}} \lambda_i c_i(x)\}$$

Note that if x is feasible, $F(x) = f(x)$ and λ_i should be zero. Otherwise $F(x)$ turns to infinity, and λ_i can be chosen arbitrary large. Consequently, F is not smooth, so it is not practical to minimize directly. We replace F by a smooth approximated function

$$\widehat{F}(x; \lambda^k, \mu_k) = \max_{\lambda \geq 0} \{f(x) - \sum_{i \in \mathcal{I}} \lambda_i c_i(x) - \frac{1}{2\mu_k} \sum_{i \in \mathcal{I}} (\lambda_i - \lambda_i^k)^2$$

where the last term can enforce the mew maximizer λ close to the previous estimate λ^k.
By above, we can obtain the explicit maximization of λ. Then we have

$$\widehat{F}(x; \lambda^k, \mu_k) = f(x) + \sum_{i \in \mathcal{I}} \psi(c_i(x), \lambda_i^k; \mu_k)$$

where the functionψ is defined as

$$\psi(t, \sigma; \mu) := \begin{cases} -\sigma t + \frac{\mu}{2} t^2 & \text{if} t - \sigma/\mu \leq 0, \\ -\frac{1}{2\mu}\sigma^2 & \text{otherwise,} \end{cases}$$

Hence, we can obtain x_k by minimizing $\widehat{F}$, and update Lagrange multiplier estimates repeatedly.