Constrained Nonlinear Optimization

Maestría en Cómputo Estadístico

Centro de Investigación en Matemáticas A.C.

What is constrained nonlinear optimization? I

- Roughly speaking, it is a nonlinear optimization problem with either equallity or inequallity constrains
- Constrains can be linear or nonlinear
- Mathematically, a nonlinear constrained optimization problem is defined as:

$$\begin{aligned} \min_{\mathrm{x}} \quad & f(\mathrm{x}) \\ \text{subject to:} \quad & h_i(\mathrm{x}) = 0, i = 1, \ldots, m \\ & g_j(\mathrm{x}) \leq 0, j = 1, \ldots, l \end{aligned} \tag{1}$$

First-Order Optimality Conditions I

- A necessary condition to consider x^* as an optimal solution can be stated as follows:

$$\nabla f(x^*) - \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) - \sum_{j=1}^{l} \gamma_j \nabla g_j(x^*) = 0 \tag{2}$$

- This is known as the Karush-Kuhn-Tucker (KKT) condition

Second-Order Optimality Conditions I

- Let $\mathcal{L}(x, \gamma, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i h_i(x) - \sum_{j=1}^{l} \gamma_j g_j(x)$
- A second-order condition should satisfy:

$$d^T \nabla^2 \mathcal{L}(x, \gamma, \lambda)\, d \geq 0 \tag{3}$$

Quadratic Programming I

- Quadratic programming (QP) is the simplest constrained nonlinear optimization problem
- It is a special case with quadratic objective function and linear constraints
- Mathematically, a QP problem is defined as:

$$\begin{aligned} \min_{\mathrm{x}} \quad & \frac{1}{2}\mathrm{x}^T \mathrm{Q}\mathrm{x} + \mathrm{q}^T\mathrm{x} \\ \text{subject to:} \quad & \mathrm{a}_i^T\mathrm{x} = b_e, i = 1, \dots, m \\ & \mathrm{a}_j\mathrm{x} \geq b_i, j = m+1, \dots, m+l \end{aligned} \tag{4}$$

- where Q is a symmetric $n \times n$ matrix

Quadratic Programming II

- If matrix Q is positive semi-definite, then the optimization problem is convex; thus, any local minimum is a global minimum
- a typical technique to solve QP is through the conjugate gradient method
- Remarkable applications of QP into machine learning are the optimization problem to train support vector machines and least squares regression

Example I

Find the optimal solution for the following QP problem:

$$\begin{aligned} \min_{x} \quad & \frac{1}{2}x^T Q x + q^T x \\ \text{subject to:} \quad & A_e{}^T x = 0 \\ & A_i{}^T x \geq 1 \end{aligned} \tag{5}$$

where

- $x = [x_1, x_2]^T$
- $Q = \begin{bmatrix} 3 & -1 \\ -1 & 1 \end{bmatrix}$
- $q = [-2, 0]^T$
- $A_e = [1, -2]^T$
- $A_i = [1, -1]^T$

Equality Constrained QP

- When only equality constraints are considered, QP is reduced to an equality constrained QP
- Equality constrained QP can be reduced to the following problem:

$$\begin{bmatrix} Q & -A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -q \\ b_e \end{bmatrix} \tag{6}$$

Active Set Methods I

- Most QP problems involve inequality constraints
- Constrained QP can be converted to an equality constrained QP form and be solved with most effective methods
- A constraint is said to be active if satisfies the equality condition
- The active set is the set of all constraints that are active

Active Set Methods II

- Intuitively, inactive inequality constraints do not play any role near the solution, so they can be dropped
- The active inequality constraints have zero values at solution, and so they can be replaced by equality constraints
- The active set methods are a feasible point method, that is, all iterates remain feasible
- The idea is that, in each iteration, a QP subproblem is solved with a subset of equality constraints, known as the working set, $\mathcal{S}_k$

Active Set Methods III

- Let x^* be a local minimizer for the QP problem, then x^* is a local minimizer of the problem

$$\min_{x^*} \ \frac{1}{2}x^T Qx + q^T x \quad (7)$$
$$\text{subject to: } {Ae_i}^T x = be_i, i \in E \cup I(x^*)$$

- If x^* is a feasible point and a KKT point of the above problem, and the corresponding Lagrangian multiplier vector λ^* satisfies:

$$\lambda_i^* \geq 0 : i \in I(x^*) \quad (8)$$

Then x^* is also a KKT point of the QP problem

CIMAT

Active Set Methods IV

- If the solution of the equality-constrained QP subproblem on the working set, $\mathcal{S}_t$ is feasible for the original QP problem, then it needs to be checked if satisfies the Lagrangian condition
- A search direction is found as:

$$\min_{\mathrm{d}} \ \frac{1}{2}(\mathrm{x_t} + \mathrm{d})^T \mathrm{Q}(\mathrm{x_t} + \mathrm{d}) + \mathrm{g}^T(\mathrm{x_t} + \mathrm{d}) \\ \text{subject to: } \mathrm{Ae_i}^T \mathrm{d} = 0, i \in \mathcal{S}_t \tag{9}$$

Active Set Methods V

- The value of α_t can be computed as follows:

$$\alpha_t = \min \left\{ 1, \min_{i \notin \mathcal{S}_t} \frac{b_i - a_t^T x_t}{a_t^T d_t} : a_t^T d_t \text{is infeasible} \right\} \tag{10}$$

Active Set Methods VI

Algorithm 1 Active Set Method

Set an initial feasible search point $\mathrm{x}^{(0)}$
Set the working set, $\mathcal{S}_0 \leftarrow E \cup I(\mathrm{x}^{(0)})$
Set $t \leftarrow 0$
while a stop criterion is not met **do**
    Find the search direction by solving the optimization problem
    if d_t is zero **then**
        Compute $\lambda_i^{(t)}$
        if $\exists \lambda_i < 0, i \in \mathcal{S}_t \cap I$ **then**
            $\mathcal{S}_t \leftarrow \mathcal{S}_t \setminus i_t, \mathrm{x}_{t+1} \leftarrow \mathrm{x}_t$
        else
            Stop
        end if
    else
        $\mathrm{x}_{t+1} \leftarrow \mathrm{x}_t + \alpha_t d_t$
        if $\alpha_t < 1$ **then**
            Add the violated constraint into the working set
        end if
    end if
    $t \leftarrow t + 1$
end while
return $\mathrm{x}^{(t)}$

Example I

Using the Active Set Method, find the optimal solution of a QP problem given by:

- $Q = \begin{bmatrix} 17 & 14 & -3 \\ 14 & 13 & 0 \\ -3 & 0 & 10 \end{bmatrix}$
- $q = [-2 \quad -2 \quad -1]^T$
- $E = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 3 \end{bmatrix}$
- $b_e = [10 \;\; 5]^T$
- $I = \begin{bmatrix} 0 & 3 & 7 \\ 4 & 7 & 1 \end{bmatrix}$
- $b_i = [3 \;\; 9]^T$

Example II

- $x^{(0)} = [10 \;\; 5 \;\; 0]^T$
- Solving, we obtain that:
  $d^{(0)} = [7.4444 \;\; -22.3333 \;\; 7.4444]^T$; and
  $\lambda^{(0)} = [29.5556 \;\; 16.8889]^T$
- That results in $\alpha = 0.5541$, and then
  $x^{(1)} = [14.1250 \;\; -7.3750 \;\; 4.1250]^T$
- $x^{(1)}$ activates one of the inequality constraint, which is further added to the working set

Example III

- Computing the new direction search and Lagrangian multipliers result in:
  $d^{(1)} = [0.000 \;\; 0.0000 \;\; 0.000]^T$; and
  $\lambda^{(0)} = [77.6875 \;\; 21.4531 \;\; 11.2031]^T$
- Since $d^{(1)}$ is zero and the Lagrangian are greater or equal than zero, then the optimal solution is given by $x^{(1)}$, i.e.,

$$x^* = [14.1250 \quad -7.3750 \;\; 4.1250]^T$$

Exercise I

Using the Active Set Method, find the optimal solution of the QP problems given by:

$$\begin{aligned} \min f(x) = \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right) - 3x_2 - x_3 \\ \text{s.t. } x_1 + x_2 + x_3 \leq 1 \\ x_2 - x_3 \leq 1 \end{aligned} \tag{11}$$

Exercise II

- $Q = \begin{bmatrix} 32 & -17 & 18 & -5 & 0 \\ -17 & 59 & -6 & 42 & 14 \\ 18 & -6 & 58 & 4 & -31 \\ -5 & 42 & 4 & 41 & 12 \\ 0 & 14 & -31 & 12 & 31 \end{bmatrix}$
- $q = [1 \;\; 1 \;\; 4 \;\; 0 \;\; 0]^T$
- $E = \begin{bmatrix} 2 & 3 & 5 & 4 & -3 \\ -3 & 1 & 2 & 4 & -2 \\ -4 & 0 & 2 & 1 & 4 \end{bmatrix}$

CIMAT

Exercise III

- $b_e = [-9 \;\; 0 \;\; -6]^T$
- $$I = \begin{bmatrix} -3 & 2 & 2 & -2 & 4 \\ -3 & -3 & -1 & -1 & 5 \\ 5 & -1 & -3 & 1 & 3 \\ -3 & 0 & 0 & -2 & -1 \\ -4 & 5 & 0 & -2 & 1 \\ 1 & -3 & -3 & 2 & -3 \\ 4 & 4 & 1 & -2 & 5 \end{bmatrix}$$
- $b_i = [8 \;\; 7 \;\; -4 \;\; 2 \;\; -9 \;\; -1 \;\; -3]^T$

Interior-Point Methods I

- Interior-point methods are able to solve quadratic programming problems
- These methods reach the solution by traversing the interior of the feasible region
- For simiplicity, we consider the inequality-constrained QP:

$$\begin{aligned} \min_{x^*} \ & \frac{1}{2}x^T Q x + q^T x \\ \text{subject to: } & A_i x \geq b_i \end{aligned} \tag{12}$$

The Central Path I

- The central path $\mathcal{C}$ is an arc of strictly feasible points
- It is parametrized by a scalar $\tau > 0$, and each point $(x_\tau, \lambda_\tau, y_\tau)$ solves the following system:

$$\begin{aligned} A^T\lambda + \mathrm{y} &= c \\ A\mathrm{x} &= b \\ x_i y_i &= \tau, i = 1, \ldots, m \\ (\mathrm{x}, \mathrm{y}) &> 0 \end{aligned} \tag{13}$$

CIMAT

The Interior-Point Method I

- The first optimality condition:

$$
\begin{aligned}
Qx + q - A^T\lambda &= 0 \\
Ax - b &\geq 0 \\
(Ax - b)_i \lambda_i &= 0 \\
\lambda &\geq 0
\end{aligned}
\tag{14}
$$

The Interior-Point Method II

- By introducing slack variables, $\mathrm{y} \geq 0$, we obtain:

$$\begin{aligned} Q\mathrm{x} + q - A^T\lambda &= 0 \\ A\mathrm{x} - b - y &= 0 \\ y_i\lambda_i &= 0 \\ (\lambda, \mathrm{y}) &\geq 0 \end{aligned} \tag{15}$$

The Interior-Point Method III

- Rewritten:

$$\begin{bmatrix} Qx + q - A^T\lambda \\ Ax - b - y \\ \mathcal{Y}\mathcal{A}e \end{bmatrix} \tag{16}$$

where $\mathcal{Y}$ is a diagonal matrix of the vector y, $\mathcal{A}$ is the diagonal matrix of the vector λ, and e is vector of ones.

The Interior-Point Method IV

- Biasing the search towards the central path:

$$\begin{bmatrix} Qx + q - A^T\lambda \\ Ax - b - y \\ \mathcal{Y}\mathcal{A}e - \sigma\mu e \end{bmatrix} = 0 \tag{17}$$

where $\mu = \dfrac{y^T\lambda}{m}$ and $\sigma = [0, 1]$

The Interior-Point Method V

- Solving the nonlinear system using Newton's method:

$$\begin{bmatrix} G & 0 & -A^T \\ A & -I & 0 \\ 0 & \mathcal{A} & \mathcal{Y} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} -r_d \\ -r_p \\ -\mathcal{A}\mathcal{Y}e + \sigma\mu e \end{bmatrix} \tag{18}$$

where $r_d = Qx + q - A^T$ and $r_p = Ax - b - y$

- The next point is obtained by setting:

$$\left(x^{t+1}, y^{t+1}, \lambda^{t+1}\right) = \left(x^{t}, y^{t}, \lambda^{t}\right) + \alpha\left(\Delta x, \Delta y, \Delta\lambda\right) \tag{19}$$

where α is chosen to keep the inequality $\left(y^{t+1}, \lambda^{t+1}\right) > 0$

The Interior-Point Method VI

- The greatest reduction in the residuals r_d and r_p is obtained by choosing the largest admissible primal (α^p) and dual $\left(\alpha^d\right)$ steplenght
- The residuals satisfy the following conditions:

$$
\begin{aligned}
r_p^{t+1} &= (1-\alpha^p)\, r_p^t \\
r_d^{t+1} &= (1-\alpha^p)\, r_d^t + \left(\alpha^p - \alpha^d\right) Q\Delta x
\end{aligned}
\tag{20}
$$

- The steplenght α is set to $\alpha = \min\left(\alpha_\tau^p, \alpha_\tau^d\right)$, where:

$$
\begin{aligned}
\alpha_\tau^p &= \max\left\{\alpha \in (0,1] : y + \alpha\Delta y \geq (1-\tau)\, y\right\} \\
\alpha_\tau^d &= \max\left\{\alpha \in (0,1] : \lambda + \alpha\Delta\lambda \geq (1-\tau)\, \lambda\right\}
\end{aligned}
\tag{21}
$$

CIMAT

The Interior-Point Method VII

- The most popular interior-point method for convex QP is based on Mehrotra's predictor-corrector
- First, we compute an affine scaling step $\left(\Delta x^{aff}, \Delta y^{aff}, \Delta \lambda^{aff}\right)$ by setting $\sigma = 0$
- Next, we compute the centering parameter σ
- The total step is obtained by solving the following system

$$\begin{bmatrix} G & 0 & -A^T \\ A & -I & 0 \\ 0 & \Lambda & \mathcal{Y} \end{bmatrix} \begin{bmatrix} \Delta x \\ \Delta y \\ \Delta \lambda \end{bmatrix} = \begin{bmatrix} -r_d \\ -r_p \\ -\Lambda \mathcal{Y} e - \Delta \Lambda^{aff} \Delta \mathcal{Y}^{aff} + \sigma \mu e \end{bmatrix} \tag{22}$$

CIMAT

The Interior-Point Method VIII

Algorithm 2 Interior Point Predictor-Corrector for QP

Set an initial search point x^0
Compute $\left(x^0, y^0, \lambda^0\right)$ with $\left(y^0, \lambda^0\right) > 0$
$t \leftarrow 0$
while a stop criterion is not meet **do**
    Set $(x, y, \lambda) \leftarrow (x^t, y^t, \lambda^t)$ and solves (18) with $\sigma = 0$ for $\left(\Delta x^{aff}, \Delta y^{aff}, \Delta \lambda^{aff}\right)$
    $\mu \leftarrow \frac{y^t \alpha}{m}$
    $\hat{\alpha}^{aff} \leftarrow \max\left\{\alpha \in (0, a] : (y, \lambda) + \alpha\left(\Delta y, \Delta \alpha\right) \geq 0\right\}$
    $\mu^{aff} \leftarrow \left(y + \hat{\alpha}^{aff} \Delta y^{aff}\right)^T \left(\lambda + \hat{\alpha}^{aff} \Delta \lambda^{aff}\right)/m$ and set $\sigma \leftarrow \left(\mu^{aff}/\mu\right)^3$
    Solves (22) for $(\Delta x, \Delta y, \Delta \lambda)$
    Choose $\tau_t \in (0, 1)$ and set $\alpha = \min\left(\alpha_\tau^p, \alpha_\tau^d\right)$
    $\left(x^{t+1}, y^{t+1}, \lambda^{t+1}\right) \leftarrow (x^t, y^t, \lambda^t) + \alpha\left(\Delta x, \Delta y, \Delta \lambda\right)$
    $t \leftarrow t + 1$
end while
return $x^{(t)}$

Penalty Function Methods I

- The penalty function methods are an important class of methods for constrained optimization problem
- In this class of methods we replace the original constrained problem by a sequence of unconstrained subproblems that minimizes the penalty functions
- The so-called "penalty" property requires that the penalty function, $\mathcal{P}(x)$, equals to $f(x)$ for all feasible points
- Conversely, $\mathcal{P}(x)$ has to be much larger than $f(x)$ when the constraint violations are severe

Penalty Function Methods II

- A constraint violation can be defined as:

$$
c_i(x) = \begin{cases} c_i(x) & \text{if } i \text{ is an equality constraint} \\ \min\{c_i(x), 0\} & \text{if } i \text{ is an inequality constraint} \end{cases} \tag{23}
$$

- The penalty function will thus consist in a sum of the original objective function and a penalty term, i.e.,

$$
\mathcal{P}(x) = f(x) + h(c(x)) \tag{24}
$$

- The key in this kind of methods relies on the definition of $h(c(x))$

Penalty Function Methods III

- One of the first approaches consisted in penalizing based on the norm of the constraints violation vector, i.e.,

$$\mathcal{P}(x) = f(x) + \sigma \parallel c(x) \parallel^k \tag{25}$$

  where $k > 0$ and $\sigma > 0$
- The basic idea of the penalty function method is that the penalty parameter, σ, is increased in each iteration until $\parallel c(x) \parallel^k$ is greater than a given tolerance, δ

Penalty Function Methods IV

The properties of the penalty methods:

- They substitute unconstrained optimization problems for constrained ones; however, the effective use of numerical methods for unconstrained problems requires the penalty function to be *sufficiently differentiable*
- Unlimited increment of the penalty parameter increases ill-conditioning which is often inherent in the minimization process of the penalty function

Applications

Some applications of nonlinear optimization in the field of machine learning encompass:

- Least-square regression
- Training of neural networks
- Training of support vector machines
- Ensemble learning

Questions?