Constrained Optimization

Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

CUEE

December 30, 2023

Outline

Lagrangian multiplier theorem

Constrained problems

a general contrained optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \leq 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

inequality constraints can be converted to equality constraints

- introduce additional variables $z_1, \ldots, z_m$
- constraints $f_i(x) \leq 0$ for $i = 1, \ldots, m$, are equivalent to

$$f_1(x) + z_1^2 = 0, \quad \ldots, \quad f_m(x) + z_m^2 = 0$$

- a problem with inequality constraints can be regarded as the problem with equality contraints only

Equality-constrained optimization

this lecture consider problems with equality constraints of the form

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

we can consider two approaches of handling the equality constraints

- penalty approach
- elimination approach

Lagrangian function

the **Lagrangian function** $L : \mathbf{R}^n \rightarrow \mathbf{R}^{n+p}$ is defined by

$$L(x, \lambda) = f(x) + \sum_{i=1}^{p} \lambda_i h_i(x)$$

denote $x^\star$ a local minimizer of f

the subspace of first-order feasible directions is defined as

$$S = \{ y \in \mathbf{R}^n \mid \nabla h_i(x^\star)^T y = 0, \quad i = 1, 2, \ldots, p \}$$

- $y \in S$ if y is orthogonal to all p gradients of constraint functions

Lagrange multiplier theorem

regularity assumption: $\nabla h_1(x^\star), \ldots, \nabla h_p(x^\star)$ are linearly independent

if $x^\star$ is a local minimizer of the problem on page 5 then

- **first-order condition:** there exists a unique $\lambda^\star \in \mathbf{R}^p$ called a **Lagrange multiplier vector** such that

$$\nabla f(x^\star) + \sum_{i=1}^{p} \lambda_i^\star \nabla h_i(x^\star) = 0$$

  - at optimum, $\nabla f(x^\star)$ is a linear combination of $\nabla_i h_i(x^\star)$
  - equivalent to the zero gradient of $\mathcal{L}$ forming a total $n+p$ equations in (x, λ)
- **second-order necessary condition**

  moreover, if f and h are twice continuously differentiable, we have

$$y^T \left(\nabla^2 f(x^\star) + \sum_{i=1}^{p} \lambda_i^\star \nabla^2 h_i(x^\star) \right) y \geq 0, \quad \forall y \in S$$

Second-order sufficient condition

assume that f and h are twice continuously differentiable

if $x^\star$ and $\lambda^\star$ satisfy the zero-gradient condition of L:

$$\nabla_x L(x^\star, \lambda^\star) = 0, \quad \nabla_\lambda L(x^\star, \lambda^\star) = 0$$

and satisfy the second-order condition:

$$y^T \nabla_x^2 L(x^\star, \lambda^\star) y > 0, \quad \forall y \neq 0 \text{ and } y \in S$$

then $x^\star$ is a strict local minimum of f subject to $h_i(x) = 0$ for $i = 1, \ldots, p$

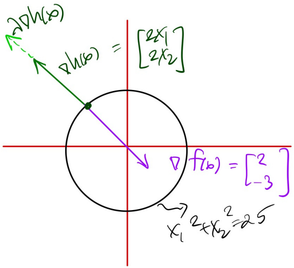
this provides a **sufficient condition** for local optimality of x

Example

minimize $2x_1 - 3x_2$ subject to $x_1^2 + x_2^2 = 25$

the zero-gradient conditions of L are

$$\nabla_x L = \begin{bmatrix} 2 \\ -3 \end{bmatrix} + 2\lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0, \quad \nabla_\lambda L = x_1^2 + x_2^2 - 25 = 0$$



solving the first-order condition gives

$$x^\star = \left(-\frac{10}{\sqrt{13}}, \frac{15}{\sqrt{13}}\right), \quad \lambda^\star = \frac{\sqrt{13}}{10}$$

and the second-order condition is

$$y^T \nabla_x^2 L(x^\star, \lambda^\star) y = 2\lambda^\star y^T y > 0, \quad \forall y \neq 0$$

- the necessary condition suggests that at optimum, $\nabla f(x^\star)$ must be a linear combination of $\nabla h(x^\star)$
- such linear combination exists if $\lambda^\star$ exists
- the sufficient condition guarantees that $x^\star$ is locally optimal
- the sufficient condition only requires that $y^T \nabla_x L(x^\star, \lambda^\star) y > 0$ for all y that perpendicular to $\nabla h(x^\star)$

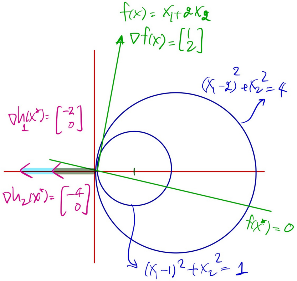
$$y^T \nabla h(x^\star) = 0 \quad \Rightarrow \quad y \in \mathrm{span}\{(3,2)\}$$

(but in this example, the positiveness of $\nabla_x^2 L(x^\star, \lambda^\star)$ holds for all $y \neq 0$)

No Lagrange multiplier

the Lagrange multiplier might not exist in some problem

$$\begin{array}{ll} \text{minimize} & x_1 + 2x_2 \\ \text{subject to} & (x_1-1)^2 + x_2^2 = 1 \\ & (x_1-2)^2 + x_2^2 = 4 \end{array}$$



- there is only one feasible point at $x^\star = 0$
- $\nabla h_1(x^\star)$ and $\nabla h_2(x^\star)$ are not independent
- there is no $\lambda^\star$ for the necessary condition to hold
- we cannot express $\nabla f(x^\star)$ as a linear combination of $\nabla h_1(x^\star)$ and $\nabla h_2(x^\star)$

from the necessary first-order condition, in order for a Lagrange multiplier to exist, $\nabla f(x^\star)$ must be orthogonal to S (subspace of first-order feasible variation)

Quadratic program with linear equality constraint

given $A \in \mathbf{R}^{p \times n}$ of rank p, consider

$$\underset{x}{\text{minimize}} \;\; (1/2)x^T P x - q^T x \quad \text{subject to} \;\; Ax = b$$

assume that P is positive definite on the nullspace of A (more relaxed)

results:

1. it can be shown that the KKT matrix $\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}$ is non-singular (please verify)
2. the zero-gradient of Lagrangian condition is the system of $n + p$ equations

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} q \\ b \end{bmatrix}$$

from 1) and gives a unique $x^\star$ as the global minimizer

➽ QP with linear equality constrained is solved from a linear system to show the result

1. $\nabla h(x) = A$ is full row rank (regularity assumption holds); S is the nullspace of A; and $\nabla_x^2 L = P$ which is positive definite on S (by assumption)
2. from the second-order sufficient condition, a solution $x^\star$ to the linear system is a local minimizer
3. from 2), since the linear system has a unique solution, the local minimizer of this problem is also a global minimizer
4. typically, a gloal minimum is obtained when the problem is convex
5. we did not assume that the problem is convex because the positive definiteness of P is not required on $\mathbf{R}^n$

Example: least-squares with linear constraints

given a full rank $A \in \mathbf{R}^{p \times n}$

$$\text{minimize } (1/2)\|Fx - g\|_2^2 \quad \text{subject to } \; Ax = b$$

the zero-gradient of the Lagrangian: $L(x, \lambda) = (1/2)\|Fx - g\|_2^2 + \lambda^T(Ax - b)$ is

$$\begin{bmatrix} F^T F & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} F^T g \\ b \end{bmatrix}$$

a set of $n + p$ linear equations in variables x and λ

➽ no need to use iterative algorithms

Example: least-norm problem

given a fat and full row rank A

$$\text{minimize } (1/2)\|x\|_2^2 \quad \text{subject to } \; Ax = y$$

meaning: find x that lies on intersections of hyperplanes and is closest to the origin

✎ after applying the Lagrange multiplier theorem,

$$x = A^T(AA^T)^{-1}y$$

- the least-norm problem has a closed-form solution
- ✎ the condition for AA^T to be invertible is from the full rank assumption of A

Equality constraint elimination

Parametrization

when the linear constraints are all linear

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad Ax = b$$

$(A \in \mathbf{R}^{m \times n}, \quad m < n)$ we parametrize the affine feasible set

$$\{x \mid Ax = b\} = \{Fz + \hat{x} \mid z \in \mathbf{R}^{n-p}\}, \quad F \in \mathbf{R}^{n \times n-p}$$

where $\hat{x}$ is a particular solution to $Ax = b$ and $\mathbf{range}(F) \in \mathcal{N}(A)$

we reparametrize and obtain an eliminated optimization problem:

$$\text{minimize} \quad \tilde{f}(z) = f(Fz + \hat{x})$$

the optimization variable is $z \in \mathbf{R}^{n-p}$ (with lower dimension)

Example: least-norm problem with a simplex constraint

$$\text{minimize} \quad \|x\|_2^2 \quad \text{subject to} \quad \mathbf{1}^T x = 1$$

is equivalent to solving

$$\text{minimize} \quad x_1^2 + x_2^2 + \cdots + x_{n-1}^2 + (1 - x_1 - \cdots - x_{n-1})^2$$

with $n - 1$ variables

example: solve the problem

$$\begin{array}{ll} \text{minimize} & -x_1 x_2 x_3 \\ \text{subject to} & \frac{x_1}{a_1} + \frac{x_2}{a_2} + \frac{x_3}{a_3} = 1 \end{array}$$

where $a_1, a_2, a_3 > 0$

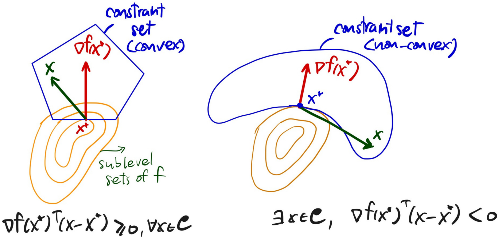
Convex constraints

Optimization over a convex set

we consider a special case of convex-constrained problem

$$\text{minimize} \quad f(x) \quad \text{subject to} \quad x \in \mathcal{C}$$

where f is continuously differentiable over a closed-convex set $\mathcal{C}$



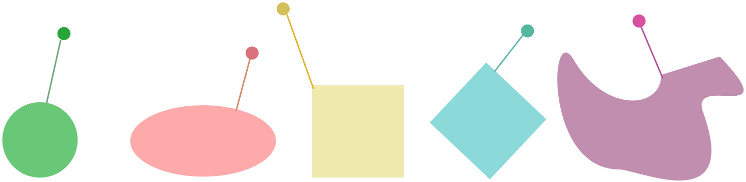
optimality condition: if $x^\star$ is a local minimizer of f over $\mathcal{C}$ then

$$\nabla f(x^\star)^T(x - x^\star) \geq 0, \quad \forall x \in \mathcal{C}$$

Projection onto a convex set

definition: a problem of finding x in $\mathcal{C}$ that is closest to a given vector u

$$\underset{x}{\text{minimize}} \quad \|u - x\|_2^2 \quad \text{subject to} \quad x \in \mathcal{C}$$



- the projection of u on $\mathcal{C}$ is denoted by $\Pi_{\mathcal{C}}(u)$
- here, ℓ_2-norm is used to measure the distance, but this concept can be re-defined using other norms
- when $\mathcal{C}$ is convex, some theoretical results are available

Projection theorem

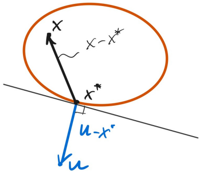
let $\mathcal{C}$ be a non-empty closed-convex set

- for every $u \in \mathbf{R}^n$, the projection $\Pi_{\mathcal{C}}(u)$ exists and is unique
- the mapping $g : \mathbf{R}^n \to \mathcal{C}$ defined by $g(u) = \Pi_{\mathcal{C}}(u)$ is continuous and nonexpansive

$$\|g(u) - g(v)\| \leq \|u - v\|, \quad \forall u, v \in \mathbf{R}^n$$

- given $u \in \mathbf{R}^n$, a vector $x^\star \in \mathcal{C}$ is equal to the projection $\Pi_{\mathcal{C}}(u)$ if and only if

$$(u - x^\star)^T (x - x^\star) \leq 0, \quad \forall x \in \mathcal{C}$$



- in case where $\mathcal{C}$ is a **subspace**, $x^\star$ is equal to $\Pi_{\mathcal{C}}(u)$ if and only if

$$(u - x^\star)^T x = 0, \quad \forall x \in \mathcal{C}$$

Projection on simple convex sets

a closed-form projection can be obtaind if $\mathcal{C}$ is simple

- **non-negative orthant:** $\mathcal{C} = \mathbf{R}^n_+$, we have $\Pi_{\mathcal{C}}(z) = z_+ := \max(0, z)$
- **box or hyper-rectangle:** $\mathcal{C} = \{x \mid l \leq x \leq u\}$

$$(\Pi_{\mathcal{C}}(z))_k = \begin{cases} l_k, & z_k \leq l_k \\ z_k, & l_k \leq z_k \leq u_k \\ u_k, & z_k \geq u_k \end{cases}$$

- ℓ_∞**-norm ball:** $\mathcal{C} = \{x \mid \|x\|_\infty \leq \lambda\}$

$$[\Pi_{\mathcal{C}}(z)]_i = \begin{cases} \lambda, & z_i > \lambda, \\ z_i, & |z_i| \leq \lambda, \\ -\lambda, & z_i < -\lambda \end{cases}$$

Projection on simple convex sets

- **euclidean unit norm ball:** $\mathcal{C} = \{x \mid \|x\|_2 \leq 1\}$

$$\Pi_{\mathcal{C}}(z) = \begin{cases} z/\|z\|_2, & \|z\|_2 \geq 1, \\ z, & \|z\|_2 \leq 1 \end{cases}$$

- **simplex:** $\mathcal{C} = \{x \mid x \succeq 0, \mathbf{1}^T x = 1\}$

$$\Pi_{\mathcal{C}}(z) = (z - \nu\mathbf{1})_+ \triangleq \max(0, z - \nu\mathbf{1})$$

for some $\nu \in \mathbf{R}$ (can find ν using bisection to solve $\mathbf{1}^T(z - \nu\mathbf{1})_+ = 1$)

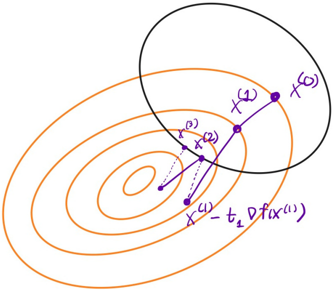
more expressions can be found in Parikh et al. 2013

Gradient projection methods

Gradient projection methods

a simple gradient projection method takes the form

$$x^{(k+1)} = \Pi_{\mathcal{C}}[x^{(k)} - t_k \nabla f(x^{(k)})]$$



- t_k can be fixed, by diminishing rule or by line search (see Bertsekas Chapter 2)
- it takes the gradient-descent direction and project it on $\mathcal{C}$
- the method is practical if the projection is fairly simple
- the convergence properties are essentially the same as those of unconstrained steepest descent method

Step size selection

- fixed step size: $0 < t < 2/L$ where L is a Lipschitz constant of ∇f
- diminising step size: $t_k \to 0$ and $\sum_{k=0}^{\infty} t_k = \infty$
- **Armijo rule** along the projection arc: given factors $\beta, \alpha \in (0,1)$, initialize t
  1. compute a new projection point with step size t
  $$x^+ = \Pi_{\mathcal{C}}(x^{(k)} - t\nabla f(x^{(k)}))$$
  2. check if the condition is satisfied
  $$f(x^+) \leq f(x^{(k)}) - \alpha \nabla f(x^{(k)})^T (x^{(k)} - x^+)$$
  3. if the above condition does not hold, decrease $t := \beta t$ and repeat step 1)

Scaled gradient projection

a basic scaled version of gradient projection is

$$x^{(k+1)} = \underset{x \in \mathcal{C}}{\operatorname{argmin}} \left\{ \nabla f(x^{(k)})^T (x - x^{(k)}) + \frac{1}{2t_k} (x - x^{(k)})^T H_k (x - x^{(k)}) \right\}$$

where H_k is a positive definite matrix (of iteration k) to be chosen by user

- the update step can be regarded as a *generalized* projection problem

$$\underset{x \in \mathcal{C}}{\text{minimize}}\ (x-u)^T H_k (x-u) \quad \text{where} \;\; u = x^{(k)} - t_k H_k^{-1} \nabla f(x^{(k)})$$

- it is equivalent to the problem in transformed coordinate as

$$\underset{y}{\text{minimize}} \;\; f(H_k^{-1/2} y) \quad \text{subject to} \;\; y \in \{ v \mid H_k^{-1/2} v \in \mathcal{C} \}$$

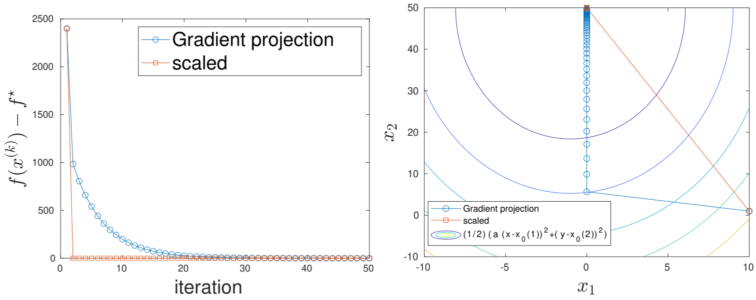
- the convergence rate is governed by the smallest and largest eigenvalues of $H_k^{-1/2}\nabla^2 f(x^{(k)})H_k^{-1/2}$
- this suggests that one should choose $H_k \approx \nabla f^2(x^{(k)})$ but in a diagonal form to maintain simplicity of the generalized projection step
- if $\nabla^2 f(x^{(k)}) \succ 0$ for all $x \in \mathcal{C}$, we can use

$$H_k = \nabla^2 f(x^{(k)})$$

  and this is called **constrained Newton's method** which has a superlinear convergence for $t_k = 1$ (see more results in Bertsekas ex 2.3.2)
- a non-diagonal scaling can improve the convergence but the projection step may not be longer simple
- for non-negative orthant set, a **two-metric projection method** uses a non-diagonal scaling matrix while maintaining the simplicity of the projection on the orthant

Example: quadratic over non-negative orthant

minimize $f(x) = (1/2)(x-c)^T H(x-c)$ over $\mathbf{R}^2_+$ with $H = \begin{bmatrix} \gamma & 0 \\ 0 & 1 \end{bmatrix}$ and $\gamma = 20$, $c = (-1, 50)$



- the gradient projection was implemented with $t = 1.9/\gamma$ (Lipschitz constant is γ)
- the scaled version used $H_k = \nabla^2 f = H$ and $t = 1$ (converged faster)
- both methods was initialized with $x^{(0)} = (10, 1)$; the optimum must occur at $x^\star = (0, c_2)$ (geometrically)

Example: algorithm update details

the scaled gradient projection step is to minimize (over $\mathbf{R}^2_+$)

$$
\begin{aligned}
&(1/2)(x - x^{(k)})^T H_k (x - x^{(k)}) + t\nabla f(x^{(k)})^T (x - x^{(k)}) \\
&= (1/2)[(x - x^{(k)} + tH_k^{-1}\nabla f(x^{(k)}))^T H_k (x - x^{(k)} + tH_k^{-1}\nabla f(x^{(k)})) \\
&\triangleq (1/2)(x-u)^T H_k (x-u), \quad u = x^{(k)} - tH_k^{-1}\nabla f(x^{(k)})
\end{aligned}
$$

- this is a generalized projection on $\mathbf{R}^n_+$ using a **weighted** euclidean norm
- when choosing $H_k = H$ (which is diagonal in this example), the projection has the same closed-form as when $H_k = I$ (a diagonal choice simplifies projections)
- gradient projection step (to $\mathbf{R}^n_+$) for this example is

$$x^+ = \Pi(x - tH(x-c))$$

- the scaled gradient projection step (to $\mathbf{R}^n_+$) is

$$z^+ = \Pi(z - tH_k^{-1}H(z-c)) = \Pi(z - t(z-c))$$

General constrained problems

most of the methods required tools in duality theory and approximation methods

- penalty method
- the method of multipliers
- Lagrangian methods
- Newton-like method
- sequential quadratic programming (SQP)
- interior-point methods

Lagrange multiplier theory can be read in Bertsekas Chapter 3.3

connections among these methods are given in Bertsekas Chapter 4

References

1. Chapter 3 and 4 in D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, 2nd edition, 2003
2. Chapter 11 in D.G. Luenberger and Y. Ye, *Linear and Nonlinear Programming*, 4th edition, Springer, 2008
3. Chapter 6 in N. Parikh and S. Boyd, *Proximal Algorithms*, Foundations and Trends in Optimization, 2013