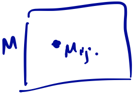
UIUC IE510 Applied Nonlinear Programming

Lecture 11: Lagrangian Multipliers Part a: Equality Constraints

Ruoyu Sun

Review Question for Lecture 10

- **Question 1**: If x^* is a local-min of $\min_x f(x)$, s.t. $x \in X$ (convex), what condition does it satisfy?



$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall x \in X.$$

- **Question 2**: You work on recommendation systems in Netflix, which uses $x_i^T y_j$ to estimate rating M_{ij}, where x_i, y_j are vectors.
  - ◇ You noticed an implicit requirement: rating is often positive.
  - ◇ But your team is using SGD.
  - ◇ What can you say to your team leader?

$$\min_{x_i, y_j's} \sum_{1 \le i,j \le n} (M_{ij} - x_i^T y_j)^2.$$

$$\text{s.t.} \quad x_i^T y_j \geq 0, \ \forall i,j.$$

Here $x_i, y_j \in \mathbb{R}^{r \times 1}$, $i,j = 1, \ldots, n$.

Wrong method → right answer.

First check: at optimum, $x_i^{*T} y_j^* \geq 0$, $\forall i,j$.

If so, then not an issue. Otherwise, an issue.

Second, how to resolve the issue?

- SGD with projection; or CD with projection; or ADMM (to be learned).

Not easy answer.

Summary of Last Lecture

- Optimality condition for

$$\min_x f(x), \text{ s.t. } x \in X.$$

Suppose X convex, f cts-diffentiable. If x^* is a local-min, then

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in X.$$

When X convex, necessary; when f convex, sufficient.

- Gradient projection method

$$\text{GP1}: x^{r+1} = \text{proj}_X(x^r - s_r \nabla f(x^r)).$$

x^r $\quad x^{r+1}$ $\quad \bar{x}^{r+1}$

GP2: pick new iterate along the direction given by GP1.

- Projection simple for: bounds, ball, simplex; sometimes linear
- Stepsize rules: constant; line search for s_r or α_r

This Lecture

set X: implicit description.
e.g. $X = \{\lambda \mid \lambda$ is an eigenvalue of a matrix A^3+2A^2+A for some matrix $A\}$.

- From today: **constrained optimization** with explicit equality/inequality constraints
- After the lecture, you should be able to
  - Write down the 1st and 2nd order conditions for equality constrained problems
  - Define Lagrangian multipliers and Lagrangian functions
  - Compute/verify optimal solutions for simple equality constrained problems
- Advanced goal: explain the two proof ideas and why they are useful

Outline

Optimization Over Convex-set: Criticism

- What is the drawback of the optimality condition

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \forall\, x \in X? \tag{1}$$

- Back to the origin: what should be an "optimality condition"?
  - Of course, $f(x) \geq f(x^*)$, $\forall x \in X$ is also optimality condition
  - Issue: hard to check
  - Is the condition (1) really better? Sort of; unclear
- There can be multiple optimality conditions. How to judge?
  - Easily checkable
  - Leading to algorithm design

Optimization Over Convex-set: Criticism

- What is the drawback of the optimality condition

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \forall\, x \in X? \tag{1}$$

- Back to the origin: what should be an "optimality condition"?
  - Of course, $f(x) \geq f(x^*),\ \forall x \in X$ is also optimality condition
  - Issue: ________
  - Is the condition (1) really better? Sort of; unclear
- There can be multiple optimality conditions. How to judge?
  - Easily checkable
  - Leading to algorithm design

Equality Constrained Problem

Let us first consider the following equality constrained problem

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

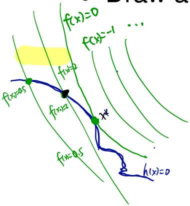
e.g. $\|x\|^2 = 1$, $Ax = b$, $XY^T = A$, etc.

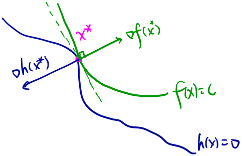
where $f : R^n \mapsto R$, $h_i : R^n \mapsto R, i = 1, ..., m$, are continuously differentiable functions.

Is this "optimization over convex set"? No, unless $h_i(x^*)$'s are affine.

Intuition: One Constraint

- Consider one constraint $\min_x f(x),$ s.t. $h(x) = 0$.
- Draw a plot of level set of f and the set $h(x) = 0$





- Observation: at an optimal solution, $\nabla f(x^*) \,//\, \nabla h(x^*)$,

  or $\nabla f(x^*) = \lambda^* \nabla h(x^*)$, for some λ^*.

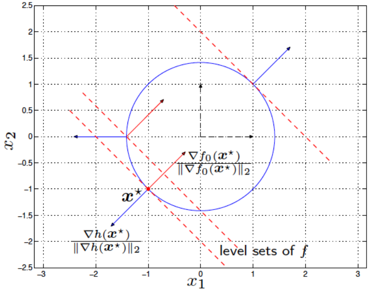
Example 1

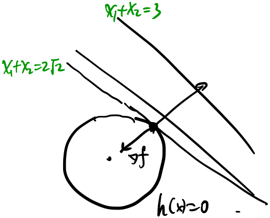
$$\nabla f = \begin{pmatrix} 1 \\ 1 \end{pmatrix},$$

$$\nabla h = \begin{pmatrix} 2x_1 \\ 2x_2 \end{pmatrix} = 2\begin{pmatrix} x_1^* \\ x_1^* \end{pmatrix}.$$

They touch at point

$$(x_1^*, x_2^*) = (x_1^*, x_1^*) = (\sqrt{2}, \sqrt{2}).$$





Consider the problem

$$\begin{aligned} \underset{x\in\mathbb{R}^2}{\text{minimize}} \quad & f_0(x) = x_1 + x_2 \\ \text{subject to} \quad & h(x) = x_1^2 + x_2^2 - 2 = 0. \end{aligned}$$

This is a problem with a linear objective function $f(x)$ and one nonlinear equality constraint $h(x) = 0$. At the solution $x^\star$, the gradient of the constraint $\nabla h(x^\star)$ is orthogonal to the level set of the function at x^*, and hence $\nabla h(x^\star)$ and $\nabla f_0(x^\star)$ are parallel *i.e.*, there is a scalar $\nu^\star$ such that

$$\nabla f_0(x^\star) + \nu^\star \nabla h(x^\star) = \mathbf{0}.$$

Clearly, in this example $x^\star$ is regular (because $\nabla h(x^\star) \neq \mathbf{0}$).

Degenerate Case

- Still one-constraint problem $\min_x f(x)$, s.t. $h(x) = 0$.
- Is $\nabla f(x^*) = \lambda \nabla h(x^*)$ always true? Assuming f and h are smooth.

- **Example 2**: $\min_{x \in \mathbb{R}} x^2$, s.t. $(x-1)^2 = 0$.

  Feasible set $\{1\}$. So optimal solution $x^* = 1$.

  Exercise: Check whether $\nabla f(x^*) = \lambda \nabla h(x^*)$.

  $\nabla f(x^*) = 2x^* = 2$,

  $\nabla h(x^*) = 2(x^*-1) = 0$.

  $2 \neq \lambda \cdot 0, \forall \mathbb{R}$.

- Modified problem: $\min_{x \in \mathbb{R}} x^2$, s.t. $x - 1 = 0$.
  Check: $\nabla f(x^*) = 2$, $\nabla h(x^*) = 1$, so $\nabla f(x^*) = 2 \cdot \nabla h(x^*)$.

Equality Constrained Problem: Optimality Conditions

Lagrange Multiplier Theorem

- Let x^* be a local min and a regular point ($\nabla h_i(x^*)$: linearly independent). Then there exist unique scalars $\lambda_1^*, ..., \lambda_m^*$ such that

  $$\text{1st order: } \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$

  If in addition f and h are twice continuously differentiable,

  $$\text{2nd order: } y'\left(\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*)\right) y \geq 0, \quad \forall \underset{y \in \mathbb{R}^n.}{y} \text{ s.t. } \nabla h_i(x^*)' y = 0, \ \forall i.$$

Characterizes a set of necessary conditions for local min.

Optimality Conditions in Words

To learn math, you should be able to translate “math language” to “human language”.

Lagrangian multiplier theorem (in words): for (smooth) equality constrained problems, at a “regular” local-min (gradients of constraints are linearly independent), we have

- The gradient of the objective can be linearly spanned by the gradients of constraints.
- The Hessian of the objective function plus the same span of Hessian of constraints is positive semidefinite in the space orthogonal to all gradients of constraints.

Optimality Conditions in Words

To learn math, you should be able to translate "math language" to "human language".

Lagrangian multiplier theorem (in words): for (smooth) equality constrained problems, at a "regular" local-min (gradients of constraints are linearly independent), we have

- The gradient of the objective can be linearly spanned by the gradients of constraints.

- The Hessian of the objective function plus the same span of Hessian of constraints is positive semidefinite in the space orthogonal to all gradients of constraints.

Short version: linear combination of Hessians is PSD in some space (orthogonal complement space).

Outline

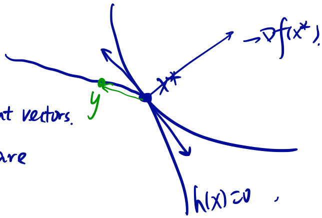
Proof Method 1: Feasible Direction

- Two ways to develop the theory of Lagrangian multipliers
  - **Feasible direction** viewpoint
  - **Penalty** approach

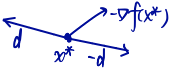
- For illustration, consider $\min_{x \in \mathbb{R}^n} f(x)$, s.t. $h(x) = 0$.

- Feasible variation $y - x^*, y \in X$ cannot be a descent variation. Draw a plot.

feasible variation: $y - x^*$, where $h(y)=0$.
As $y \to x^*$, becomes tangent vectors.
$\therefore$ Feasible variations are tangent vectors



Tangent vectors are not descent
$\Rightarrow$ tangent vectors $\perp \nabla f(x^*)$
(Since d is tangent
$\Rightarrow -d$ is tangent;
If $d \not\perp \nabla f(x^*)$, then either d or $-d$ is descent direction.)



Proof Method 1: Feasible Direction (cont'd)

- A bit more formally, use Taylor expansion:

$$f(x^* + d) = f(x^*) + \langle \nabla f(x^*), d\rangle + o(\|d\|^2) \geq f(x^*),$$

where $d = y - x^*$ in which $y \in X$ is any feasible variation.

$$\langle \nabla f(x^*), d\rangle + o(\|d\|^2) \geq 0 \qquad (2)$$

for any feasible variation d.

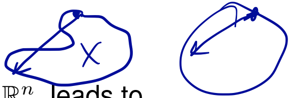
- Idea: Represent feasible variation d by Taylor expansion of h:

$$h(x^* + d) \approx h(x^*) + \langle \nabla h(x^*), d\rangle, \text{ i.e. } \quad 0 \approx \langle \nabla h(x^*), \underbrace{d}_{\text{tangent}}\rangle.$$

- If $\nabla h(x^*) \neq 0$ and $\nabla f(x^*) \nparallel \nabla h(x^*)$, then there is some d
  - orthogonal to $\nabla h(x^*)$
  - and positively related to $\nabla f(x^*)$,

violating (2).

Proof Method 1: More Formal Proof (Reading)



Remark 1: One issue of last-page proof.

- Unconstrained case: d can be anything in $\mathbb{R}^n$, leads to $\nabla f(x^*) = 0$.
    - Pick $d = -\alpha \nabla f(x^*)$ or positively related direction, let $\alpha \to 0$.
- **Issue**: for constrained case, d is restricted: cannot even scale.
    - thus finding one d is not enough
- **Correction**: there is such a sequence of d^k (norm going to zero)

Remark 2: A cleaner proof is by "elimination" [Sec. 3.1.2.]:

– represent some variables by others

– and transform to unconstrained problem.

It is essentially a "feasible direction" proof.

$h(x_1, x_2, x_3) = 0$

$\sim \ x_1 = \varphi(x_2, x_3)$

$\Rightarrow \min_{x_2, x_3} f(\varphi(x_2, x_3), x_2, x_3)$.

Check unconstrained 1st order condition.

Will get desired 1st order condition

Proof Method 2: Penalty Approach

- For illustration, consider (P1): $\min_{x\in\mathbb{R}^n} f(x)$, s.t. $h(x) = 0$.
- Consider another problem (P2)
  $\min_{x\in\mathbb{R}^n} f(x) + k\|h(x)\|^2 + \|x - x^*\|^2. \overset{\Delta}{=} F(x).$
- Wish: If x^* is a local-min of (P1), then also a local-min of (P2).
  $\Downarrow$ $\nabla F(x^*) = 0$
  - This implies the desired 1st order condition, done.

$$\text{1oC:} \quad \nabla f(x^*) + \underbrace{2kh(x^*)}_{\lambda^*}\nabla h(x^*) = 0.$$

- Fact: it only holds for $k \to \infty$.

  Claim: As $k \to \infty$, there is a sequence of local-mins of (P2) $\{x^k\}$ that converge to x^*.

Question Is x^* global-min of (P2), provided that x^* is global-min of (P1)?
Answer: quite subtle .... Will discuss in later lectures.

Why are These Proofs Useful?

Geometrical understanding of Lagrangian multiplier condition.

- Viewpoint 1: $\nabla f(x^*)$ spanned by $\nabla h_k(x^*)$
- Viewpoint 2: $\nabla f(x^*)$ orthogonal to tangent space of constraint manifold
- Quick reason: feasible direction $\neq$ descent direction

Algorithm design.

- "Penalty" is critical for constrained-opt algorithm design
- Related constrained to unconstrained

Formal Proof of the Theorem (Sec. 3.1.1.) –Reading

- Suppose x^* is a local min satisfying $h(x^*) = 0$. Pick any $\alpha > 0$. Consider
$$f^k(x) = f(x) + k|h(x)|^2 + \frac{\alpha}{2} \|x - x^*\|^2 .$$
- Let x^k be a constrained minimizer of f^k over the region $\{x \mid f(x^*) \leq f(x), \|x - x^*\| \leq 1\}$. We will show that x^k is an unconstrained local min of f^k for all large k.
- Taking limit $k \to \infty$ of
$$f^k(x^k) = f(x^k) + k|h(x^k)|^2 + \frac{\alpha}{2} \left\|x^k - x^*\right\|^2 \leq f^k(x^*) = f(x^*)$$
along any convergent subsequence of $\{x^k\}$,we get $h(\bar{x}) = \lim_{k\to\infty} h(x^k) = 0$.
- Furthermore, taking limit of $f(x^k) + \frac{\alpha}{2} \|x - x^*\|^2 \leq f(x^*)$ shows
$$f(\bar{x}) + \frac{\alpha}{2} \|\bar{x} - x^*\|^2 \leq f(x^*)$$
- Since $h(\bar{x}) = 0$, it follows that $f(x^*) \leq f(\bar{x})$.Thus, we have $\bar{x} = x^*$ and $f(x^*) = f(\bar{x})$.

Formal Proof of the Theorem (Sec. 3.1.1.) –Reading

- Since $\bar{x}$ is any limit point, we have $x^k \to x^*$,so $\left\|x^k - x^*\right\| < 1$ for large k,$\Rightarrow$ x^k for k large enough, x^k is an unconstrained local min of f^k, satisfying

$$\nabla f^k(x^k) = 0, \quad \nabla^2 f^k(x^k) \succeq 0.$$

- Taking limit of the following optimality condition

$$0 \; = \; \nabla f(x^k) + 2kh(x^k)\nabla h(x^k) + \alpha(x^k - x^*) \tag{3}$$

Since $\nabla h(x^*)$ has rank m, $\nabla h(x^k)$ also has rank m for large k, so $\nabla h(x^k)'\nabla h(x^k)$: invertible.

Multiplying (1) with $\nabla h(x^k)'$ yields

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

- Taking limit as $k \to \infty$ and $x^k \to x^*$,

$$\{kh(x^k)\} \to (\nabla h(x^*)'\nabla h(x^*))^{-1}\nabla h(x^*)'\nabla f(x^*) \equiv \lambda.$$

Formal Proof of the Theorem (Sec. 3.1.1.) –Reading

Taking limit as $k \to \infty$ in Eq.(1), we obtain

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

- **Exercise**: 2nd order L-multiplier condition: Use 2nd order unconstrained condition for x^k, and algebra.

Outline

Optimality Conditions for Equally Constrained Problems

Two Proof Methods: Feasible Direction and Penalty

Lagrangian Function and Sufficient Conditions

Lagrangian Function

- Define the Lagrangian function (λ: Lagrangian multiplier)

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

Different from $f(x) + \sum_i h_i(x)^2$.

Then, if x^* is a local minimum which is regular, $\nabla L = 0$.

1st order Condition (1oC): $\nabla_x L(x^*, \lambda^*) = 0, \quad \nabla_\lambda L(x^*, \lambda) = 0.$ $\longrightarrow$ $h_i(x^*) = 0, \forall i$.

$\nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) = 0$.

2nd Order Condition (2oC): $y' \nabla^2_{xx} L(x^*, \lambda^*) y \geq 0, \quad \forall y \ s.t. \ \nabla h(x^*)' y = 0.$

Note: $\nabla^2_{\lambda\lambda} L(x; \lambda) = 0$.

- **Remark**: $n + m$ variables, $n + m$ equations in 1oC
- **Example:**

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) \\ \text{subject to} & x_1 + x_2 + x_3 = 3. \end{array}$$

Necessary conditions

$$x_1^* + \lambda^* = 0, \quad x_2^* + \lambda^* = 0, \quad x_3^* + \lambda^* = 0, \quad x_1^* + x_2^* + x_3^* = 3.$$

Original: 3 variables, 1 equation + $\min_x \square$.

1oC: 4 variables, 4 equations

Original: 5 var's 2 equations

1oC: 7 -- , 7. equations.

Optimality Conditions in Words

Lagrangian multiplier theorem (restated, in words): for (smooth) equality constrained problems, at a “regular” local-min (gradients of constraints are linearly independent), we have

- The gradient of the Lagrangian function is zero
- The Hessian of the Lagrangian function w.r.t. x is positive semidefinite in the space orthogonal to all gradients of constraints.
  - Another way: in the nullspace defined by all gradients of constraints

Optimality Conditions in Words

Lagrangian multiplier theorem (restated, in words): for (smooth) equality constrained problems, at a “regular” local-min (gradients of constraints are linearly independent), we have

- The gradient of the Lagrangian function is zero
- The Hessian of the Lagrangian function w.r.t. x is positive semidefinite in the space orthogonal to all gradients of constraints.
  - Another way: in the nullspace defined by all gradients of constraints

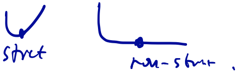
Suffciency Condition

- Second Order Suffciency Conditions: Let $x^* \in R^n$ and $\lambda \in R^m$ satisfy

$$\nabla_x L(x^*, \lambda^*) = 0, \;\; \nabla_\lambda L(x^*, \lambda^*) = 0,$$
$$y' \nabla^2_{xx} L(x^*, \lambda^*) y > 0, \forall y \neq 0 \text{ with } \nabla h_i(x^*)' y = 0. \; \forall i,$$

Then x^* is a strict local minimum.

Remark: No need to have regularity condition.



- **Example:**

$$\begin{array}{ll} \text{minimize} & -(x_1x_2 + x_2x_3 + x_1x_3) \\ \text{subject to} & x_1 + x_2 + x_3 = 3. \end{array}$$

$\nabla L = 0.$

We have that $x_1^* = x_2^* = x_3^* = 1$ and $\lambda^* = 2$ satisfy the 1st order conditions. Also

$$\nabla^2_{xx} L(x^*, \lambda^*) = \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix}$$

We have for all $y \neq 0$ with $\nabla h(x^*)' y = 0$ or $y_1 + y_2 + y_3 = 0$,

$$y' \nabla^2_{xx} L(x^*, \lambda^*) y = -y_1(y_2 + y_3) - y_2(y_1 + y_3) - y_3(y_1 + y_2) = y_1^2 + y_2^2 + y_3^2 > 0$$

Hence, x^* is a strict local minimum.

Proof Preparation: A Useful Lemma (Reading)

- Let P and Q be two symmetric matrices. Assume that $Q \succ 0$ and $P \succ 0$ on the nullspace of Q,i.e.,$x'Px > 0$ for all $x \neq 0$ with $x'Qx = 0$. Then there exists a scalar c such that

$$P + cQ : \text{positive definite}, \forall c > \bar{c}.$$

- Proof: Assume the contrary. Then for every k, there exists a vector x^k with $x^k = 1$ such that

$$(x^k)'Px^k + k(x^k)'Qx^k < 0.$$

Consider a subsequence $\{x^k\}_{k \in \mathcal{K}}$ converging to some x with $x = 1$. Taking the limit superimum,

$$x'Px + \limsup_{k \to \infty \ k \in \mathcal{K}} (k(x^k)'Qx^k) \leq 0.$$

We have $(x^k)'Qx^k \geq 0$ (since $Q \succeq 0$), so

$$\{(x^k)'Qx^k\}_{k \in \mathcal{K}} \to 0.$$

Therefore, $x'Qx = 0$ and using the hypothesis, $x'Px > 0$, a contradiction.

Proof by Penalty Approach (Reading)

Consider the *augmented Lagrangian function*

$$L_c(x, \lambda) = f(x) + \lambda' h(x) + \frac{c}{2} \|h(x)\|^2 ,$$

where c is a scalar. We have

$$\nabla_x L_c(x, \lambda) = \nabla_x L(x, \tilde{\lambda}), \ \nabla^2_{xx} L_c(x, \lambda) = \nabla^2_{xx} L(x, \tilde{\lambda}) + c \nabla h(x) \nabla h(x)'$$

where $\tilde{\lambda} = \lambda + ch(x)$. If (x^*, λ^*) satisfy the suffciency conditions, we have using the lemma (why?),

$$\nabla_x L_c(x^*, \lambda^*) = 0, \ \nabla^2_{xx} L_c(x^*, \lambda^*) > 0,$$

for suffciently large c. Hence for some $\gamma > 0, \epsilon > 0$, (x^* unconstrained local min)

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2} \|x - x^*\|^2 ,$$

if $\|x - x^*\| < \epsilon$. Since $L_c(x, \lambda^*) = f(x)$ when $h(x) = 0$,

$$f(x) \geq f(x^*) + \frac{\gamma}{2} \|x - x^*\|^2 , \ if \ h(x) = 0, \|x - x^*\| < \epsilon.$$

Summary

In this lecture, we learned the following: (think about what you have learned)

- Optimality conditions (1st and 2nd order) for equally constrained problems
- Two proofs: feasible direction and penalty
- Lagrangian multipliers λ and Lagrangian function $L(x, \lambda)$

  Express optimality conditions by derivatives of L

Summary

In this lecture, we learned the following:

- Optimality conditions (1st and 2nd order) for equally constrained problems

- Two proofs: feasible direction and penalty

- Lagrangian multipliers λ and Lagrangian function $L(x, \lambda)$

  Express optimality conditions by derivatives of L