UIUC IE510 Applied Nonlinear Programming

Lecture 10: Optimization over a Convex Set

Ruoyu Sun

One Framework to Cover Unconstrained Optimization

- The first part of the semester: **unconstrained optimization**
- Starting point: gradient descent method
- Iteration complexity (for strongly convex case): $O(\kappa \log \frac{1}{\varepsilon})$.
- Three classes of "faster" methods
  - HB or Nesterov: reduce κ to $\sqrt{\kappa}$
  - CD/SGD: reduce κ to $\frac{\lambda_{avg}}{\lambda_{min}} = \kappa_{CD}$.
  - Newton/BFGS/BB (using curvature info). : "eliminate κ", reduce $\log\frac{1}{\varepsilon}$ to $\log\log\frac{1}{\varepsilon}$. (locally) (true convergence rate unknown)

This Lecture

- Starting from today: **constrained optimization**
- Today: optimization over convex sets
- After this lecture, you should be able to
  - Apply optimality conditions for optimization over convex sets
  - Apply gradient projection method
  - Tell the pros and cons of gradient projection method

Outline

Motivating Example: Support Vector Machine

w

x_i + + + − − − $y_i = 1$ $y_i = -1$

- **Training data set**: two classes $\{\mathbf{x}_i, y_i\}_{i=1}^M$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1,\ 1\}$
- Suppose the training data are linearly separable
- **Objective**: find a hyperplane to separate the data points, i.e., find w such that

$$y_i\, x_i^T \mathbf{w} > 0, \quad \forall i.$$

$$\begin{cases} x_i^T w > 0, & \text{if } y_i > 1 \\ x_i^T w < 0, & \text{if } y_i < 1. \end{cases}$$

- Equivalent to: find $\mathbf{w}$ such that $y_i x_i^T \mathbf{w} \geq 1, \forall i.$

Scale w. s.t. $y_i x_i^T (w \cdot 10^6) \geq 1$.

Art of Constraints

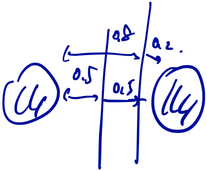
- **Formulation**: Consider the following problem

$$\min_{\mathbf{w}} \quad 0 \tag{1}$$

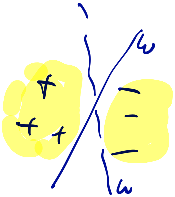
$$\text{s.t.} \quad y_i \cdot w^T x_i \geq 1 \ , \ \forall\, i \tag{2}$$

- The objective does _____ nothing
- The constraint says "no classification error"
- This is a feasibility problem

("requirement", can be viewed as constraint)



Support Vector Machine



- Infinitely many solutions, pick which one?
- **SVM**: Find the separating plane that is far away from both classes
- **Formulation** of SVM:

$\|w\|_1, \quad \|w\|_3.$

$$\min_{\mathbf{w}} \quad \|w\|^2 \tag{3}$$

Exercise: margin

$$\text{s.t.} \quad y_i \cdot w^T x_i \geq 1 \ , \ \forall\, i \tag{4}$$

- **Questions**: How to characterize the optimal solution? Have a feasible solution? Which algorithm?

Outline

Motivation: SVM

Optimality Condition of Constrained Optimization

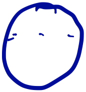
Gradient Projection Method

Constrained Optimization Problem

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

- In the most general form, X can be any set, f can be any function on X
- In most parts of the course, we assume f is continuously differentiable, X is a convex set
- Convex set X means we allow the following types of constraints
  1. $g(x) \leq 0$ where $g(x)$ is a convex function   e.g. $\|x\|^2 \leq 1$
  2. $h(x) = 0$ where $h(x)$ is an affine function: $Cx + d = 0$

  If $g(x)$ is convex,
- Why $g(x) = 0$ is not a convex set? Consider $\|x\|^2 = 1$,



Optimality Conditions

minimize $f(x)$
subject to $x \in X$,

- **Question:** How to characterize the global/local optimal solution x^*?

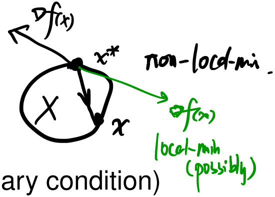
- Still $\nabla f(x) = 0$, $\nabla^2 f(x) \succ 0$?

  $+ \quad x \in X.$ Not the right condition

Fermat 17th century.

Optimality Conditions

$$\begin{array}{|ll|}\hline \text{minimize} & f(x) \\ \text{subject to} & x \in X, \\ \hline\end{array}$$



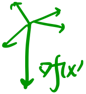
- If x^* is a local minimum of f or X, then (necessary condition)

$$\langle \nabla f(x^*), \underline{x - x^*} \rangle \geq 0, \ \forall x \in X. \tag{5}$$

$\not\Rightarrow \nabla f(x^*) = 0.$

- **Remark 1**: For general f (possibly nonconvex), solutions satisfying this condition is called stationary point (nowhere to move).

$\nabla f(x^*) = 0 \Rightarrow (5).$ If $X = \mathbb{R}^n$, $\nabla f(x^*) = 0 \Leftrightarrow (5).$



- **Remark 2**: If f convex, this condition is also sufficient for x^* to minimize f over X.

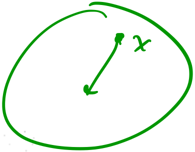
<u>Coro:</u> If $x^* \in \text{int}(X)$, then $\nabla f(x^*) = 0 \Leftrightarrow (5).$

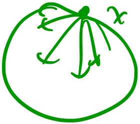
Feasible Directions

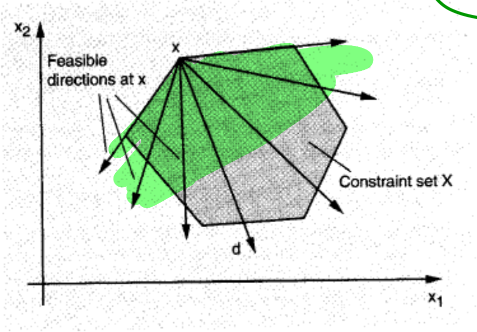
- A feasible direction at an $x \in X$ is a vector $d \neq 0$ such that $x + \alpha d$ is feasible for all sufficiently small $\alpha > 0$
- The set of feasible directions at x is the set of all

$$\alpha(z - x)$$

where $z \in X$, $z \neq x$, and $\alpha > 0$







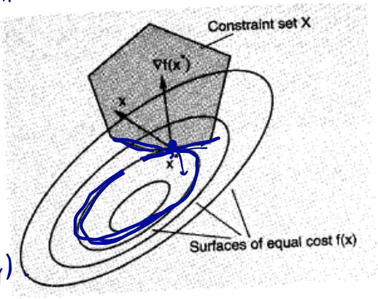
Proof of Optimality Conditions

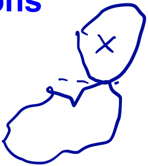
"$\exists$ Better $\Rightarrow$ not best"

If d is both descent, and feasible.

$\begin{cases} f(x^* + \alpha d) < f(x), \\ x^* + \alpha d \in X. \end{cases}$

$\Rightarrow$ x^* is NOT optimal (even locally)

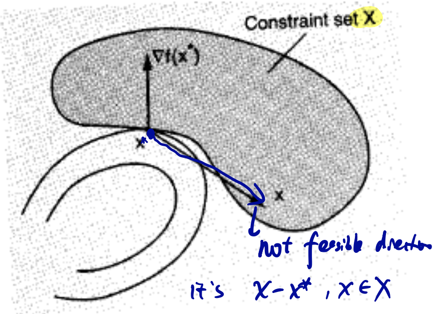




- Descent direction d: for small enough $\alpha > 0$, $f(x + \alpha d) < f(x)$.
  - Set of descent directions: $\langle -\nabla f(x), z \rangle > 0$; i.e. $\langle \nabla f(x), z \rangle < 0$
- **Proof**: At a local-min, feasible direction $\neq$ descent direction. . $x - x^*$ feasible
- Utilizing the characterizations of feasible directions and descent directions, we get

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

Graph Illustration of Optimality Conditions



Find example:
(1) x^* is Local min
(2) Condition fails.

Key: for nonconvex set X, $x - x^*$, where $x \in X$ is possibly not a descent direction at x^*.

Remark: When X is not convex, the condition is NOT necessary.

- For example, x^* is a local min but we have $\nabla f(x^*)'(x - x^*) < 0$ for some feasible vector $x \in X$.

Optimality Conditions with/without Convexity

f, X

- **Summary**: What kind of condition is (5)?

	X convex	X general
f general	**necessary**	not nec., not sufficient
f convex	suf. & nec.	? Sufficient.

Vote. A: Not meaningful

B: maybe meaningful

X convex

nec.

f convex suf.

nec.	Neither
suf. + nec.	suf.

Rigorous Proofs

a) Suppose that $\nabla f(x^*)'(x - x^*) < 0$ for some $x \in X$. By the Mean Value Theorem, for every $\epsilon > 0$ there exists an $s \in [0, 1]$ such that

$$f(x^* + \epsilon(x - x^*)) = f(x^*) + \epsilon \nabla f(x^* + s\epsilon(x - x^*))'(x - x^*).$$

Since ∇f is continuous, for sufficiently small $\epsilon > 0$,

$$\nabla f(x^* + s\epsilon(x - x^*))'(x - x^*) < 0,$$

so that $f(x^* + \epsilon(x - x^*)) < f(x^*)$. The vector $x^* + \epsilon(x - x^*)$ is feasible for all $\epsilon \in [0, 1]$ because X is convex, contradicting the local optimality of x^*.

b) Using the convexity of f

$$f(x) \geq f(x^*) + \nabla f(x^*)'(x - x^*)$$

for every $x \in X$.If the condition $\nabla f(x^*)'(x - x^*) \geq 0$ holds for all $x \in X$,we obtain $f(x) \geq f(x^*)$, so x^* minimizes f over X.

Application of (5): Optimization Subject to Bounds

$\begin{cases} \min f(x) \\ \text{s.t. } x_i \geq 0, \forall i. \end{cases}$

- Consider nonnegative orthant: $X = \{x \mid x \geq 0\}$.
  - Then the necessary condition for $x^* = (x_1^*, \ldots, x_n^*)'$ to be a local min is

$$\sum_{i=1}^n \frac{\partial f(x^*)}{\partial x_i}(x_i - x_i^*) \geq 0, \forall x_i \geq 0, i = 1, \ldots, n.$$

- Fix i. Let $x_j = x_j^*$ for $j \neq i$ and $x_i = x_i^* + 1$:

$$\frac{\partial f}{\partial x_i}(x^*) \underline{\geq} 0$$

- If $x_i^* > 0$. let also $x_j = x_j^*$ for $j \neq i$ and $x_i = \frac{1}{2}x_i^*$. Then $\frac{\partial f(x^*)}{\partial x_i} \leq 0$, so

If $x_i^* > 0$, then $\frac{\partial f}{\partial x_i}(x^*) = 0$. ("interior")

Another example: apply optimality condition to $\min_x f(x)$ s.t. $\sum_i x_i = 1$, $x_i \geq 0, \forall i$.

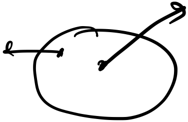
Outline

Motivation: SVM

Optimality Condition of Constrained Optimization

Gradient Projection Method

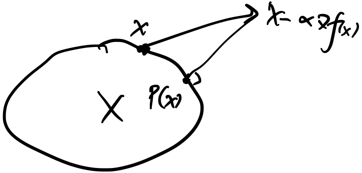
Projection Over A convex Set



- Can we use gradient descent to solve $\min_x f(x), s.t. x \in X$?
  $x \to x - \nabla f(x)$.
- **Central issue:** What if the iterate goes out of the feasible set X?
- **One solution:** "Project" it back to X!

Projection Over A convex Set

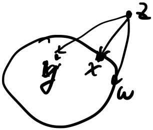
- Can we use gradient descent to solve $\min_x f(x), s.t. x \in X$?
  $x \to x - \nabla f(x)$.
- **Central issue:** What if the iterate goes out of the feasible set X?
- **One solution:** "Project" it back to X!



Projection Over A convex Set

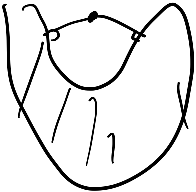
- **Projection Theorem** (part 1): Let $z \in R^n$ and a closed convex set X be given. Problem:

$$\begin{array}{ll}\text{minimize} & f(x) = \|z - x\|^2 \\ \text{subject to} & x \in X.\end{array}$$

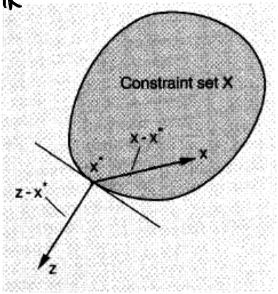


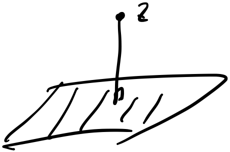
has a unique solution $x^* = \text{proj}[z]$ (the projection of z).

- $x^* = \text{proj}[z] \iff$ The angle between $z - x^*$ and $x - x^*$ is greater or equal to 90 degrees for all $x \in X$, or $(z - x^*)'(x - x^*) \leq 0$
- If X is a subspace of $\mathbb{R}^n$, $z - x^* \perp X$. $\longrightarrow$ "Orthogonality principle":



nonconvex set;
(1) projection may not be unique;
(2) Angle may be larger than 90°.





- The mapping $f : R^n \mapsto X$ defined by $f(x) = \text{proj}[x]$ is continuous and non-expansive, that is,

$$\| \text{proj}[x] - \text{proj}[y] \| \leq \| x - y \|, \forall x, y \in R^n.$$

Why? [Add $\langle x - \text{proj}[x], \text{proj}[y] - \text{proj}[x] \rangle \leq 0$ to $\langle y - \text{proj}[y], \text{proj}[x] - \text{proj}[y] \rangle \leq 0$]

- **Exercise:** Assume X is convex. A vector $x^* \in X$ is a stationary point of

$$\boxed{\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \end{array}}$$

iff x^* satisfies the following fixed point equation

$$x^* = \text{proj}[x^* - \alpha \nabla f(x^*)] \quad \Leftrightarrow (5)$$

for any $\alpha > 0$.

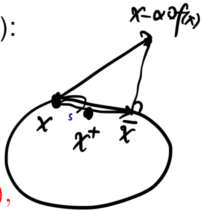
$x^* = \varphi(x^*)$

Another optimality condition.

principle of algorithm design:

global-min / desired solution should be fixed point (or close to)

Gradient Projection Methods



- Simplest version of **Gradient projection method** (GP):

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

- **Gradient projection method** (GP):

$$\text{GP2}: \quad x^{r+1} = x^r + \alpha_r(\bar{x}^r - x^r),$$

$$\text{where} \quad \bar{x}^r = \text{proj}_X[x^r - s_r \nabla f(x^r)]$$

  where, $\text{proj}_X[\cdot]$ denotes projection on the set X, $\alpha_r \in (0, 1]$ is a stepsize, and s_r is a positive scalar.

- Stepsize rule for GP1, i.e. assuming $\alpha_r \equiv 1$:
  - Armijo along the projection arc (s_r: variable)
  - constant stepsize for s_r
  - Diminishing s_r

- Stepsize rules for GP2. Allow changing α_r, but fixed $s_r \equiv s$
  - Limited minimization
  - Armijo along the feasible direction
  - constant stepsize

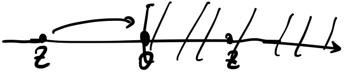
Perform the Projections

$\min f(x)$
s.t. $x_i \geq 0, \forall i.$

- **Example 1:** Projection to nonnegative orthant R_+. Solve

$$\min \frac{1}{2}\|x - y\|^2, \quad \text{s.t. } x \geq 0$$

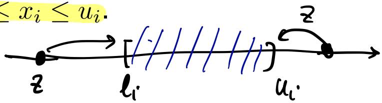
- **Solution** [graphically]



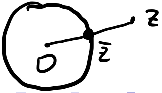
$$x_i^* = y_i, \quad \text{if } y_i \geq 0, \quad x_i^* = 0 \text{ otherwise}, \quad \forall\, i$$

or simiply denote $y = [y]^+$ (means taking non-negative part)

- **Example 1b:** Projection to bounds $l_i \leq x_i \leq u_i$.
  Answer:



- **Example 2:** Projection to ball $\|x\| \leq B$.
  Answer: $\bar{z} = z \cdot \frac{1}{\|z\|}$.



e.g. batch-normalization

Example: Nonnegative LS

- Nonnegative least square problem (we discussed it for CD methods)

$$\min \ \frac{1}{2}\|Ax - b\|^2, \quad \text{s.t.. } x \geq 0$$

CD can solve it

- Lots of practical applications, especially useful when dealing with nonnegative data
- Gradient projection?

$$x^{r+1} = \text{proj}_{x\geq 0}\left[\underbrace{x^r - \frac{1}{L}(A^T(Ax^r - b))}_{\text{denote as } y}\right]$$

Another example: NMF: $\min_{X,Y\geq 0} \|M - XY^T\|_F^2$.

Limitation of GP

- Common **misconception** (by many non-optimizers): constraints are not scary, just do projection.
- No! Constraints are often scary!
- GP is VERY restricted.

  $\min_x \|z-x\|^2$ s.t. $x \in X$. $= \mathrm{Proj}_X(z)$.

  - In general, solve a subproblem to find projection; often expensive
  - Only practical for very simple constraints: bounds, simplex, one ball

    $\min_x \|z-x\|^2$ s.t. $Ax=b$.
  - Linear constraints $Ax = b$? Closed-form projection (inverting matrix), but expensive for large dimension!
  - Two constraints $Ax = b, x \geq 0$? Solve a quadratic programming!

    $\min \|z-x\|^2$ s.t. $Ax=b, x \geq 0$.

Another example:

$\min f(x)$ s.t. $\|x\| \leq 1$, $\|x_i\|^2 \leq \frac{1}{10}$. $\forall i$. → not easy to do GP.

Convergence Analysis of GP Methods

- The first two results are for GP1.
- **Result 1** (constant stepsize): Assume f has L-Lipschitz gradient. If $\alpha_r = 1$, and $s_r = s \in (0, 2/L)$, then every limit point of the GP iterates is stationary. [Prop. 2.3.2 in book 1999]

  Result 2 (Armijo s): Fix s, if α_r is chosen by the limited minimization rule or by the Armijo rule along the feasible direction, every limit point of $\{x^r\}$ is stationary; [Prop. 2.3.3 in book 1999]

- The last result is for GP2.

- **Result 3** (fix s): Fix s, if α_r is chosen by the limited minimization rule or by the Armijo rule along the feasible direction, every limit point of $\{x^r\}$ is stationary; [Prop. 2.3.1 in book 1999]

Convergence Rate Analysis (Optional)

- Consider a strongly convex quadratic function $f(x) = \frac{1}{2}x'Ax + b'x$, with $A \succ 0$.

- $\exists$ a unique solution $x^* \in X$ satisfying $x^* = \text{proj}_X[x^* - s\nabla f(x^*)]$ (why?),so

$$\begin{aligned}
\| x^{r+1} - x^* \| &= \| \text{proj}_X[x^r - s\nabla f(x^r)] - \text{proj}_X[x^* - \alpha^r \nabla f(x^*)] \| \\
&\leq \| (x^r - x^*) - s(\nabla f(x^r) - \nabla f(x^*)) \| \\
&= \| (I - sA)(x^r - x^*) \| \\
&\leq \max\{1 - s\lambda_{\min}, 1 - s\lambda_{\max}\} \| x^r - x^* \| \\
&\leq (\frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}) \| x^r - x^* \| = (1 - \frac{2}{\kappa + 1}) \| x^r - x^* \| .
\end{aligned}$$

  In the last inequality we choose $s = \frac{2}{m+M}$.

- Convergence rate depends on $\kappa = \lambda_{max}/\lambda_{min}$, but *independent* of dimension.

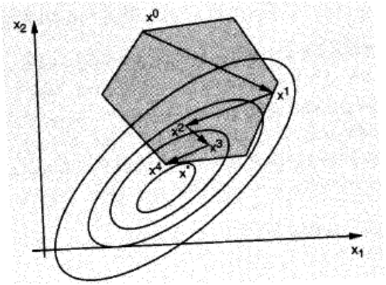
- Requires $O(1)\kappa\ln(1/\epsilon)$ to find ϵ-relative optimal solution.

Feasible Directions Method (Optional)

- A feasible direction method:

$$x^{r+1} = x^r + \alpha_r d^r,$$

where d^r: feasible descent direction, i.e., $\nabla f(x^r)' d^r < 0$, and $\alpha_r > 0$ is such that $x^{r+1} \in X$.



Feasible Directions Method (Optional)

- Alternative definition:

$$x^{r+1} = x^r + \alpha_r(\overline{x}^r - x^r)$$

where $\alpha_r \in (0,1]$, $\bar{x}^r$ is some feasible point. If x^r is nonstationary,

$$x^r \in X, \quad \nabla f(x^r)'(\overline{x}^r - x^r) < 0.$$

- Stepsize rules: Limited minimization, Constant $\alpha_r = 1$, Armijo: $\alpha_r = \beta^{m_r} s$, where m_r is the first nonnegative m for which

$$f(x^r) - f(x^r + \beta^m(\overline{x}^r - x^r)) \geq -\sigma\beta^m \nabla f(x^r)'(\overline{x}^r - x^r),$$

Convergence Analysis (Optional)

SGD for constraint $x_i \geq 0$.

?

- Similar to the one for (unstrained) gradient methods.
- The direction sequence $\{d^r\}$ is gradient related to $\{x^r\}$ if the following property can be shown: For any subsequence $\{x^r\}_{r\in K}$ that converges to a nonstationary point, the corresponding subsequence $\{d^r\}_{r\in K}$ is bounded and satisfies

$$\limsup_{r\to\infty, r\in K} \nabla f(x^r)' d^r < 0.$$

- **Proposition (Stationary of Limit Points)** Let $\{x^r\}$ be a sequence generated by the feasible direction method $x^{r+1} = x^r + \alpha_r d^r,$. Assume that:
  ⋆ $\{d^r\}$ is gradient related
  ⋆ α_r is chosen by the limited minimization rule or the Armijo rule.
  Then every limit point of $\{x^r\}$ is a stationary point.
- Proof is nearly identical to the unconstrained case.

Summary

In this lecture, we learned the following:

- Constrained optimization: various types
- Optimality condition of constrained optimization
- Gradient projection method
  - When projection is easy
  - Convergence theory

Summary

In this lecture, we learned the following:

- Constrained optimization: various types
- Optimality condition of constrained optimization
- Gradient projection method
  - When projection is easy
  - Convergence theory