DDA3020 Machine Learning
Lecture 04 Basic Optimization

JIA, Kui
School of Data Science, CUHK-SZ

September 21, 2023

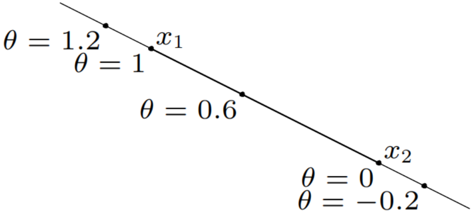
Outline

1. Convex set
2. Convex function
3. Convex optimization problem
4. Unconstrained minimization: gradient descent method
5. Constrained minimization: Lagrangian duality, KKT conditions
6. Optimization and machine learning

Affine set

- The **Affine line** through $\mathbf{x}_1, \mathbf{x}_2$: all points

$$\mathbf{x} = \theta \mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \ (\theta \in \mathbb{R})$$



- The **Affine set** contains the line through any two distinct points in the set.
- **Example**: solution set of linear equations $\{\mathbf{x} | \mathbf{A}\mathbf{x} = \mathbf{b}\}$
(conversely, every affine set can be expressed as solution set of system of linear equations)

Convex set

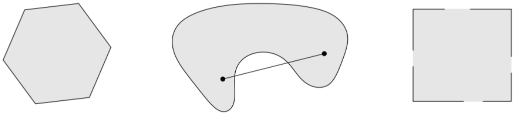
- The **line segment** between $\mathbf{x}_1, \mathbf{x}_2$: all points

$$\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$$

  with $0 \leq \theta \leq 1$
- The **convex set** contains line segment between any two points in the set.

$$\mathbf{x}_1, \mathbf{x}_2 \in C,\ 0 \leq \theta \leq 1 \quad \rightarrow \quad \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in C$$

- **Examples**: (one convex, two nonconvex sets)



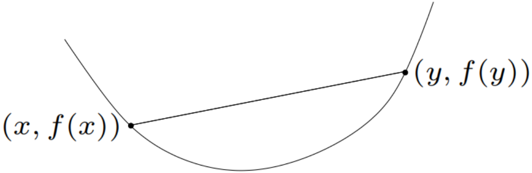
1. Convex set
2. Convex function
3. Convex optimization problem
4. Unconstrained minimization: gradient descent method
5. Constrained minimization: Lagrangian duality, KKT conditions
6. Optimization and machine learning

Convex function definition

- $f : \mathbb{R}^n \to \mathbb{R}$ is convex if $\mathbf{dom} f$ is a convex set and

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f, \ 0 \leq \theta \leq 1$



- f is concave if $-f$ is convex
- f is strictly convex if $\mathbf{dom} f$ is convex and

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

for $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f, \ \mathbf{x} \neq \mathbf{y}, \ 0 < \theta < 1$

Examples on $\mathbb{R}$

Convex:

- affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
- exponential: e^{ax}, for any $a \in \mathbb{R}$
- powers: x^α on $\mathbb{R}_+$, for $\alpha \geq 1$ or $\alpha \leq 0$
- powers of absolute value: $|x|^p$ on $\mathbb{R}$, for $p \geq 1$
- negative entropy: $x \log x$ on $\mathbb{R}_+$

Concave:

- affine: $ax + b$ on $\mathbb{R}$, for any $a, b \in \mathbb{R}$
- powers: x^α on $\mathbb{R}_+$, for $0 \leq \alpha \leq 1$
- logarithm: $\log x$ on $\mathbb{R}_+$

Examples on $\mathbb{R}^n$ and $\mathbb{R}^{m\times n}$

Affine functions are convex and concave

Examples on $\mathbb{R}^n$

- Affine function $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x} + b$
- ℓ_p norms: $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ for $p \geq 1$; $\|\mathbf{x}\|_\infty = \max_k |x_k|$
- *All norms are convex functions, and the above is obtained by checking the convexity of the domain*

Examples on $\mathbb{R}^{m\times n}$ ($m \times n$ matrices)

- Affine function

$$f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{A}^\top \mathbf{X}\right) + b = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}x_{ij} + b,$$

  where $\mathrm{tr}(\cdot)$ indicates the trace norm, *i.e.*, the summation of all diagonal values of a matrix
- Spectral (maximum singular value) norm

$$f(\mathbf{X}) = \|\mathbf{X}\|_2 = \sigma_{\max}(\mathbf{X}) = \left(\lambda_{\max}\left(\mathbf{X}^\top \mathbf{X}\right)\right)^{1/2}$$

First-order condition of convex function

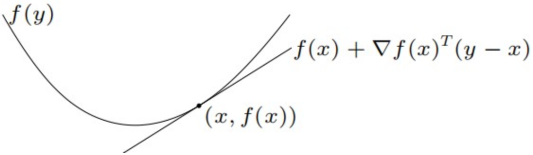
f is **differentiable** if **dom** f is open and the gradient

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \frac{\partial f(\mathbf{x})}{\partial x_2}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right)$$

exists at each $\mathbf{x} \in$ **dom** f

1st-order condition: differentiable f with convex domain is convex iff

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \text{dom}\, f$$



First-order approximation of f is global underestimator

Second-order conditions of convex function

f is **twice differentiable** if **dom** f is open and the Hessian $\nabla^2 f(\mathbf{x}) \in \mathbf{S}^{nn}$,

$$\nabla^2 f(\mathbf{x})_{ij} = \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j}, \quad i, j = 1, \ldots, n,$$

exists at each $\mathbf{x} \in \mathbf{dom}\ f$

2nd-order conditions: for twice differentiable f with convex domain

- f is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq 0 \quad \text{for all } \mathbf{x} \in \mathbf{dom}\ f$$

- if $\nabla^2 f(\mathbf{x}) \succ 0$ for all $\mathbf{x} \in \mathbf{dom}\ f$, then f is strictly convex
- Note that $\succeq$ indicates positive semi-definite, and $\succ$ indicates positive definite.
- *Note: A square matrix* $\mathbf{W}$ *is positive semi-definite if* $\mathbf{x}^\top \mathbf{W} \mathbf{x} \geq \mathbf{0}$ *for any compatiable* $\mathbf{x}$ *or if all the eigenvalues of* $\mathbf{W}$ *are non-negative.*

Examples

Quadratic function: $f(\mathbf{x}) = (1/2)\mathbf{x}^\top \mathbf{P}\mathbf{x} + \mathbf{q}^\top \mathbf{x} + r$ (with $\mathbf{P} \in \mathbf{S}^{n\times n}$)

$$\nabla f(\mathbf{x}) = \mathbf{P}\mathbf{x} + \mathbf{q}, \quad \nabla^2 f(\mathbf{x}) = \mathbf{P}$$

convex if $\mathbf{P} \succeq 0$

Least-squares objective: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$

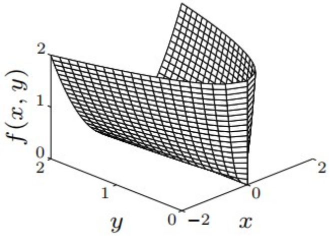
$$\nabla f(\mathbf{x}) = 2\mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b}), \quad \nabla^2 f(\mathbf{x}) = 2\mathbf{A}^\top \mathbf{A}$$

convex (for any $\mathbf{A}$)

Quadratic-over-linear: $f(x, y) = x^2/y$

$$\nabla^2 f(x,y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^\top \succeq 0$$

convex for $y > 0$



Jensen's inequality

Basic inequality: if f is convex, then for $0 \le \theta \le 1$,

$$f(\theta \mathbf{x} + (1-\theta)\mathbf{y}) \le \theta f(\mathbf{x}) + (1-\theta) f(\mathbf{y})$$

Extension: if f is convex, then

$$f(\mathbf{E}[\mathbf{z}]) \le \mathbf{E}[f(\mathbf{z})]$$

for any random variable $\mathbf{z}$

Optimization problem in standard form

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

- $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable
- $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective or cost function
- $f_i : \mathbb{R}^n \to \mathbb{R}, i = 1, ..., m$, are the inequality constraint functions
- $h_i : \mathbb{R}^n \to \mathbb{R}$ are the equality constraint functions

Optimal objective value

Optimal objective value:

$$p^* = \inf\{f_0(\mathbf{x}) | f_i(\mathbf{x}) \leq 0, i = 1, ..., m,\ h_i(\mathbf{x}) = 0, i = 1, ..., p\},$$

where $\inf\{\mathcal{S}\}$ indicates the infimum of the set $\mathcal{S}$, *i.e.*, **greatest lower bound**.

Properties:

- $p^* = \infty$ if problem is infeasible (no $\mathbf{x}$ satisfies the constraints)
- $p^* = -\infty$ if problem is unbounded below

Reference:
https://en.wikipedia.org/wiki/Infimum_and_supremum

Optimal and locally optimal points

Feasible point: $\mathbf{x}$ is **feasible** if $\mathbf{x} \in \mathbf{dom} f_0$ and it satisfies the constraints

Optimal point: A feasible $\mathbf{x}$ is **optimal** if $f_0(\mathbf{x}) = p^*$; X_{opt} is the set of optimal points

Locally optimal point: $\mathbf{x}$ is **locally optimal** if there is an $r > 0$ such that $\mathbf{x}$ is optimal for

$$\begin{array}{ll} \text{minimize}_{\mathbf{z}} & f_0(\mathbf{z}) \\ \text{subject to} & f_i(\mathbf{z}) \leq 0,\ i = 1, \ldots, m, \quad h_i(\mathbf{z}) = 0,\ i = 1, \ldots, p, \\ & \|\mathbf{z} - \mathbf{x}\|_2 \leq r \end{array}$$

Examples (with $n = 1, m = p = 0$)

- $f_0(x) = 1/x, \text{dom}\, f_0 = \mathbb{R}_+ : p^\star = 0$, no optimal point
- $f_0(x) = -\log x, \text{dom}\, f_0 = \mathbb{R}_+ : p^\star = -\infty$
- $f_0(x) = x \log x, \text{dom}\, f_0 = \mathbb{R}_+ : p^\star = -1/e, x = 1/e$ is optimal
- $f_0(x) = x^3 - 3x, p^\star = -\infty$, local optimum at $x = 1$

Implicit constraints

The standard form optimization problem has an **implicit constraint**

$$\mathbf{x} \in \mathcal{D} = \bigcap_{i=0}^{m} \operatorname{dom} f_i \cap \bigcap_{i=1}^{p} \operatorname{dom} h_i,$$

- We call $\mathcal{D}$ the **domain** of the problem
- The constraints $f_i(\mathbf{x}) \leq 0, h_i(\mathbf{x}) = 0$ are the explicit constraints
- A problem is **unconstrained** if it has no explicit constraints $(m = p = 0)$

Example:

$$\text{minimize } f_0(\mathbf{x}) = -\sum_{i=1}^{k} \log\left(b_i - \mathbf{a}_i^\top \mathbf{x}\right)$$

is an unconstrained problem with implicit constraints $\mathbf{a}_i^\top \mathbf{x} < b_i$

Convex optimization problem

Standard form convex optimization problem

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

- $f_0, f_1, \ldots, f_m$ are convex; equality constraints are affine

It is often written as

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

Important property: feasible set of a convex optimization problem is convex

Convex optimization problem

Example

$$
\begin{array}{ll}
\text{minimize} & f_0(\mathbf{x}) = x_1^2 + x_2^2 \\
\text{subject to} & f_1(\mathbf{x}) = x_1 / \left(1 + x_2^2\right) \leq 0 \\
 & h_1(\mathbf{x}) = \left(x_1 + x_2\right)^2 = 0
\end{array}
$$

- f_0 is convex; feasible set $\{(x_1, x_2) \mid x_1 = -x_2 \leq 0\}$ is convex
- *Originally*, not a convex problem (according to our definition): f_1 is not convex, h_1 is not affine
- Equivalent (but not identical) to the convex problem

$$
\begin{array}{ll}
\text{minimize} & x_1^2 + x_2^2 \\
\text{subject to} & x_1 \leq 0 \\
 & x_1 + x_2 = 0
\end{array}
$$

Local and global optima of the convex problem

Theorem: Any locally optimal point of a convex problem is globally optimal

Proof:

Step 1: suppose $\mathbf{x}$ is locally optimal, but there exists a feasible $\mathbf{y}$ with

$$f_0(\mathbf{y}) < f_0(\mathbf{x}) \tag{1}$$

And, $\mathbf{x}$ locally optimal means there is a $r > 0$ such that

$$\mathbf{z} \text{ is feasible}, \quad \| \mathbf{z} - \mathbf{x} \|_2 \leq r \quad \Rightarrow \quad f_0(\mathbf{z}) \geq f_0(\mathbf{x}) \tag{2}$$

Step 2: we construct that

$$\mathbf{z} = \theta \mathbf{y} + (1 - \theta)\mathbf{x} \text{ with } \theta = r/(2 \| \mathbf{y} - \mathbf{x} \|_2) \tag{3}$$

If we set $\| \mathbf{y} - \mathbf{x} \|_2 = 1.5r$, then we have $\| \mathbf{z} - \mathbf{x} \|_2 = 0.5r$. It implies that $\mathbf{y}$ is out of the local region of $\mathbf{x}$, while $\mathbf{z}$ is within the local region.

Step 3: According to the basic property of convex function, we have

$$f_0(\mathbf{z}) \leq \theta f_0(\mathbf{y}) + (1 - \theta) f_0(\mathbf{x}) < \theta f_0(\mathbf{x}) + (1 - \theta) f_0(\mathbf{x}) = f_0(\mathbf{x}),$$

where the second $<$ utilizes (1), which contradicts our assumption that $\mathbf{x}$ is locally optimal, *i.e.*, (2). It means that there doesn't exist a feasible $\mathbf{y}$ to satisfy (1), thus $\mathbf{x}$ is also globally optimal

Unconstrained convex minimization

Unconstrained convex minimization problem

$$\text{minimize } f(\mathbf{x})$$

- f convex, twice continuously differentiable (hence **dom** f open)
- We assume optimal value $p^\star = \inf_{\mathbf{x}} f(\mathbf{x})$ is attained (and finite)

Unconstrained convex minimization methods

- Produce sequence of points $\mathbf{x}^{(k)} \in \textbf{dom } f, k = 0, 1, \ldots$ with

$$f(\mathbf{x}^{(k)}) \to p^\star$$

General descent Method

One step update of general descent method:

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)} \Delta \mathbf{x}^{(k)} \text{ with } f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$$

- $\Delta \mathbf{x}$ is the search direction; t is the step size
- We also define the notation $\mathbf{x}^{+} = \mathbf{x} + t\Delta \mathbf{x}$
- Recall **1st-order condition** of convex function,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\top}(\mathbf{y} - \mathbf{x}) \quad \text{for all } \mathbf{x}, \mathbf{y} \in \text{dom } f$$

  Thus, we have

$$f(\mathbf{x}^{+}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^{\top}(\mathbf{x}^{+} - \mathbf{x}) = f(\mathbf{x}) + t\nabla f(\mathbf{x})^{\top} \Delta \mathbf{x}$$

- If $f(\mathbf{x}^{+}) < f(\mathbf{x})$, then it implies $\nabla f(\mathbf{x})^{\top} \Delta \mathbf{x} < 0$, *i.e.*, $\Delta \mathbf{x}$ is a descent direction

General descent Method

General descent method

Given a starting point $\mathbf{x} \in \mathbf{dom} f$.
repeat

1. Determine a descent direction $\Delta\mathbf{x}$
2. Choose a step size $t > 0$, such as using *Line search method*
3. **Update**. $\mathbf{x} := \mathbf{x} + t\Delta\mathbf{x}$.

until stopping criterion is satisfied.

Line search method

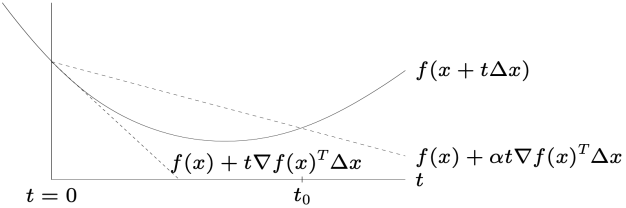
Exact line search: $t = \arg\min_{t>0} f(\mathbf{x} + t\Delta\mathbf{x})$

Backtracking line search (inexact) (with parameters $\alpha \in (0, 1/2), \beta \in (0, 1)$)

- Starting at $t = 1$, repeat $t := \beta t$ until

$$f(\mathbf{x} + t\Delta\mathbf{x}) < f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^\top \Delta\mathbf{x}$$

- Graphical interpretation: backtrack until $t \le t_0$



Gradient descent method

General descent method with $\Delta \mathbf{x} = -\nabla f(\mathbf{x})$ is called gradient descent method

---

Given a starting point $\mathbf{x} \in \mathbf{dom} f$.
repeat
1. $\Delta \mathbf{x} := -\nabla f(\mathbf{x})$.
2. Choose step size t via exact or backtracking line search
3. **Update**. $\mathbf{x} := \mathbf{x} + t\Delta \mathbf{x}$.

until stopping criterion is satisfied.

- Stopping criterion usually of the form $\|\nabla f(\mathbf{x})\|_2 \leq \epsilon$
- Note that although here we consider the convex minimization problem, gradient descent and its variants (*e.g.*, stochastic gradient descent) can also be directly applied to solve non-convex optimization problem, such as training deep neural networks
- In this course, gradient descent method will be used in linear regression and logistic regression

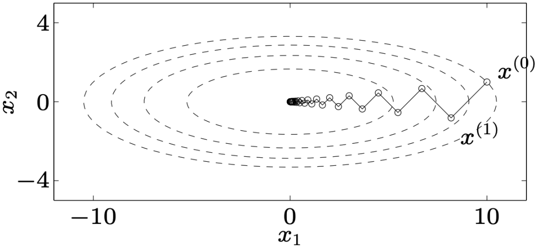
Example: quadratic problem in $\mathbb{R}^2$

$$\min_{\mathbf{x}} f(\mathbf{x}) = (1/2)(x_1^2 + \gamma x_2^2),$$

where $\gamma > 0$. Solve the above problem using gradient descent with exact line search, starting at $\mathbf{x}^{(0)} = (\gamma, 1)$, we can derive the following update:

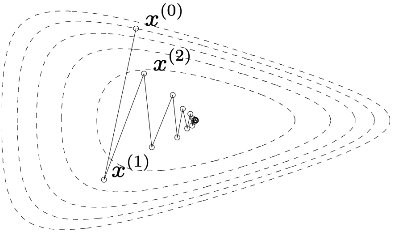
$$x_1^{(k)} = \gamma \left(\frac{\gamma - 1}{\gamma + 1}\right)^k, \qquad x_2^{(k)} = \left(-\frac{\gamma - 1}{\gamma + 1}\right)^k$$

- very slow if $\gamma \gg 1$ or $\gamma \ll 1$
- example for $\gamma = 10$:

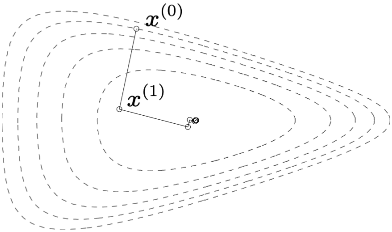


Example: non-quadratic example

$$\min_{x_1,x_2} f(x_1,x_2) = e^{x_1+3x_2-0.1} + e^{x_1-3x_2-0.1} + e^{-x_1-0.1}$$



backtracking line search



exact line search

Constrained minimization and Lagrange duality

- Given a general minimization problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, \quad i = 1, \dots, m \\ & \ell_j(\mathbf{x}) = 0, \quad j = 1, \dots, r \end{aligned}$$

- The **Lagrangian function**:

$$L(\mathbf{x}, \mathbf{u}, \mathbf{v}) = f(\mathbf{x}) + \sum_{i=1}^{m} u_i h_i(\mathbf{x}) + \sum_{j=1}^{r} v_j \ell_j(\mathbf{x})$$

- The **Lagrange dual function**:

$$g(\mathbf{u}, \mathbf{v}) = \min_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x}, \mathbf{u}, \mathbf{v})$$

- The **dual problem** (an easier, convex problem w.r.t.$\mathbf{u}$ and $\mathbf{v}$):

$$\max_{\mathbf{u}\in\mathbb{R}^m, \mathbf{v}\in\mathbb{R}^r} g(\mathbf{u}, \mathbf{v})$$

$$\text{subject to } \mathbf{u} \geq 0$$

- Let $p^* = \min f(\mathbf{x})$ and $d^* = \max g(\mathbf{u}, \mathbf{v})$, by definition we have $p^* \geq d^*$.

KKT conditions

- Given general problem

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^n} \quad & f(\mathbf{x}) \\ \text{subject to} \quad & h_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & \ell_j(\mathbf{x}) = 0, \quad j = 1, \ldots, r \end{aligned}$$

- The **Karush-Kuhn-Tucker conditions** or **KKT conditions** are:
  - $0 \in \partial f(\mathbf{x}) + \sum_{i=1}^{m} u_i \partial h_i(\mathbf{x}) + \sum_{j=1}^{r} v_j \partial \ell_j(\mathbf{x})$ (stationarity)
  - $u_i \cdot h_i(\mathbf{x}) = 0$ for all i (complementary slackness)
  - $h_i(\mathbf{x}) \leq 0, \ell_j(\mathbf{x}) = 0$ for all i, j (primal feasibility)
  - $u_i \geq 0$ for all i (dual feasibility)
- *Note: For a convex problem, if* $\mathbf{x}, \mathbf{u}, \mathbf{v}$ *satisfy the KKT conditions, then they are optimal.*

Note: Lagrangian function and KKT conditions will be used later in support vector machines, K-means Gaussian mixture models, and principal component analysis in this course

Optimization and machine learning

Optimization is one of the basis techniques in machine learning:

- Convex minimization will be directly utilized in linear regression, logistic regression, support vector machine in this course
- Gradient descent method will be adopted to solve linear regression, logistic regression and neural networks
- Lagrangian function and KKT conditions will be adopted to solve support vector machine, K-means, Gaussian mixture models, and principal component analysis

Given the objective function and constraints of a machine learning model, you should be able to determine

- whether it is convex or non-convex optimization problem
- whether there is local or global optima
- which optimization method could be adopted to solve the problem

Acknowledgment

Credit to Professor Stephen Boyd, Stanford University.

https://web.stanford.edu/class/ee364a/lectures.html