2021 ZJU-CSE Summer School

Lecture I: Introduction & Convexity

Jinming Xu

Zhejiang University

August 02, 2021

Course Goals and Evaluation

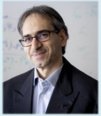
- Goal of the course
  - Prepare graduate students with advanced distributed control, optimization and learning methods for large-scale networked systems arising from modern control engineering and data science
- Topics
  - Distributed Control and Estimation; Intelligent Autonomous Systems; Distributed Convex Opitmization; Acceleration Methods and ADMM; Distributed Stochastic Optimization; Distributed Learning in Non-convex World
- Reference Books
  - There is no required specific textbook. All course mateirals will be presented in class and will be available online as notes.
  - Stephen Boyd and Lieven Vandenberghe. Convex optimization. Cambridge university press, 2004.
    Avaiable for free at https://web.stanford.edu/ boyd/cvxbook/
- Evaluation
  - A certificate will be granted after completion of 80% of the course

Course Content

- Lectures (First Week, Aug 02-06)
  - Convex Optimization
  - Graph Basics and Consensus
  - (Distributed) Stochastic Optimization
  - Operator Splitting and ADMM
  - Acceleration methods
- Seminar/Tutorials (Second Week, Aug 09-13)
  - Distributed Convex Optimization
  - Statistical Inference over Networks
  - Distributed Stochastic Nonconvex Optimization
  - Intelligent Unmanned Systems
  - Distributed Load Frequency Control in Smart Grids

Invited Speakers[1]

Prof. Lihua Xie
(IEEE/IFAC Fellow)
Nanyang Technological University



Prof. Gesualdo Scutari
(IEEE Fellow)
Purdue University

Prof. Angelia Nedich
Arizona State University

Prof. Usman Khan
Tufts University

Prof. Shichao Liu
Carleton University

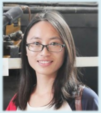
Prof. Ying Sun
Pennsylvania State University

Prof. Hoi To Wai
Chinese University of Hong Kong

Prof. César Uribe
Rice University

Prof. Wenchan Meng
Zhejiang University

Prof. Huan Li
Nankai University



Prof. Chengcheng Zhao
Zhejiang University

Dr. Kun Yuan
Damo Academy (达摩院)

[1]Refer to jinmingxu.github.io for more details.

Time Schedule for Lectures[2]

Week One (Aug 02 - Aug 06)

Time	Monday (Aug 02)	Tuesday(Aug 03)	Wednesday(Aug 04)	Thursday(Aug 05)	Friday(Aug 06)
8.30 am - 12.00 pm (GMT+8)	Lecture I **Introduction to the course** Speaker Jinming Xu, ZJU （Yuquan Campus） **Tencent Meeting ID:** 752 593 984	Lecture II **Convex Optimization** Speaker Ying Sun, PSU （online） **Tencent Meeting ID:** 755 843 514	Lecture IV **Distributed Convex Optimization** Speaker Ying Sun, PSU （online） **Tencent Meeting ID:** 642 952 965	Lecture V **Stochastic Optimization** Speaker Ying Sun, PSU （online） **Tencent Meeting ID:** 708 668 998	Lecture VIII **Advanced Topics(Operator Splitting, ADMM)** Speaker Jinming Xu, ZJU （Yuquan Campus） **Tencent Meeting ID:** 479 145 622
12.00 pm - 2.30 pm (GMT+8)	Lunch Break	Lunch Break	Lunch Break	Lunch Break	Lunch Break
2.30 pm - 5.30 pm (GMT+8)	**Lab Tour** Shining Gao/Anjun Chen (Yuquan Campus)	Lecture III **Graph Basics and Consensus** Speaker Prof Chengcheng Zhao, ZJU （Yuquan Campus） **Tencent Meeting ID:** 624 997 500	Research & Discussion	Lecture VI **Distributed Stochastic Optimization** Speaker Kun Yuan, Damo Academy (Yuquan Campus) **Tencent Meeting ID:** 202 219 103	Lecture IX **Advanced Topics(Acceleration)** Speaker Huan Li, NKU (Yuquan Campus) **Tencent Meeting ID:** 962 195 511

[2] Refer to jinmingxu.github.io for more details.

Time Schedule for Tutorial/Seminar[3]

Week Two (Aug 09 - Aug 13)

Time	Monday (Aug 09)	Tuesday(Aug 10)	Wednesday(Aug 11)	Thursday(Aug 12)	Friday(Aug 13)
8.30 am - 12.00 pm (GMT+8)	T/S I Prof. Usman Khan, Tufts Univ T/S II Prof. Hoi To Wai, CUHK	T/S III Prof Cesar Uribe, Rice T/S IV Prof Angelia Nedich, ASU	T/S V Part I/Part II Prof Gesualdo Scutari, Purdue	T/S VI Prof Shichao liu, Carleton T/S VII Prof Xie Lihua, NTU	Group Sharing & Discussion
12.00 pm - 2.30 pm (GMT+8)	Lunch Break	Lunch Break	Lunch Break	Lunch Break	Lunch Break
2.30 pm - 5.30 pm （GMT+8）	Lecture VII **Distributed Control** Speaker Prof Meng Wenchao, ZJU （Yuquan Campus）	Research & Discussion	Research & Discussion	Research & Discussion	

[3]Refer to jinmingxu.github.io for more details.

Outline for Lecture I

Introduction
- Optimization problems
- Convex sets and functions
- Operations that preserve convexity
- Convex problem and first-order optimality

Duality
- Weak duality
- Strong Duality

KKT conditions

Summary

Outline for Lecture I

Structure of Optimization Problems

- (Mathematical) optimization problem

$$\begin{aligned} p^\star := \min_{x \in \mathbb{R}^n} \ & f(x) \\ \text{subject to } & h_i(x) \le 0, \ i = 1, 2, ..., m, \\ & l_j(x) = 0, \ j = 1, 2, ..., r \end{aligned}$$

  - $x := [x_1, ..., x_n]^T$: optimization variables
  - $f : \mathbb{R}^n \to \mathbb{R}$: objective function
  - $h_i, l_j : \mathbb{R}^n \to \mathbb{R}$: constraint functions
- Feasible solution set (assume $\mathbf{dom}\, f = \mathbb{R}^n$)

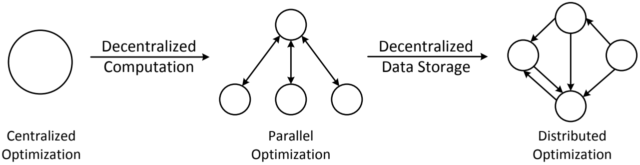
$$\mathcal{X} := \{x | h_i(x) \le 0, \ i = 1, 2, ..., m, l_j(x) = 0, \ j = 1, 2, ..., r\}$$

- Algorithms solving the above problem
  - first order primal (dual) methods, second order methods,...

The Goal is to find a point that minimizes f among all feasible points

A brief history of optimization

- Theory (convex analysis): back to 1900s
- Algorithms [B & V 2004],
  - 1940s: simplex algorithm for linear programming (Dantzig)
  - 1970s: subgradient methods; proximal point method (Rockafellar,...)
  - 1980s: polynomial-time interior-point methods (Karmarkar, Nesterov & Nemirovski)
  - 1990s-now: accelerated method; parallel and distributed methods
- Applications
  - before 1990: mostly in operations research; few in enginneing
  - since 1990: many new applications in engineering, such as control, signal processing, communicaitons, and machine learning...
- Structure: from centralized to distributed (2010s-now)



Examples: l_1-regularized least square problem

- Measurement Model

$$y = Mx + v$$

  - $x \in \mathbb{R}^d$: the unknown parameter assumed to be sparse
  - $M \in \mathbb{R}^{s \times d}$: measurement matrix
  - $v \in \mathbb{R}^s$: measurement noise
  - $y \in \mathbb{R}^s$: the observation of a sensor

- Least square problem for a sensor

$$\min_{x \in \mathbb{R}^d} \|y - Mx\|^2 + \|x\|_1$$

  - $\|\cdot\|$ encoding the sparsity,
  - arising from compressive sensing, image processing, etc.

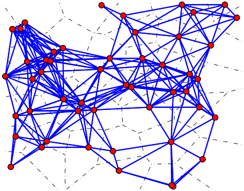


Figure: A sensor network of 50 nodes

- Distributed Estimation

$$\min_{x \in \mathbb{R}^d} \sum_{i=1}^{m} \|y_i - M_i x\|^2 + \|x\|_1$$

how to solve it when there is no center knowing all M_i, y_i?

Examples: Support Vector Machine (SVM)

Consider m training samples $(x_1, y_1), ..., (x_m, y_m)$ with $y_i \in \{-1, +1\}$

- Look for a separating hyperplane $\{x \in \mathbb{R}^d | w^T x + b = 0\}$ such that

$$\begin{cases} w^T x_i + b > 0, \ \forall i \text{ such that } y_i = +1, \\ w^T x_i + b < 0, \ \forall i \text{ such that } y_i = -1 \end{cases}$$

- The min. point-to-hyperplane distance

$$d = \min_i \frac{|w^T x_i + b|}{\|w\|}$$

  - scaling w, b such that $d\|w\| = 1$

- Want to solve the following problem

$$\max_{\{w,b\}} d = \frac{1}{\|w\|} \quad (\text{or } \min_w (1/2)\|w\|^2)$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1, \ i = 1, 2, ..., N$$

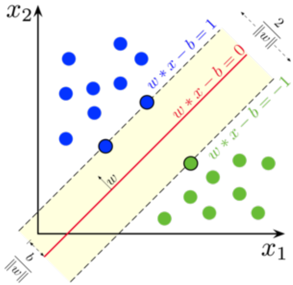


Figure: A hyperplane that separates the positive samples from negative ones (from Google)

How to solve it when the data set is distributed across several data centers?

Examples: Economic Dispatch of Power Systems

- Economic Dispatch Problem

$$\min_{\mathbf{p}\in\mathbb{R}^m} C(\mathbf{p}) = \sum_{i=1}^{m} C_i(p_i)$$

$$\textbf{s.t.} \sum_{i=1}^{m} p_i = \sum_{i=1}^{m} l_i,\ \underline{p}_i \le p_i \le \overline{p}_i,\ \forall i \in \mathcal{V}.$$

  - p_i: power generation of bus i,
  - l_i: the load demand from bus i,
  - $\underline{p}_i, \overline{p}_i$: capacity limit of bus i.

- Power generation model

$$C_i(p_i) = a_i p_i^2 + b_i p_i + c_i,$$

  - a_i, b_i, c_i are some coefficients related to bus i.

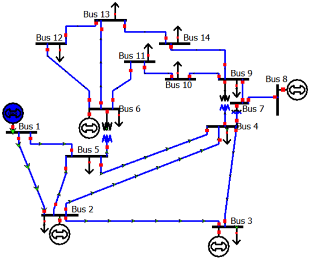


Figure: IEEE 14-Bus System[4]

$\boxed{\text{how to solve it when there is no center knowing all } C_i\text{?}}$

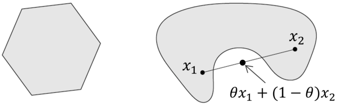
[4]more details at http://icseg.iti.illinois.edu/ieee-14-bus-system/

Outline for Lecture I

Convex Sets and Functions

- **Convex set**: for any $x_1, x_2 \in \mathcal{C}$ and any $\theta \in [0\ 1]$, we have

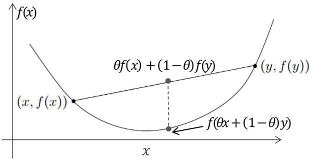
$$\boxed{\theta x_1 + (1-\theta)x_2 \in \mathcal{C}}$$



  - examples: $\mathcal{S} := \{x | Ax = b\}$ or $\mathcal{S} := \{x | Ax \preceq b\}$

- **Convex function**: for all $x, y \in \mathbb{R}^n$, and any $\theta \in [0\ 1]$, we have

$$\boxed{\begin{aligned} & f(\theta x + (1-\theta)y) \\ & \quad \leq \theta f(x) + (1-\theta) f(y) \end{aligned}}$$

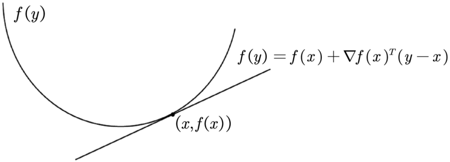


  - examples: $x^2, e^x, -\log x, x \log x$
  - f is concave if $-f$ is convex.

Convex Sets and Functions

- **First-order condition** for differentiable f

f is convex if and only if **dom** f is convex and
$f(y) \geq f(x) + \nabla f(x)^T(y - x)$



- **Second-order condition** for twice differentiable f

f is convex if and only if **dom** f is convex and its Hessian
is positive semi-definite, i.e., for all $x \in$ **dom** f, $\nabla^2 f(x) \geq 0$

  - Example: $f(x) := (1/2)x^T Px + q^T x + r$
  - f is convex if and only if $P \succeq 0$

Jesen's Inequality

Lemma(Jesen's Inequality): Let f be convex, $x_1, x_2, ..., x_m \in \mathbb{R}^n$ and $\lambda_1, \lambda_2, ..., \lambda_m \in \mathbb{R}^+$ such that $\sum_{i=1}^m \lambda_i = 1$. Then,
$f(\sum_{i=1}^m \lambda_i x_i) \leq \sum_{i=1}^m \lambda_i f(x_i)$

- For $m = 2$, the above reduces to convexity.
- Examples: Let x be a random variable and ϕ be a convex function. Then, $\phi(E[x]) \leq E[\phi(x)]$.

Outline for Lecture I

Operations that preserve convexity

Let Γ denotes the class of convex functions.

- **Nonnegative weighted sum**

$$f_1, f_2 \in \Gamma \Rightarrow w_1 f_1 + w_2 f_2 \in \Gamma$$

  - negative entropy function: $\sum_i x_i \log x_i$
  - sparsity prior (l_1-norm): $\sum_i |x_i|$

- **Composition with affine function**

$$f \in \Gamma \Rightarrow f(Ax + b) \in \Gamma$$

  - quatratic function: $\|Ax + b\|^2$
  - log barrier function: $-\log(b - a^T x)$

Operations that preserve convexity

- **Pointwise maximum**

$$f_1, f_2 \in \Gamma \Rightarrow \max\{f_1, f_2\} \in \Gamma$$

  - piecewise linear function: $\max_i\{a_i x + b_i\}$
  - Nesterov test function: $\max_{1\le i\le d} x_i$

- **Pointwise maximum over a set**
  If f convex in x for each $z \in \mathcal{Z}$, then

$$g(x) := \max_{z\in\mathcal{Z}} f(x,z) \in \Gamma$$

  - support function: $\sigma_{\mathcal{C}}(x) = \max_{z\in\mathcal{C}} x^T z$
  - dual norm $\|x\|_* := \max_{\|z\|\le 1} x^T z$

Outline for Lecture I

Convex Optimization Problems

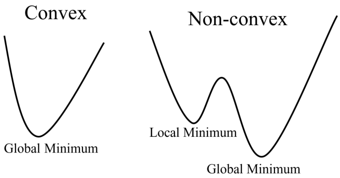
- Standard **convex optimization problem**

$$p^\star := \min_{x \in \mathbb{R}^n} \; f(x)$$

$$\text{subject to } h_i(x) \le 0, \; i = 1, 2, ..., m,$$
$$l_j(x) = 0, \; j = 1, 2, ..., r$$

Assumption 1:
The objective f and all $\{h_i\}, \{l_i\}$ are convex and the optimal value $p^\star$ is finite

- Why convexity?
  - can understand and solve a broad class of convex problems
  - nonconvex probems are mostly treated on a case-by-case basis



Special property: for a convex problem, local optima are global optima

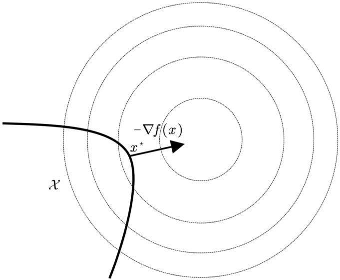
First-order optimality condition

- For a convex problem

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

and differentiable f, a feasible point x is optimal if and only if

$$\boxed{\nabla f(x)^T (y - x) \geq 0 \text{ for all } y \in \mathcal{X}}$$



- all feasible directions from x are aligned with gradient $\nabla f(x)$
- If $\mathcal{X} = \mathbb{R}^n$ (unconstrained optimization), then the first-order optimality condition reduces to

$$\boxed{\nabla f(x) = 0.}$$

Example: quadratic minimization

- Consider minimizing a quadratic problem as follows

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^T Q x + c^T x$$
$$\text{s.t. } Ax = b$$

  with $Q \succ 0$. The first-order optimality condition says that the solution $x^\star$ satisfies $Ax^\star = b$ and

$$\langle Qx^\star + c, y - x^\star \rangle \geq 0, \quad \forall y \text{ such that } Ay = b$$

  which is equivalent to

$$\langle Qx^\star + c, z \rangle \geq 0, \forall z \in \mathbf{null}\{A\}$$

- If the equality constraint is vacuous, the condition becomes

$$Qx^\star + c = 0, \quad \text{or namely,} \quad x^\star = -Q^{-1}c$$

Outline for Lecture I

Convex Optimization Problems

- Standard **convex optimization problem**

$$p^\star := \min_{x \in \mathbb{R}^n} \; f(x)$$
$$\text{subject to } h_i(x) \leq 0, \; i = 1, 2, ..., m,$$
$$l_j(x) = 0, \; j = 1, 2, ..., r$$

  - where the objective f and all $\{h_i\}, \{l_i\}$ are convex and the optimal value $p^\star$ is finite

- We define the **Lagrangian** as

$$L(x, \lambda, \nu) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \nu_j l_j(x)$$

  - λ_i is Lagrange multiplier associated with $h_i(x) \leq 0$
  - ν_i is Lagrange multiplier associated with $l_i(x) = 0$

Weak Duality

- **Lagarange dual function**

$$g(\lambda, \nu) := \min_x L(x, \lambda, \nu) = \min_x \left(f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \nu_j l_j(x) \right)$$

- **Lower bound property**:

$$\boxed{\text{If } \lambda \succeq 0, \text{ then } p^\star \geq g(\lambda, \nu)}$$

Remark: this always holds (even if primal problem is nonconvex)
proof: Let $\bar{x}$ be a feasible solution. Since $\lambda \succeq 0$, we have

$$\begin{aligned} f(\bar{x}) \geq f(\bar{x}) + \sum_{i=1}^m \underbrace{\lambda_i h_i(\bar{x})}_{\leq 0} + \sum_{i=1}^r \underbrace{\nu_i l_i(\bar{x})}_{=0} \\ = L(\bar{x}, \lambda, \nu) \geq \min_x L(x, \lambda, \nu) = g(\lambda, \nu) \end{aligned}$$

Then, minimizing over all feasible $\bar{x}$ gives $p^\star \geq g(\lambda, \nu)$

Weak Duality

- **Lagrange dual problem**

$$d^\star := \max_{\lambda,\nu}\ g(\lambda,\nu) = \underbrace{\min_x \left(f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \nu_j l_j(x) \right)}_{\text{point-wise minimum of convex functions in } (\lambda,\nu)}$$

subject to $\lambda_i \geq 0, i = 1, 2, ..., m$

  - $d^\star$ is the "best" estimate for the primal optimal value

Remark: always concave (even when primal problem is not convex)

- **Duality gap**

$$\boxed{G := p^\star - d^\star \geq 0}$$

  - always have $G \geq 0$ due to weak duality
  - if $d^\star = p^\star$, we say zero duality gap (or strong duality holds).

An Example for Duality Gap

- Consider a two-dimensional problem

$$\min_{x \in \mathbb{R}^2} \; e^{x_2}$$
$$\text{subject to } \; \|x\| - x_1 \leq 0$$

- Feasible solution $x_1 \geq 0, x_2 = 0 \Rightarrow p^\star = 1$.
- Consider now the dual function

$$g(\lambda) = \min_{x \in \mathbb{R}^2} e^{x_2} + \lambda \underbrace{\left(\sqrt{x_1^2 + x_2^2} - x_1 \right)}_{\geq 0}$$

which is positive for all $\lambda \geq 0$

An Example for Duality Gap

- Also, we can show that $g(\lambda) \leq 0 \; \forall \lambda \geq 0$. Let us restrict x to vary such that $x_1 = x_2^4$:

$$\sqrt{x_1^2 + x_2^2} - x_1 = \frac{x_2^2}{\sqrt{x_1^2 + x_2^2} + x_1} = \frac{x_2^2}{\sqrt{x_2^8 + x_2^2} + x_2^4} \leq \frac{1}{x_2^2}$$

- Thus, we have $x_2 \to -\infty \;\Rightarrow\; \sqrt{x_1^2 + x_2^2} - x_1 \to 0$
- It follows that

$$g(\lambda) \leq \min_{x_2 < 0, x_1 = x_2^4} e^{x_2} + \lambda \left(\sqrt{x_1^2 + x_2^2} - x_1 \right) = 0 \; \forall \lambda \geq 0$$

  which, together with $g(\lambda) \geq 0 \; \forall \lambda \geq 0$, shows that $g(\lambda) = 0$ for all $\lambda \geq 0$ and thus $d^\star = \max_{\lambda \geq 0} g(\lambda) = 0$.
- There is a duality gap $G = p^\star - d^\star = 1$!

Strong Duality

- **Slater condition**

$$\boxed{\begin{array}{l}\text{There exists a feasible } \bar{x} \in \mathbb{R}^n \text{ such that} \\ \qquad\qquad h_i(\bar{x}) < 0 \text{ (strictly feasible)}, \text{ for all } i = 1, 2, ..., m.\end{array}}$$

  Remark: linear inequalities do not need to be strict!

  Theorem: Let Assumption 1 and the Slater condition hold. Then,
  - There is no duality gap, i.e., $d^\star = p^\star$
  - The set of dual optimal solutions is nonempty and bounded
- If strong duality holds
  - KKT conditions (which are always sufficient) becomes necessary.
  - since $p^\star = d^\star$, instead of solving primal problem with complex constraints, we can

    $$\boxed{\text{Solve it from the dual}}$$

    which usually have simpler constraints, smaller dimension and thus algorithmically favorite!

The Dual of a Quadratic Program

- Consider the quadratic programming problem

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + c^T x, \text{ subject to } Ax \preceq b$$

where $Q \in \mathbb{R}^{n \times n}, A \in \mathbb{R}^{r \times n}$ with $r << n$

- **Lagrange dual function**

$$g(\lambda) = \min_x \frac{1}{2} x^T Q x + c^T x + \lambda^T (Ax - b)$$

which attains its minimum at $x = -Q^{-1}(c + A^T \lambda)$

- **Dual problem** becomes:

$$\max_\lambda g(\lambda) := -\frac{1}{2} \lambda^T P \lambda - a^T \lambda, \text{ subject to } \lambda \succeq 0$$

where $P = AQ^{-1}A^T, a = b + AQ^{-1}c$ with $P \in \mathbb{R}^{r \times r}$

$\boxed{\text{much smaller dimension and simpler constraints!}}$

Duality in Linear Programs

Given $c \in \mathbb{R}^n, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$:

primal problem

$$\min_{x \in \mathbb{R}^n} \; c^T x$$

subject to $Ax \preceq b$

dual problem

$$\max_{\lambda \in \mathbb{R}^m} \; -b^T \lambda$$

subject to $A^T\lambda + c = 0, \lambda \succeq 0$

- Lagrange dual function

$$g(\lambda) = \min_x \left\{(c + A^T\lambda)^T x - \lambda^T b)\right\} = \begin{cases} -b^T\lambda, & A^T\lambda + c = 0 \\ -\infty, & \text{otherwise} \end{cases}$$

  - Slater condition holds for linear constraints, thus $p^\star = d^\star$
  - the primal variable $x \in \mathbb{R}^n$ Versus the dual variable $\lambda \in \mathbb{R}^m$.

The Dual of SVM problem

- Recall the SVM problem

$$\min_{w} \frac{1}{2}\|w\|^2$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1,\ i = 1, 2, ..., m$$

- Lagrange dual function

$$g(\lambda) = \min_{w,b} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \lambda_i (1 - y_i(w^T x_i + b))$$

  – attaining optimality at $w = \sum_{i=0}^{m} \lambda_i y_i x_i, \sum_{i=0}^{m} \lambda_i y_i = 0$

- Dual problem becomes:

$$\max_{\{\lambda_i\}}\ g(\lambda) := \sum_{i=0}^{m} \lambda_i - \frac{1}{2}\sum_{i=0}^{m}\sum_{j=0}^{m} \lambda_i \lambda_j y_i y_j \underbrace{\langle x_i, x_j \rangle}_{\text{kernel}},$$

$$\text{s.t. } \sum_{i=0}^{m} \lambda_i y_i = 0, \lambda_i \geq 0, \forall i$$

Outline for Lecture I

Convex Optimization Problems

- Standard **convex optimization problem**

$$
\begin{aligned}
p^\star := \min_{x \in \mathbb{R}^n} \ & f(x) \\
\text{subject to } & h_i(x) \leq 0, \ i = 1, 2, ..., m, \\
& l_j(x) = 0, \ j = 1, 2, ..., r
\end{aligned}
$$

- The KKT (Karush-Kuhn-Tucker) conditions
  - (stationarity) $0 \in \nabla_x L(x, \lambda, \nu) := \nabla_x \left(f(x) + \sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \nu_j l_j(x) \right)$
  - (complementary slackness) $\lambda_i \cdot h_i(x) = 0$, for all i
  - (primal feasibility) $h_i(x) \leq 0, l_i(x) = 0$, for all i, j
  - (dual feasibility) $\lambda_i \geq 0$, for all i

- For unconstrained problems, the KKT conditions reduces to the ordinary optimality condition, i.e., $0 \in \partial f(x^\star)$.

Implication of KKT conditions

KKT conditions always sufficient; also necessary under strong duality.

Theorem: For a problem with strong duality,
$x^\star$ is a primal optimal and $\lambda^\star, \nu^\star$ a dual optimal solution
if and only if $x^\star$ and $\lambda^\star, \nu^\star$ satisfy the KKT conditions.

Why KKT conditions?

- provide a certificate of optimality for primal-dual pairs
- exploited in algorithm design and analysis
  - to verify optimality/suboptimality
  - as design principle (algorithms designed for solving KKT equations)
- **Limitations**: sometimes, KKT conditions do not really give us a way to find solution, but gives a better understanding and allow us to screen away some improper points before performing optimization.

Proof for Necessity

Let $x^\star$ and $\lambda^\star, \nu^\star$ be primal and dual optimal solutions with zero duality gap ($p^\star = g^\star$). Then

$$\begin{aligned} f(x^\star) &= g(\lambda^\star, \nu^\star) \\ &= \min_x f(x) + \sum_{i=1}^m \lambda_i^\star h_i(x) + \sum_{j=1}^r \nu_j^\star l_j(x) \\ &\le f(x^\star) + \sum_{i=1}^m \lambda_i^\star h_i(x^\star) + \sum_{j=1}^r \nu_j^\star l_j(x^\star) \\ &\le f(x^\star) \end{aligned}$$

- The point $x^\star$ achieves the Lagrangian Optimality in x, i.e., $x^\star = \inf_x L(x, \lambda^\star, \nu^\star)$, which is the staionary condition.
- By feasibility of $x^\star$, we must have $\sum_{i=1}^m \lambda_i^\star h_i(x^\star) = 0$, which in turn implies that $\lambda_i^\star h_i(x^\star) = 0$ (note that $h_i(x^\star) \le 0, \forall i$), which is the complementary slackness.

Proof for Sufficiency

If there exists $x^\star$, $\lambda^\star, \nu^\star$ that satisfies the KKT conditions, then

$$\begin{aligned} g(\lambda^\star, \nu^\star) &= \min_x L(x, \lambda^\star, \nu^\star) \\ &\overset{(a)}{=} f(x^\star) + \sum_{i=1}^m \lambda_i^\star h_i(x^\star) + \sum_{j=1}^r \nu_j^\star l_j(x^\star) = f(x^\star) \\ &\overset{(b)}{\geq} f(x^\star) + \sum_{i=1}^m \lambda_i h_i(x^\star) + \sum_{j=1}^r \nu_j l_j(x^\star) = L(x^\star, \lambda, \nu) \\ &\geq \min_x L(x, \lambda, \nu) = g(\lambda, \nu), \forall \lambda \geq 0 \end{aligned}$$

where (a) holds from the stationary condition and (b) holds from complementary slackness.

- The above together with the dual feasibility implies that the dual solution pair $(\lambda^\star, \nu^\star)$ is dual optimal.
- The above together with strong duality and primal feasibility also leads to the fact that $f^\star = g(\lambda^\star, \nu^\star) = f(x^\star)$, which implies that the primal solution $x^\star$ is primal optimal.

Example: quadratic with equality constraints

- Consider the following problem with $Q \succeq 0$

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^T Q x + c^T x$$
$$\text{s.t. } Ax = b$$

- Lagrangian function

$$L(x, \lambda) = \frac{1}{2} x^T Q x + c^T x + \lambda^T (Ax - b)$$

- KKT conditions
  - stationary condition

$$Qx + A^T \lambda = -c, Ax - b = 0, \text{or equivalently} \begin{bmatrix} Q & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ \lambda \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}$$

  - the above probelm reduces to solving linear system of equations (complementary slackness and dual feasiblity are vacuous)

Example: support vector machines

- Recall the SVM problem

$$\min_w \frac{1}{2}\|w\|^2$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1,\ \ i = 1, 2, ..., m$$

- Lagrangian function

$$L(w, \lambda) = \frac{1}{2}\|w\|^2 + \sum_i \lambda_i(1 - y_i(w^T x_i - b))$$

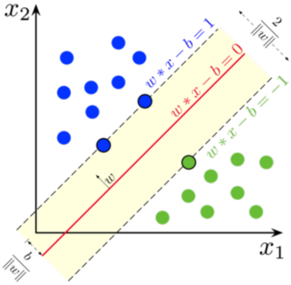
- KKT conditions
  - stationary condition

$$\sum_{i=1}^{m} \lambda_i y_i = 0, \quad w = \sum_i^{m} \lambda_i y_i x_i$$

  - complementary slackness

$$\lambda_i(1 - y_i(w^T x_i - b)) = 0, \forall i = 1, 2, ..., m$$

  - $\lambda_i \neq 0$ only when $1 = y_i(w^T x_i - b)$;



Such points are called the support vectors

Summary

- Convex sets and functions
  - convex set: $\theta x_1 + (1-\theta)x_2 \in \mathcal{C}$
  - convex function: $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$
  - operations that preserve conexity
- Weak duality and duality gap
  - $G = p^\star - d^\star \geq 0$
  - always true even primal problem is nonconvex
- Strong duality and its implication
  - $p^\star = d^\star$
  - solve the primal from the dual that is usually simpler
- KKT conditions and its implication

References

- Boyd, Stephen, and Lieven Vandenberghe. *Convex optimization*. Cambridge University press, 2004.
- Dimitri P., Bertsekas, Angelia, Nedich and Asuman E., Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.
- Angelia, Nedich. *Lecture Notes for Convex Optimization*. University of Illinois Urbana-Champaign, 2008.
- Ryan Tibshirani, *Lecture Notes for Convex Optimzation*. Carnegie Mellon University, 2018.