CS:E4830 Kernel Methods in Machine Learning

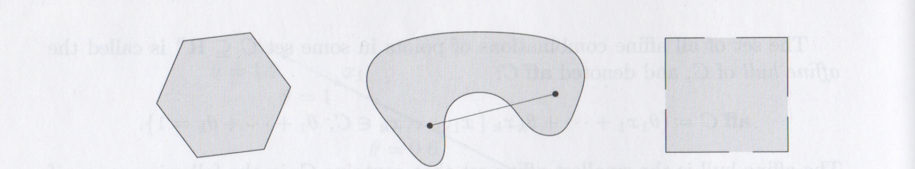
Lecture 5 : Convexity and Duality

Rohit Babbar

30th March, 2022

Convex sets

- A **line segment** between $x_1 \in \mathbb{R}^d$ and $x_2 \in \mathbb{R}^d$ is defined as all points that satisfy
$$x = \theta x_1 + (1-\theta)x_2, 0 \leq \theta \leq 1$$
- A **convex set** contains the line segment between any two distinct points in the set
$$x_1, x_2 \in C, 0 \leq \theta \leq 1 \Rightarrow \theta x_1 + (1-\theta)x_2 \in C$$
- Below: Convex and non-convex sets. Q: Which ones are convex?



Operations that preserve convexity of sets

There are two main ways of establishing the convexity of a set C

1. apply the definition of convexity:

$$x_1, x_2 \in C, 0 \leq \theta \leq 1 \Rightarrow \theta x_1 + (1-\theta)x_2 \in C$$

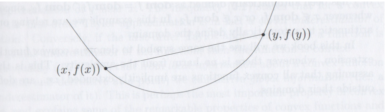
2. or show that the set can be obtained from simpler convex sets with operations that preserve convexity, most importantly:
    - intersection: if S_1, S_2 are convex, their intersection $S_1 \cap S_2$ is convex.
    - affine functions: If S is a convex and $f(x) = Ax - b$ is affine, the image of S under f, $f(S) = \{f(x)|x \in S\}$ is convex

Convex functions

- A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if (i) the domain $\mathcal{D}$ of f is a convex set and (ii) for all $x, y \in \mathcal{D}$, and $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y).$$

- Geometrical interpretation: the graph of the function lies below the line segment from $(x, f(x))$ to $(y, f(y))$



- A function f is
  - strictly convex if strict inequality holds above
  - concave if $-f$ is convex.

First order conditions

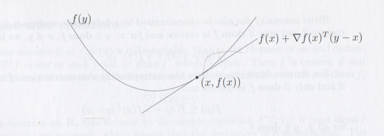
- Suppose $f : \mathbb{R}^d \mapsto \mathbb{R}$ is differentiable. Then f is convex if and only if

$$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

  holds for all $x, y \in \mathbb{R}^d$.

- The right-hand side, the first order Taylor approximation of f, is a global underestimator of f, i.e. at every point the function lies above the 1st order approximation.

- Geometrical interpretation: a convex function lies above each the tangent at any point.



- Corresponding forms of the equation can be written for strictly convex (replace $\geq$ with $>$ and concave functions (replace $\geq$ with $\leq$)

Second order conditions

- Assume $f : \mathbb{R}^d \mapsto \mathbb{R}$ is twice differentiable. Then f is convex if and only if its Hessian matrix (matrix of second derivatives)

$$H = [\nabla^2 f(x)]_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, i = 1, \dots, d, j = 1, \dots, d,$$

  is positive semi-definite: for all $y^T H y \geq 0$
- Geometrically the condition means that the function has positive curvature at x.
- Strict convexity is partially characterized by second order conditions: if $H = \nabla^2 f(x)$ is positive definite, $y^T H y > 0$ for all y, then f is strictly convex.
- For function defined on $\mathbb{R}$, the condition reduces to the simple condition $f''(x) \geq 0$, that is, that the first derivative is non-decreasing.
- Analogous conditions can be written for (strictly) concave functions and negative (semi-)definite Hessians

Operations that preserve convexity of functions

- Nonnegative weighted sums:
  - Nonnegative weighted sum of convex functions:

    $$f = w_1 f_1 + \dots w_m f_m,$$

    where $w_j \geq 0$ is convex.
  - Similarly, nonnegative weighted sum of concave functions is concave.
  - These properties extend to infinite sums and integrals
- Pointwise maximum and supremum:
  - Pointwise maximum $f(x) = \max(f_1(x), f_2(x), \dots, f_m(x))$, of a set $f_1, \dots, f_m$ of convex functions is convex.
  - Pointwise supremum of an infinite set of convex functions is convex
  - Similarly: Pointwise minimum (infimum) of concave functions is concave

Convex optimization problem

Standard form of a convex optimization problem

$$\min_{x \in \mathcal{D}} f_0(x)$$
$$\text{s.t. } f_i(x) \leq 0, i = 1, \ldots, m$$
$$h_i(x) = 0, i = 1, \ldots, p$$

- The problem is composed of the following components:
  - The variable $x \in \mathcal{D}$, $\mathcal{D} = \mathbb{R}^d$ is typical.
  - The **objective function** $f_0 : \mathbb{R}^d \mapsto \mathbb{R}$ to be minimized, a convex function of the variable x
  - The **constraint functions** $f_i : \mathcal{D} \mapsto \mathbb{R}$ related to **inequality constraints**, convex functions of x
  - The constraint functions $h_i(x) = a_i^T x - b_i$ related to **equality constraints**, affine (linear) functions of x

Convex optimization problem

Standard form of a convex optimization problem

$$\min_{x \in \mathcal{D}} \; f_0(x)$$
$$\text{s.t. } f_i(x) \leq 0, i = 1, \ldots, m$$
$$h_i(x) = 0, i = 1, \ldots, p$$

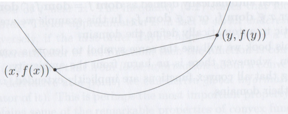
- Value of x that satisfy the constraints is called **feasible**, the set of all feasible points is called the **feasible set**.
- x is **optimal** if it has the smallest objective function value among all feasible $z \in \mathcal{D}$.
- Lets denote by p^*, the optimal value of the above problem, i.e.

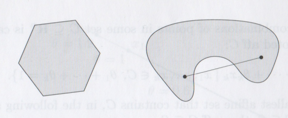
$$p^* = \min\{f_0(x) | f_i(x) \leq 0, i = 1, \ldots, m; h_i(x) = 0, i = 1, \ldots, p\}$$

Why convexity?

- Convex objective:
  - We can always improve a sub-optimal objective value by stepping towards negative gradient
  - All local optima are global optima
- Convex constraints i.e. convex feasible set
  - Any point between two feasible points is feasible
  - Updates remain inside the feasible set as long as the update direction is towards a feasible point

$\implies$ fast algorithms based on the principle of feasible descent





Duality

- Principle of viewing an optimization problem from two interchangeable views, primal and dual views
- Intuitively:
  - Minimization of a primal objective $\Leftrightarrow$ Maximization of the dual objective
  - Primal constraints $\Leftrightarrow$ Dual variables
  - Dual constraints $\Leftrightarrow$ Primal variables

Duality: Lagrangian

- Consider the primal optimisation problem

$$\begin{aligned} \min_{x \in \mathcal{D}} \; & f_0(x) \\ \text{s.t.} \; & f_i(x) \le 0, i = 1, \ldots, m \\ & h_i(x) = 0, i = 1, \ldots, p \end{aligned}$$

  with variable $x \in \mathbb{R}^d$

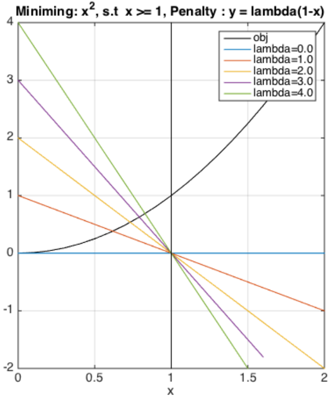
- Augment the objective function with the weighted sum of the constraint functions to form the **Lagrangian** of the optimization problem:

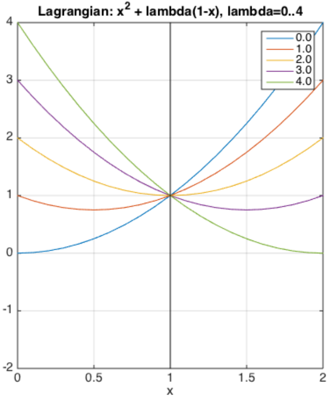
$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

- $\lambda_i, i = 1, \ldots, m$ and $\nu_i, i = 1, \ldots, p$ (ν is the greek letter 'nu') are called the **Lagrange multipliers** or **dual variables**

Example: Plotting the Lagrangian

- Minimizing $f_0(x) = x^2$, s.t. $f_1(x) = 1 - x \leq 0$
- Lagrangian: $L(x, \lambda) = x^2 + \lambda(1 - x)$





Lagrange dual function

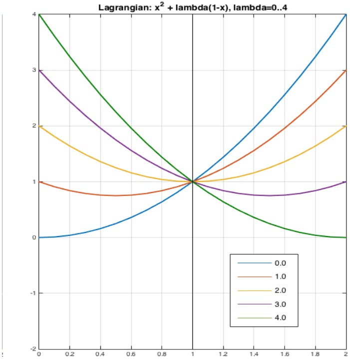
- The **Lagrange dual function** $g : \mathbb{R}^m \times \mathbb{R}^p \mapsto \mathbb{R}$ is the minimum value of the Lagrangian over x:

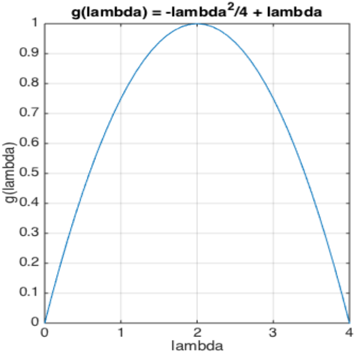
$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x \{f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \nu_i h_i(x)\}$$

- Intuitively:
  - Fixing coefficients (λ, ν) corresponds to certain level of penalty,
  - The infimum returns the optimal x for that level of penalty
  - $g(\lambda, \nu)$ is the corresponding value for the Lagrangian
  - $g(\lambda, \nu)$ is a concave function as a pointwise infimum of a family of affine functions of (λ, ν)

Example: Plotting the Lagrange dual function

- Minimizing x^2, s.t. $x \geq 1$
- Lagrange dual function : $g(\lambda) = \inf_x(x^2 + \lambda(1 - x))$
- Set derivatives to zero $\nabla_x(x^2 + \lambda(1 - x)) = 2x - \lambda = 0 \implies x = \lambda/2$
- Plug back to the Lagrangian: $g(\lambda) = \frac{\lambda^2}{4} + \lambda(1 - \lambda/2) = -\frac{\lambda^2}{4} + \lambda$





Lower bounds on optimal value

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x \{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\}$$

- In general, it holds that $g(\lambda, \nu) \leq p^*$ for any non-negative λ ($\lambda_i \geq 0, i = 1, \ldots, m$) and for any ν
- To see this, let $\tilde{x}$ be a feasible point of the original problem, thus all primal constraints are satisfied:
- We have $\lambda_i f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$ (Why?)

Lower bounds on optimal value

$$g(\lambda,\nu) = \inf_x L(x,\lambda,\nu) = \inf_x \{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\}$$

- In general, it holds that $g(\lambda,\nu) \leq p^*$ for any non-negative λ $(\lambda_i \geq 0, i = 1, \ldots, m)$ and for any ν
- To see this, let $\tilde{x}$ be a feasible point of the original problem, thus all primal constraints are satisfied:
- We have $\lambda_i f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$ (Why?) since $\lambda_i > 0$ by assumption and $f_i(\tilde{x}) \leq 0$

Lower bounds on optimal value

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x \{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\}$$

- In general, it holds that $g(\lambda, \nu) \leq p^*$ for any non-negative λ ($\lambda_i \geq 0, i = 1, \ldots, m$) and for any ν
- To see this, let $\tilde{x}$ be a feasible point of the original problem, thus all primal constraints are satisfied:
- We have $\lambda_i f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$ (Why?) since $\lambda_i > 0$ by assumption and $f_i(\tilde{x}) \leq 0$
- Similarly $\nu_i h_i(\tilde{x}) = 0$ for $i = 1, \ldots, p$ (Why?)

Lower bounds on optimal value

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x \{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\}$$

- In general, it holds that $g(\lambda, \nu) \leq p^*$ for any non-negative λ ($\lambda_i \geq 0, i = 1, \ldots, m$) and for any ν
- To see this, let $\tilde{x}$ be a feasible point of the original problem, thus all primal constraints are satisfied:
- We have $\lambda_i f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$ (Why?) since $\lambda_i > 0$ by assumption and $f_i(\tilde{x}) \leq 0$
- Similarly $\nu_i h_i(\tilde{x}) = 0$ for $i = 1, \ldots, p$ (Why?) $h_i(\tilde{x}) = 0$

Lower bounds on optimal value

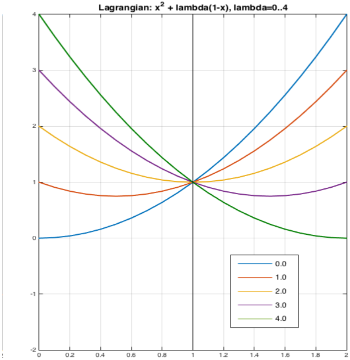
$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu) = \inf_x \{f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)\}$$

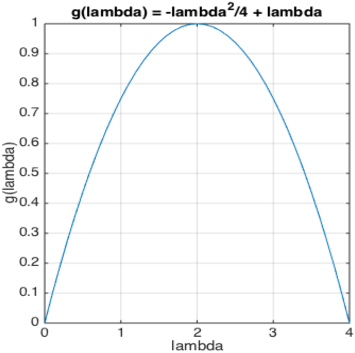
- In general, it holds that $g(\lambda, \nu) \leq p^*$ for any non-negative λ ($\lambda_i \geq 0, i = 1, \ldots, m$) and for any ν
- To see this, let $\tilde{x}$ be a feasible point of the original problem, thus all primal constraints are satisfied:
- We have $\lambda_i f_i(\tilde{x}) \leq 0$, $i = 1, \ldots, m$ (Why?) since $\lambda_i > 0$ by assumption and $f_i(\tilde{x}) \leq 0$
- Similarly $\nu_i h_i(\tilde{x}) = 0$ for $i = 1, \ldots, p$ (Why?) $h_i(\tilde{x}) = 0$
- Thus the value of the Lagrangian is less than the objective function at $\tilde{x}$:

$$L(\tilde{x}, \lambda, \nu) = f_0(\tilde{x}) + \sum_{i=1}^{m} \lambda_i f_i(\tilde{x}) + \sum_{i=1}^{p} \nu_i h_i(\tilde{x}) \leq f_0(\tilde{x})$$

- Now, $g(\lambda, \nu) = \inf_{x \in \mathcal{D}} L(x, \lambda, \nu) \leq L(\tilde{x}, \lambda, \nu) \leq f_0(\tilde{x})$ as infimum is computed over a set containing $\tilde{x}$

- The Lagrange dual function gives us **lower bounds** on the optimal value $p^*$of the primal problem: below, $g(\lambda) \leq 1 = p^*$,





The Lagrange dual problem

- For each pair $(\lambda, \nu), \lambda \geq 0$, the Lagrange dual function gives a lower bound on the optimal value of p^*.
- What is the tightest lower bound that can be achieved? We need to find the maximum
- This gives us an optimization problem

$$\max_{\lambda,\nu} g(\lambda, \nu)$$
$$\text{s.t.} \lambda, \nu \geq 0$$

- It is called the **Lagrange dual problem** of the original optimization problem.
- It is a convex optimisation problem, since it is equivalent to minimising $-g(\lambda, \nu)$ which is a convex function

Properties of convex optimisation problems

We will look at further concepts to understand the properties of convex optimisation problems

- Weak and strong duality
- Duality gap
- Complementary slackness
- KKT conditions

Weak and strong duality

- Let p^* and d^* denote primal and dual optimal values of an optimization problem.
- Weak duality
$$d^* \leq p^*$$
always holds, even when primal optimization problem is non-convex
- Strong duality
$$d^* = p^*$$
holds for convex optimization problems that we are interested in this course

Duality gap

- A pair $x, (\lambda, \nu)$ where x is primal feasible and (λ, ν) is dual feasible is called primal dual feasible pair
- For primal dual feasible pair, the quantity

$$f_0(x) - g(\lambda, \nu),$$

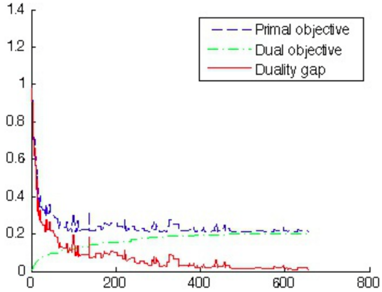
  is called the **duality gap**
- A primal dual feasible pair localizes the primal and dual optimal values

$$g(\lambda, \nu) \leq d^* \leq p^* \leq f_0(x)$$

  into an interval the width of which is given by the duality gap
- If the duality gap is zero, we know that x is primal optimal and (λ, ν) is dual optimal
- We can use duality gap as a stopping criterion for optimisation

Stopping criterion for optimization

- Suppose the algorithm generates a sequence of primal feasible $x^{(k)}$ and dual feasible $(\lambda^{(k)}, \nu^{(k)})$ solutions for $k = 1, 2, \ldots$
- Then the duality gap can be used as the stopping criterion: e.g. stop when $|f_0(x^{(k)}) - g(\lambda^{(k)}, \nu^{(k)})| \leq \epsilon$, for some $\epsilon > 0$



Complementary slackness

- Let x^* be a primal optimal (and thus also feasible) and (λ^*, ν^*) be a dual optimal (and thus also feasible) solution and let strong optimality hold, i.e. $d^* = p*$
- Then, at optimum

$$
\begin{aligned}
d^* = g(\lambda^*, \nu^*) = \inf_x \{ f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \nu_i^* h_i(x) \} \\
\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) \leq f_0(x^*) = p^*
\end{aligned}
$$

- First inequality: definition of infimum, second: inequality from x^* being a feasible solution
- Since $d^* = p^*$, the inequalities must hold as equalities $\implies$ penalty terms must equate to zero

Complementary slackness

- We have

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \nu_i^* h_i(x^*) = 0$$

- Since $h_i(x^*) = 0, i = 1, \ldots, p$ we conclude that

$$\sum_{i=1}^{m} \lambda_i^* f_i(x^*) = 0$$

  and since each term is non-positive

$$\lambda_i^* f_i(x^*) = 0, i = 1, \ldots, m$$

- This condition is called the **complementary slackness**

Complementary slackness

- Intuition: at optimum there cannot be both slack in the dual variable $\lambda_i > 0$ and the constraint $f_i(x^*) < 0$ at the same time:

$$\lambda_i^* > 0 \Rightarrow f_i(x^*) = 0$$

and

$$f_i(x^*) < 0 \Rightarrow \lambda_i^* = 0$$

- At optimum, positive Lagrange multipliers are associated with active constraints

Karush-Kuhn-Tucker (KKT) conditions

- For convex or non-convex optimization, at optimum the following conditions must hold true:
  - Inequality constraints satisfied: $f_i(x^*) \leq 0, i = 1, \ldots, m$
  - Equality constraints satisfied: $h_i(x^*) = 0, i = 1, \ldots, p$
  - Non-negativity of dual variables of the inequality constraints: $\lambda_i^* \geq 0, i = 1, \ldots, m$
  - Complementary slackness: $\lambda_i^* f_i(x^*) = 0, i = 1, \ldots, m$
  - Derivative of Lagrangian vanishes: $\nabla_x L(x^*, \lambda^*, \nu^*) = \nabla_x f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x f_i(x^*) + \sum_{i=1}^{p} \nu_i^* \nabla_x h_i(x^*)) = 0$
- These conditions are called the Karush-Kuhn-Tucker conditions
  - However, for convex problems, KKT conditions is also sufficient for optimality

- Convex Optimization by Boyd and Vandenberghe (available online with Videos)
  - Convex sets - chapter 2
  - Convex functions - chapter 3
  - Convex optimization problems - chapter 4
  - Duality - chapter 5