CS5275 Lecture 4: Convexity

Jonathan Scarlett

February 18, 2025

Useful references:

- Blog posts on Lagrange multipliers,[1] duality for linear programming,[2] and general Lagrange duality[3]
- Boyd and Vandenberghe's "Convex Optimization" book[4]
- Boyd's lectures on convex optimization, available on YouTube
- Nesterov's lecture notes on convex optimization

Categorization of material:

- Core material: Sections 1–4 except parts involving the Hessian, maxflow-mincut duality, and support vector machine.
- Extra material: Section 5 (KKT conditions) and the above-mentioned parts from Sections 1–4.

(Exam will strongly focus on "Core". Take-home assessments may occasionally require consulting "Extra".)

1 Convex Sets and Functions

Basic definitions.

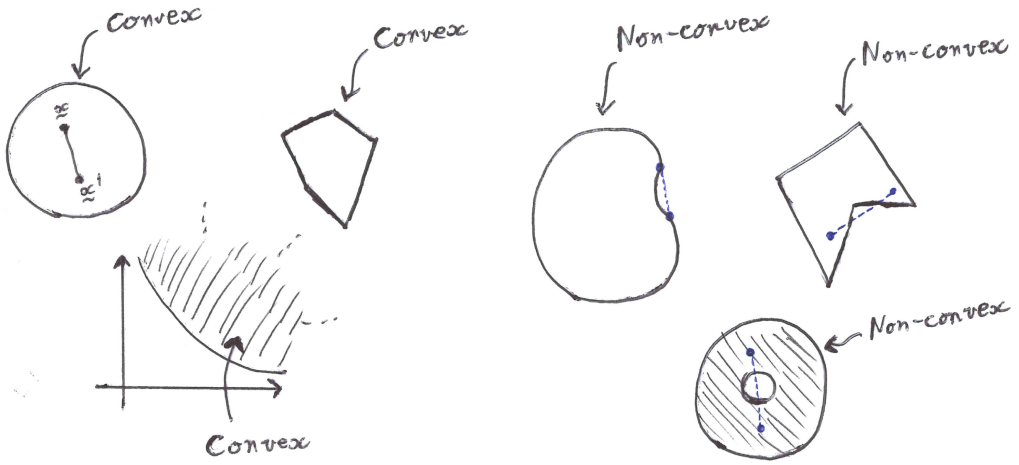
- A set D (e.g., a subset of $\mathbb{R}^d$) is said to be a *convex set* if, for all $\mathbf{x} \in D$ and $\mathbf{x}' \in D$, it holds that

$$\lambda \mathbf{x} + (1-\lambda)\mathbf{x}' \in D$$

  for all $\lambda \in [0,1]$

  - In words (roughly): Draw a straight line between any two points in D. This whole line segment must also lie within D.
  - Examples:

[1] http://jeremykun.com/2013/11/30/lagrangians-for-the-amnesiac/
[2] http://jeremykun.com/2014/06/02/linear-programming-and-the-most-affordable-healthy-diet-part-1/
[3] http://blogs.princeton.edu/imabandit/2013/02/21/orf523-lagrangian-duality/
[4] http://web.stanford.edu/~boyd/cvxbook/

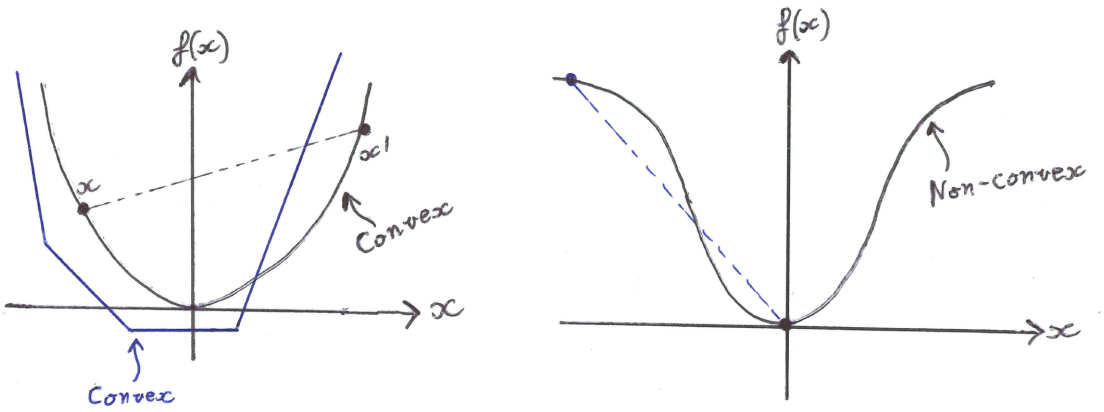


- A function $f : D \to \mathbb{R}$ is said to be a *convex function* if, for all $\mathbf{x} \in D$ and $\mathbf{x}' \in D$, it holds that

$$f(\lambda \mathbf{x} + (1-\lambda)\mathbf{x}') \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{x}')$$

  for all $\lambda \in [0, 1]$. Implicitly, this requires that the domain D is a convex set.

  - In words (roughly): Draw a straight line between $(\mathbf{x}, f(\mathbf{x}'))$ and $(\mathbf{x}', f(\mathbf{x}'))$. For inputs in between $\mathbf{x}$ and $\mathbf{x}'$, the function lies below this straight line.
  - Illustration:



  - We say that $f(\mathbf{x})$ is a *concave function* if $-f(\mathbf{x})$ is a convex function.
  - Convex = "bowl-shaped" ($\cup$), concave = "arch-shaped" ($\cap$)
  - A function is simultaneously convex and concave $\iff$ it is affine (i.e., a "straight line" (or plane)).
  - Key property. For a convex function, any local minimum is also a global minimum.

Other examples.

- Convex functions: $\|\mathbf{x}\|^2$, e^x, e^{-x}, $\log \sum_{i=1}^d e^{x_i}$, and many more.
- Concave functions: $-\|\mathbf{x}\|^2$, $\log x$, $\log \det \mathbf{X}$, $\sum_{i=1}^d x_i \log \frac{1}{x_i}$, and many more.

Equivalent definitions of convexity.

- Recall the notions of gradient and Hessian for $\mathbf{x} = [x_1, \ldots, x_d]^T$:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}, \qquad \nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1} & \frac{\partial^2 f}{\partial x_d x_2} & \cdots & \frac{\partial^2 f}{\partial x_d^2} \end{bmatrix}.$$

- (First order) If f is differentiable, then it is convex if and only if

$$f(\mathbf{x}') \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{x}' - \mathbf{x})$$

  for all $\mathbf{x}, \mathbf{x}'$. (The function lies above its tangent plane)

- (Second order) If f is twice differentiable, then it is convex if and only if

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{0}$$

  for all $\mathbf{x} \in D$. (The Hessian is positive semi-definite)

Operations that preserve convexity.

- If $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are convex, and α_1 and α_2 are positive, then $f(\mathbf{x}) = \alpha_1 f_1(\mathbf{x}) + \alpha_1 f_2(\mathbf{x})$ is convex. By induction, a similar statement holds for $\sum_{\ell=1}^{L} \alpha_\ell f_\ell(\mathbf{x})$ also for $L > 2$.

- If $f_1(\mathbf{x}), \ldots, f_L(\mathbf{x})$ are convex, then so is $f(\mathbf{x}) = \max_{\ell=1,\ldots,L} f_\ell(\mathbf{x})$.

- Certain compositions of the form $f(\mathbf{x}) = g(h(\mathbf{x}))$ are convex under certain conditions on g and h (see Section 3.2 of Boyd and Vandenberghe's book)

  - Simplest case: If h is a linear (or affine) function and g is convex, then f is convex.

Jensen's inequality.

- *Jensen's inequality* states that, for any random vector $\mathbf{X}$ and convex function f, it holds that

$$f(\mathbb{E}[\mathbf{X}]) \leq \mathbb{E}[f(\mathbf{X})].$$

  This is used in countless proofs in machine learning, statistics, information theory, etc.

- Note that the inequality is true directly from the definition of convexity when $\mathbf{X}$ equals one value $\mathbf{x}$ with probability λ, and another value $\mathbf{x}'$ with probability $1 - \lambda$. Jensen's inequality states the more general form for an arbitrary distribution on $\mathbf{X}$.

2 Convex Optimization

- In numerous fields, we are frequently interested in minimizing some cost function (or maximizing some utility function), possibly subject to certain constraints (see below for a variety of examples).

- Consider the following general form of optimization problem:

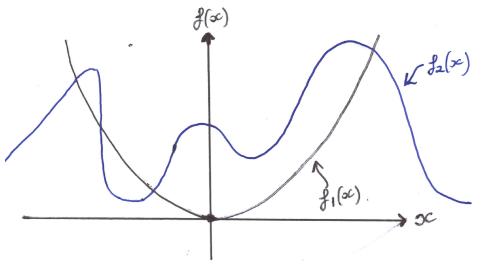
$$\begin{array}{lll} \text{minimize}_{\mathbf{x}} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, & i = 1, \ldots, m_{\text{ineq}} \\ & h_i(\mathbf{x}) = 0, & i = 1, \ldots, m_{\text{eq}}. \end{array} \tag{1}$$

  There are m_{ineq} *inequality constraints* and m_{eq} *equality constraints*.

- **Definition.** We say that (1) is a *convex optimization problem* if (i) $f_0(\mathbf{x})$ is convex; (ii) $f_i(\mathbf{x})$ is convex for all $i = 1, \ldots, m_{\text{ineq}}$; (iii) $h_i(\mathbf{x})$ is affine for all $i = 1, \ldots, m_{\text{eq}}$.

- This definition is very useful because, although solving (constrained or unconstrained) optimization problems is extremely hard in general, convexity is usually enough to permit finding a solution (sometimes analytically, but more often numerically).

- We can get some intuition by looking at the 1D case – which of these functions is easier to optimize using gradient descent techniques?

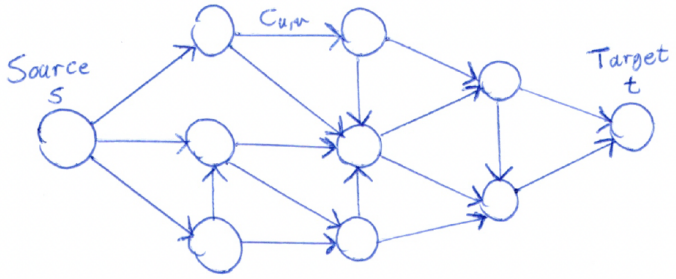


3 Examples of Convex Optimization Problems

Convexity may initially seem like a property that is rarely encountered in practice, but in fact it is surprisingly widespread and far-reaching. Just a few examples are given below.

Maximum flow via linear programming:

- Consider a directed graph with a number of nodes, two of which are the *source* and *target* (the latter is also known as the sink or destination):



- Let V denote the set of nodes and E denote the set of edges in the graph (i.e., $(u, v) \in E$ indicates an edge pointing from u to v), and let s and t denote the source and target nodes.

- Suppose that each edge $(u, v) \in E$ has a capacity c_{uv}, and consider the problem of creating as much "flow" (e.g., of information) from the source to target without exceeding any edge's capacity constraint.

- This naturally gives rise to a linear program: Letting f_{uv} represent the flow from u to v,

$$\begin{aligned} \text{maximize}_{\{f_{uv}\}_{(u,v)\in E}} \quad & \sum_{v\,:\,(s,v)\in E} f_{sv} \\ \text{subject to} \quad & 0 \le f_{uv} \le c_{uv} \quad \forall (u,v) \in E \\ & \sum_{u\,:\,(u,v)\in E} f_{uv} = \sum_{w\,:\,(v,w)\in E} f_{vw} \quad \forall v \in (V \setminus \{s,t\}), \end{aligned} \tag{2}$$

  where the first constraint simply constrains the total flow through each edge to be non-negative and not exceed capacity, and the second one constrains the total flow into any node to equal the total flow out (except for the source and target). The objective is to maximize the flow coming out of the source (which, by the constraints imposed, matches the amount coming into the target).

- Lagrange duality (covered below) for this problem is closely related to the famous *max-flow min-cut theorem* – we briefly explore this in an example later.

Estimation/regression via loss minimization:

- In the problem of linear regression, one is given n examples $(\mathbf{x}_i, y_i)$ (for $i = 1, \ldots, n$) with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$, and seeks to find linear weights $\boldsymbol{\theta}$ such that $\boldsymbol{\theta}^T \mathbf{x}_i$ approximates y_i well for all i (and similarly for unseen $(\mathbf{x}, y)$ pairs).

- A famous approach is *least squares*:

$$\text{minimize}_{\boldsymbol{\theta}} \sum_{i=1}^{n} (y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2,$$

  which is an unconstrained convex optimization problem (and a rare example of one that has an explicit closed-form solution). Other convex functions can also be used, e.g., using $|y_i - \boldsymbol{\theta}^T \mathbf{x}_i|$ instead of $(y_i - \boldsymbol{\theta}^T \mathbf{x}_i)^2$ provides better robustness to outliers.

- Constraints can also naturally enter, e.g., we may have prior information that the weights sum to one (so constrain $\sum_{i=1}^d \theta_i = 1$) or lie within a ball of a certain radius r (so constrain $\sum_{i=1}^d \theta_i^2 \le r$).

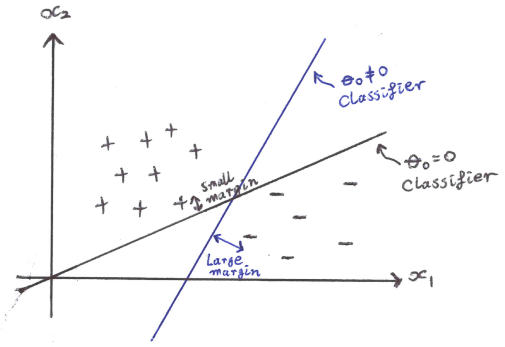
Maximum margin classification:

- The goal of binary classification seeks to construct a classifier that can reliably separate one class from another (e.g., distinguish spam vs. non-spam emails).

- Similarly to regression, we can represent inputs by vectors $\mathbf{x}_i \in \mathbb{R}^d$, and let $y_i \in \{-1, 1\}$ denotes its binary label (unlike regression where y was a general real value). Suppose that we are given a dataset with n such pairs, i.e., $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$.

- The class of *linear classifiers* takes the form $\hat{y} = \text{sign}(\boldsymbol{\theta}^T \mathbf{x} + \theta_0)$ for some weights $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\theta_0 \in \mathbb{R}$. (Here θ_0 is an extra *offset* that was omitted in the regression example.)

- When the data set being learned from is perfectly separable, a natural approach is to find the classifier with the *largest margin* (i.e., maximize the distance between the margin and the closest data point):



- Using some geometric analysis and manipulation, this turns out to be a convex optimization problem:

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{\theta}^T \mathbf{x}_i + \theta_0) \geq 1, \quad \forall i = 1, \ldots, n. \tag{3}$$

  (See Lectures 2 and 6a of `https://www.comp.nus.edu.sg/~scarlett/CS5339_notes/` for further details.) Lagrange duality for this problem will be discussed briefly below.

Power allocation in communication:

- A famous formula in information theory gives the maximum possible rate of communication over a communication channel that adds Gaussian noise to its input: $R = \frac{1}{2}\log\left(1+\frac{P}{\sigma^2}\right)$, where P is the input power and σ^2 is the noise variance.
  - Note: Due to the (very slow) sub-linear growth of the $\log(\cdot)$ function, this means that we have *diminishing returns* as we increase the power level P
- Now imagine that we have K communication channels with different noise levels $\sigma_1^2, \ldots, \sigma_K^2$, and the constraint is on the *total power*: $P_1 + \ldots + P_K \leq P_{\text{total}}$. How should we allocate power to maximize the total rate $R_1 + \ldots + R_K$?
- This is a concave maximization problem:

$$\begin{aligned} \text{maximize}_{P_1,\ldots,P_K} \quad & \sum_{i=1}^{K} \frac{1}{2}\log\left(1+\frac{P_i}{\sigma_i^2}\right) \\ \text{subject to} \quad & \sum_{i=1}^{K} P_i \leq P_{\text{total}}, \\ & P_i \geq 0, \qquad i = 1, \ldots, K. \end{aligned} \tag{4}$$

- Lagrange dual analysis (covered below) gives rise to a well-known solution termed *waterfilling*.

Portfolio optimization:

- Here we give one simple example from finance applications, known as Markowitz portfolio design.
- Suppose that we have a model for price changes in the form of a distribution on a random vector $\mathbf{p} \in \mathbb{R}^d$, where d is the number of assets (e.g., stocks) we can invest in, and p_i is the random price change for asset i. Specifically, let $\mu_{\mathbf{p}} \in \mathbb{R}^d$ denote the mean vector of $\mathbf{p}$, and $\Sigma_{\mathbf{p}} \in \mathbb{R}^{d\times d}$ denote

the covariance matrix (i.e., the diagonal entries give the variances, and the off-diagonals capture the correlations between price changes of different assets).

- Suppose that we would be satisfied with any average return of $r_{\min}$ or higher, but we are risk-averse so want to keep the variance low. Based on this idea, the Markowitz portfolio design is as follows:

$$\begin{aligned} \text{minimize}_{\mathbf{x}\in\mathbb{R}^d} \quad & \mathbf{x}^T\Sigma_{\mathbf{p}}\mathbf{x} \\ \text{subject to} \quad & \mu_{\mathbf{p}}^T\mathbf{x} \geq r_{\min} \\ & \sum_{i=1}^{n} x_i = 1, \end{aligned}$$

  where $\mathbf{x} = [x_1, \ldots, x_d]^T$ with x_i being the proportion invested into asset i. The last constraint arises from the total proportion being 1; constraining $\mu_{\mathbf{p}}^T\mathbf{x} \geq r_{\min}$ captures the goal of attaining the desired average return (or higher); and minimizing $\mathbf{x}^T\Sigma_{\mathbf{p}}\mathbf{x}$ captures the goal of lowering variance.

- Once again, this is a convex optimization problem.

4 Lagrange Multipliers and Duality

- Warm-up: We can get some rough intuition behind Lagrange multipliers by consider the two-function case: Minimize $f(\mathbf{x})$ subject to $g(\mathbf{x}) \leq 0$. Suppose both are differentiable. Let $\mathbf{x}^*$ be a point that we believe to be a minimizer.

  If we could find a direction $\mathbf{v}$ such that $\mathbf{v}^T\nabla g(\mathbf{x}^*) = \mathbf{0}$ but $\mathbf{v}^T\nabla f(\mathbf{x}^*) \neq \mathbf{0}$, then (*a bit informally / hand-wavey*) we could move a tiny amount in some direction to decrease f while still satisfying the constraint on g. If $\mathbf{x}^*$ is a minimizer then this should be impossible, and thus either (i) $\nabla f(\mathbf{x}^*) = \mathbf{0}$, or (ii) ∇f and ∇g are parallel,[5] i.e., $\nabla f(\mathbf{x}^*) + \lambda\nabla g(\mathbf{x}^*) = \mathbf{0}$. Case (i) may occur when the constraint is "inactive" (i.e., removing it makes no difference), whereas Case (ii) gives a different type of optimality condition stating that $\mathbf{x}^*$ minimizes $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$ for suitably-chosen λ (*Lagrange multiplier*).

- We proceed with a formal treatment of the general case. For an optimization problem of the form (1), the *Lagrangian* is defined as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i h_i(\mathbf{x}), \tag{5}$$

  where we have introduced extra parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_{m_{\text{ineq}}})$ and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_{m_{\text{eq}}})$. These are known as *Lagrange multipliers*.

  - We assume that $\lambda_i \geq 0$ for all i, whereas $\nu_i \in \mathbb{R}$ may be positive or negative.
  - Intuition: We no longer insist that $f_i(\mathbf{x}) \leq 0$, but we pay a penalty (scaled by λ_i) if it fails to hold. Conversely, we are "rewarded" if $f_i(\mathbf{x}) < 0$, i.e., strict inequality.

- **Important observation.** For any $\mathbf{x}$ feasible in (1), and any $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ with $\lambda_i \geq 0$, we have

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}). \tag{6}$$

[5] Two non-zero vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are parallel (i.e., one is a multiple of the other) if and only if, for all $\mathbf{u}$, the inner products $\mathbf{u}^T\mathbf{v}_1$ and $\mathbf{u}^T\mathbf{v}_2$ are either both zero or both non-zero.

- Proof: Follows immediately from $\lambda_i \geq 0$, $f_i(\mathbf{x}) \leq 0$, and $h_i(\mathbf{x}) = 0$.

- Minimizing both sides of (6) over $\mathbf{x}$ gives

$$\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \leq f_0(\mathbf{x}^*), \tag{7}$$

where $\mathbf{x}^*$ is an optimal solution to (1).

  - The function

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

is called the *Lagrange dual function.*

- Since $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ lower bounds $f_0(\mathbf{x}^*)$ according to (7), it is natural to look for the *best (highest) lower bound*. This leads to the *Lagrange dual problem*:

$$\begin{aligned} &\text{maximize}_{\boldsymbol{\lambda},\boldsymbol{\nu}} && g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \\ &\text{subject to} && \lambda_i \geq 0, \qquad i = 1, \ldots, m_{\text{ineq}}. \end{aligned} \tag{8}$$

Henceforth, let $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ denote the maximizer.

- **Duality.**

  - Since (7) holds for all $(\boldsymbol{\lambda}, \boldsymbol{\nu})$, it holds in particular for $(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$, yielding

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq f_0(\mathbf{x}^*).$$

This is known as *weak duality.*

  - One of the most important results in convex optimization is that, if the original optimization problem is convex (i.e., f_0 and f_i are convex functions, and h_i is are linear functions), and a mild regularity condition holds, then

$$g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = f_0(\mathbf{x}^*). \tag{9}$$

This is known as *strong duality.*

    * There are many possible "mild regularity conditions"; the most well-known is known as *Slater's condition*: There exists at least one feasible point $\mathbf{x}$ satisfying the constraints of (1) with strict inequality (i.e., $f_i(\mathbf{x}) < 0$ and $h_i(\mathbf{x}) = 0$).
    * Another (more restrictive) sufficient condition is that the constraint functions f_i ($i = 1, \ldots, m_{\text{ineq}}$) are not only convex, but linear (and a feasible point exists).

  - Minimax theorem viewpoint: One way to understand duality is to interpret the original constrained optimization problem as solving

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

This is because the inner maximization (more precisely a supremum) equals ∞ whenever $f_i(\mathbf{x}) > 0$ or $h_i(\mathbf{x}) \neq 0$, because any arbitrarily large value can be achieved by taking the corresponding λ_i

or ν_i to be huge. In addition, when $\mathbf{x}$ satisfies the constraints (i.e., each $f_i(\mathbf{x}) \leq 0$ and $h_i(\mathbf{x}) = 0$), it is not hard to show that that $\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x})$ (achieved by $\boldsymbol{\lambda} = \mathbf{0}$ and $\boldsymbol{\nu} = \mathbf{0}$).

In contrast, the Lagrange dual problem solves

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}).$$

So almost everything is the same – just the max and min are swapped!

It a well-known fact of optimization and game theory that $\min_A \max_B f(A, B) \geq \max_B \min_A f(A, B)$. Strong duality is related to the *minimax theorem* (see `https://en.wikipedia.org/wiki/Minimax_theorem`), which states that in fact

$$\min_A \max_B f(A, B) = \max_B \min_A f(A, B)$$

in the case that $f(A, \cdot)$ is concave in B, $f(\cdot, B)$ is convex in A, and some other "mild" conditions hold (e.g., $\min_A$ and $\max_B$ are optimizations over compact sets).

- Optimality certificate viewpoint: Notice that any primal feasible $\mathbf{x}$ provides an *upper bound* to the optimal solution. Correspondingly, any dual feasible $(\boldsymbol{\lambda}, \boldsymbol{\mu})$ provides a *lower bound* to the optimal solution. When strong duality holds, we can get *matching upper and lower bounds*, which provide us a proof of optimality, so we call this a *certificate*.
- Other viewpoints (**Optional**): See Section 5 Boyd/Vandenberghe for other interpretations, including geometric (Section 5.3; see also `https://www.argmin.net/p/ends-in-a-draw` for a summary) and sensitivity analysis based (Section 5.6). In the latter, some statements are given on how the value of the Lagrange multiplier relates to how much the optimal value changes upon tightening/loosening the constraints.

- **Example 1 (Linear programming).**
  - Consider a linear program of the form

$$\begin{aligned} \text{maximize}_{\mathbf{x}} \quad & \mathbf{c}^T \mathbf{x} && (10) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0} && (11) \end{aligned}$$

    for some matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ and vectors $\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{c} \in \mathbb{R}^d$. The inequality $\mathbf{x} \geq \mathbf{0}$ should be interpreted as holding element-wise.
  - Interpreting this as being in the form (1) with $m_{\text{ineq}} = m$ and $m_{\text{eq}} = d$, we have the Lagrangian

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) &= -\mathbf{c}^T \mathbf{x} - \sum_{i=1}^{d} \lambda_i x_i + \sum_{i=1}^{m} \nu_i (\mathbf{a}_i^T \mathbf{x} - b_i) \\ &= -\mathbf{c}^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{x} + \boldsymbol{\nu}^T (\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= -\mathbf{b}^T \boldsymbol{\nu} + (\mathbf{A}^T \boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c})^T \mathbf{x}, \end{aligned}$$

    where $\mathbf{a}_i$ is the i-th row of $\mathbf{A}$, b_i is the i-th entry of $\mathbf{b}$, etc.
    - Note: Switching from "maximize" to "minimize" requires taking $f_0(\mathbf{x}) = -\mathbf{c}^T \mathbf{x}$. Similarly, we rewrite $\mathbf{x} \geq \mathbf{0}$ as $-\mathbf{x} \leq \mathbf{0}$.

– Minimizing over $\mathbf{x}$, we find that $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ (which we recall is $\min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$) takes the form

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \begin{cases} -\mathbf{b}^T \boldsymbol{\nu} & \mathbf{A}^T \boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c} = \mathbf{0} \\ -\infty & \text{otherwise.} \end{cases}$$

This is because whenever $\mathbf{A}^T \boldsymbol{\nu} + \boldsymbol{\lambda} + \mathbf{c} \neq 0$, one can just make a suitable entry of x_i arbitrarily large in either the positive or negative direction.

– Substituting this expression for $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ into (8) yields the *dual problem*:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & -\mathbf{b}^T \boldsymbol{\nu} \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}, \\ & \mathbf{A}^T \boldsymbol{\nu} - \boldsymbol{\lambda} - \mathbf{c} = \mathbf{0}, \end{aligned}$$

where the second constraint can be introduced since all other values yield a (certainly suboptimal) value of $-\infty$. Since $\boldsymbol{\lambda}$ does not appear in the objective function, we can further simplify the above maximization to

$$\begin{aligned} \text{minimize}_{\boldsymbol{\nu}} \quad & \mathbf{b}^T \boldsymbol{\nu} \\ \text{subject to} \quad & \mathbf{A}^T \boldsymbol{\nu} \geq \mathbf{c}. \end{aligned}$$

– If we replace $\mathbf{Ax} = \mathbf{b}$ by $\mathbf{Ax} \leq \mathbf{b}$ in the original formulation, then we arrive at a similar dual expression but with the added constraint $\boldsymbol{\nu} \geq 0$.

– **An intuitive interpretation:**

* The original problem constrains $\mathbf{Ax} = \mathbf{b}$; multiplying both sides on the left by $\boldsymbol{\nu}^T$ gives $\boldsymbol{\nu}^T \mathbf{Ax} = \boldsymbol{\nu}^T \mathbf{b}$, or equivalently $(\mathbf{A}^T \boldsymbol{\nu})^T \mathbf{x} = \mathbf{b}^T \boldsymbol{\nu}$ (by standard properties of the transpose)
* Now, since $\mathbf{x} \geq \mathbf{0}$ and we are maximizing $\mathbf{c}^T \mathbf{x}$, we find that if $\mathbf{A}^T \boldsymbol{\nu} \geq \mathbf{c}$, it holds that $(\mathbf{A}^T \boldsymbol{\nu})^T \mathbf{x}$ is at least as high as $\mathbf{c}^T \mathbf{x}$. Then, by the previous dot point, $\mathbf{b}^T \boldsymbol{\nu}$ is at least as high as $\mathbf{c}^T \mathbf{x}$.
* Hence, for any $\boldsymbol{\nu}$ that satisfies $\mathbf{A}^T \boldsymbol{\nu} \geq \mathbf{c}$, we have that $\mathbf{b}^T \boldsymbol{\nu}$ is at least as high as the original problem's optimal value, i.e., it is an *upper bound* to the optimal value.
* By minimizing over all such $\boldsymbol{\nu}$ (as is done in the dual expression), we are finding the *lowest (best) possible upper bound*, and this turns out to make the upper bound hold with equality.

• **Example 2 (Maxflow-mincut duality).**

– A famous theorem for the maxflow problem is *maxflow-mincut duality*: The value of the maximum flow is the same as the value of the minimum cut (i.e., the minimum total weight of edges whose removal makes the target unreachable from the source).

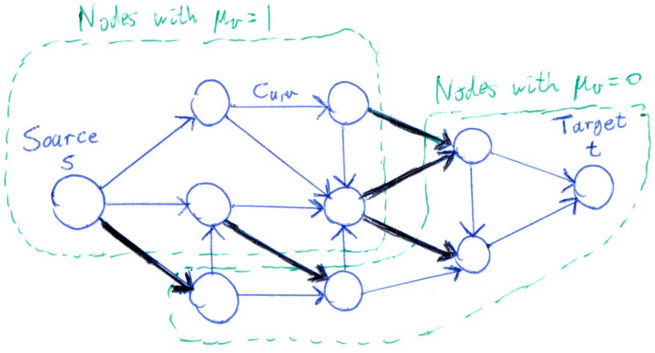
– A Lagrange dual analysis of (2) similar to Example 1 gives rise to the following:

$$\begin{aligned} \text{maximize}_{\{\lambda_{uv}\}_{(u,v)\in E}, \{\mu_v\}_{v \in V}} \quad & \sum_{(u,v)\in E} c_{uv} \lambda_{uv} \\ \text{subject to} \quad & \mu_s = 1, \mu_t = 0, \\ & \lambda_{uv} \geq \mu_u - \mu_v, \ \lambda_{uv} \geq 0 \ \ (\forall (u,v) \in E). \end{aligned}$$

* Each λ_{uv} arises as a Lagrange multiplier for the edge's capacity constraint, and each μ_v arises as a Lagrange multiplier for the node's inflow=outflow constraint (except s and t, which is why their μ_v values are instead fixed to 1 and 0).

- Note that once the μ_v's are all specified, the λ_{uv}'s are trivially given by $\lambda_{uv} = \max\{0, \mu_u - \mu_v\}$. So the problem "minimize over all μ_v and λ_{uv}" is essentially just one of "minimize over all μ_v".
- If we additionally constrain $\mu_v \in \{0,1\}$ and $\lambda_{uv} \in \{0,1\}$ for all u and v, then the mincut interpretation is immediate: The nodes with $\mu_u = 1$ are on the "source side", those with $\mu_v = 0$ are on the "target side", and the goal is to minimize the sum of c_{uv} over all edges from the former to the latter (for which $\lambda_{uv} = 1$).

  An illustration (with bold edges being the 'cut' ones, whose c_{uv} values are summed):



- In other words, the Lagrange dual analysis has given a *linear programming relaxation* of the min-cut problem. But it turns out to be a case where the relaxation is *tight* – it gives the same answer as the integer-constrained problem. (See the next lecture for more on this.)

- **(**Optional**) Example 3 (Support vector machine).**
  - As given in the above examples, the maximum-margin hyperplane problem in classification can be cast as

$$\text{minimize}_{\boldsymbol{\theta},\theta_0} \quad \frac{1}{2}\|\boldsymbol{\theta}\|^2 \quad \text{subject to} \quad y_i(\boldsymbol{\theta}^T\mathbf{x}_i + \theta_0) \geq 1, \quad \forall i = 1,\ldots,n. \tag{12}$$

    This is a convex optimization problem with affine constraints, so strong duality holds.
  - By a similar analysis Example 1, the following dual optimization problem can be derived:

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^n \alpha_t - \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n \alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j \\ \text{subject to} \quad & \alpha_i \geq 0 \quad \forall i \in \{1,\ldots,n\}, \\ & \sum_{i=1}^n \alpha_i y_i = 0. \end{aligned}$$

    * Each $\alpha_i \geq 0$ arises as a Lagrange multiplier corresponding to the i-th data point's constraint, and the condition $\sum_{i=1}^n \alpha_i y_i = 0$ is related to an optimization over θ_0 (if θ_0 is removed from the problem altogether, then so is this constraint in the dual).
  - In addition, the Lagrange duality analysis reveals that $\boldsymbol{\theta} = \sum_{i=1}^n \alpha_i y_i \mathbf{x}_i$, and complementary slackness (covered below) can used to derive $\theta_0 = \frac{1}{y_t} - \boldsymbol{\theta}^T\mathbf{x}_t$ for any t such that $\alpha_t > 0$.

– A significant advantage of the dual formulation is that it depends on the $\mathbf{x}$'s only via inner products (e.g., $\mathbf{x}_i^T \mathbf{x}_j$), which allows for an application of the *kernel trick* where each inner product is replaced by a more general function $k(\mathbf{x}_i, \mathbf{x}_j)$. The idea is that we could "change the representation" of the input by replacing each $\mathbf{x}$ by some function $\phi(\mathbf{x})$, but the kernel trick allows us to do so *implicitly* instead of explicitly – there are many spaces where $\phi(\mathbf{x})$ is "complicated" (e.g., infinite-dimensional) and yet $\phi(\mathbf{x})^T \phi(\mathbf{x}')$ is "simple" (e.g., can be evaluated in linear time).

– See Lecture 6a of `https://www.comp.nus.edu.sg/~scarlett/CS5339_notes/` for the complete details of this example.

5 The Karush-Kuhn-Tucker (KKT) Conditions

- In the case that strong duality holds as per (9), we have the following chain of inequalities:

$$\begin{aligned} f_0(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \\ &= \min_{\mathbf{x}} \left\{ f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}) \right\} \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*), \end{aligned}$$

  where we first applied the definition of g, then upper bounded the minimum by the specific value $\mathbf{x}^*$, then used the fact that $f_i(\mathbf{x}^*) \leq 0$ and $h_i(\mathbf{x}^*) = 0$.

- Since we ended up with $f_0(\mathbf{x}^*) \leq f_0(\mathbf{x}^*)$, both of the inequalities must hold with equality. Let's look at these in more detail:

  – The first inequality holding with equality gives

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* h_i(\mathbf{x}).$$

    Assuming the functions are differentiable, the fact that $\mathbf{x}^*$ is a minimizer means that the derivative must vanish:

$$\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* \nabla h_i(\mathbf{x}^*) = \mathbf{0}.$$

  – The second inequality holding with equality gives

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \qquad i = 1, \ldots, m_{\text{ineq}}.$$

    This means that either $f_i(\mathbf{x}^*) = 0$ (i.e., the constraint holds with equality) or $\lambda_i^* = 0$. This property is known as *complementary slackness*.

- Summarizing the above leads to a set of conditions on $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ known as the *KKT conditions*:

  1. (Primal feasibility) $f_i(\mathbf{x}^*) \leq 0$ for $i = 1, \ldots, m_{\text{ineq}}$, and $h_i(\mathbf{x}^*) = 0$ for $i = 1, \ldots, m_{\text{eq}}$.

2. (Dual feasibility) $\lambda_i^* \geq 0$ for $i = 1, \ldots, m_{\text{ineq}}$.
3. (Complementary slackness) $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for $i = 1, \ldots, m_{\text{ineq}}$.
4. (Vanishing gradient) $\nabla f_0(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{ineq}}} \lambda_i^* \nabla f_i(\mathbf{x}^*) + \sum_{i=1}^{m_{\text{eq}}} \nu_i^* \nabla h_i(\mathbf{x}^*) = \mathbf{0}$.

These generalize the requirement that the *unconstrained* maximizer of $f_0(\mathbf{x})$ should satisfy $\nabla f_0(\mathbf{x}^*) = 0$.

- General case: If strong duality holds, it is necessary that $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfy the KKT conditions.
- Convex case: If strong duality holds *and the primal problem is convex*, then $(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$ satisfying the KKT conditions are *also sufficient for optimality* (the proof of this is omitted).

- **(**Optional**)** As a final note (without proof), *geometric* optimality conditions can also be formed: For constraint sets with smooth boundaries,the negative gradient vector $-\nabla f(\mathbf{x}^*)$ points perpendicular to the constraint set boundary (left figure below). A similar statement can be made for "pointy" constraint sets like polyhedra, except that we get an *entire cone* of possible directions (right figure below).