ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Alex Mikhael, Rushabha Balaji, Vinay Kanakeri April 30, 2024
Editor/TA: Marialena Sfyraki

Lecture 9: Duality Theory, Part I

1 Review

We begin by reviewing some results from last lecture.

Lemma 1. *Let $F(x) = \mathbb{E}_z[f(x;z)] : \mathbb{R}^d \to \mathbb{R}$ be a convex function. Consider the update*

$$x_{k+1} = x_k - \eta g_k,$$

where $\mathbb{E}_z[g_k] = \nabla F(x_k)$. Suppose $x_* = \arg\min F(x)$ exists and the initial distance is bounded, i.e., $\|x_1 - x_*\| \leq D$. Then,

$$\frac{1}{K}\mathbb{E}\left[\sum_{k=1}^{K} F(x_k) - F(x_*)\right] \leq \frac{\eta}{2K}\left(\sum_{k=1}^{K}\mathbb{E}\left[\|g_k\|_2^2\right]\right) + \frac{\|x_1 - x_*\|^2}{2\eta K}$$

Remark: The variance is a critical factor in the progress of SGD.

Theorem 1. *Suppose each $f_i(\cdot)$ is a L-smooth and μ strongly convex. Setting $\eta = \frac{1}{8L}$ and $K = 64\frac{L}{\mu}$. Then, at each stage s, given v_s,*

$$\mathbb{E}\left[F(v_{s+1}) - F(x_*)\right] \leq \frac{3}{4}\left(F(v_s) - F(x_*)\right)$$

where $x_* \in \arg\min_x F(x)$

Remark: The optimality gap at each stage decays with a constant factor as shown in Theorem 1.

The condition for SVRG to be faster than GD is when

$$\left(\frac{L}{\mu} + n\right) \ll n\frac{L}{\mu} \;\Leftrightarrow\; n \ll (n-1)\frac{L}{\mu}$$

This holds if the function is smooth, strongly convex, and the problem is finite-sum. It can also be shown that SVRG is faster than SGD for smooth, convex, finite-sum problems.

Remark: SVRG is not applicable if we get the data in a stream fashion. In fact, in this case, we would not know all the samples and thus would not have access to the full gradient. In this case, SGD is better.

2 Lagrangian, Dual Problem, and Duality

2.1 Motivation

In practice, the optimal solution is unknown and we need a way to verify the solution obtained after solving the optimization problem is indeed optimal. For that, we derive an upper bound for the optimality gap. Let $f(\cdot) : C \to \mathbb{R}$ and $C \subseteq \mathbb{R}^d$ where the optimality gap is denoted by $\delta_k := f(x_k) - \inf_{x \in C} f(x)$.

An upper bound for the optimality gap provides a lower bound for the optimal value. The estimated lower bound, denoted by y_*, is characterized by $y_* \leq \inf_{x \in C} f(x)$.

2.2 Optimization with functional constraints

The constrained optimization problem consisting of m inequality constraints and p equality constraints is shown below.

$$\begin{aligned} &\inf_{x \in \text{dom} f} f(x) \\ &\text{s.t. } f_j(x) \leq 0, \quad j = 1, \ldots, m \\ &\quad \textbf{affine } \ h_i(x) = 0, \quad i = 1, \ldots, p \end{aligned}$$

Note that the above problem is convex if $f(\cdot), f_i(\cdot)$, $\forall i \in [n]$ are convex and $h_i(\cdot)$, $\forall i$ are affine. Note that if $h_i(\cdot)$, $\forall i$ are not affine, we might end up with a non-convex feasible region. As a counterexample, consider a set that contains only a single equality constraint, i.e. $S_0 = \{x : x^2 = 1\}$. The roots of $x^2 = 1$ are $1, -1 \in S_0$ and clearly this set is non-convex, since $\frac{1}{2} \cdot (-1) + \frac{1}{2} c \cdot 1 = 0 \notin S_0$.

Recall that:

- **Linear Function** : $h_i(x) = a_i^\top x$
- **Affine function** : $h_i(x) = a_i^\top x + d$, where $d \in \mathbb{R}$

Note that both linear and affine functions are convex. Additionally, a set S such that $S = \{x : a^\top x = c\}$ defines a **hyperplane**.

We can formulate the domain of the function to be a constraint using the indicator function denoted by $I(\cdot)$. The indicator function is defined as follows

$$I(x \in \text{dom} f) = \begin{cases} 0 & \text{if } x \in \text{dom} f \\ \infty & \text{if } x \notin \text{dom} f. \end{cases}$$

The optimization problem can be equivalently formulated as

$$\begin{aligned} &\inf_{x \in \mathbb{R}^d} f(x) \\ &\text{s.t.} f_j(x) \leq 0, \quad j = 1, \ldots, m \\ &\quad \textbf{affine} \;\; h_i(x) = 0, \quad i = 1, \ldots, p \\ &\quad I(x \in \text{dom} f) \leq 0 \end{aligned}$$

Observe that in this case the infimum is taken over the whole $\mathbb{R}^n$. If the original problem is convex, then adding the indicator function constraint does not change the convex nature of the problem because the indicator function is convex. Also, note that the minimum value is unchanged between the above two problems because the feasible solutions to the above program will always be in $\text{dom} f$.

We will next present the method of Lagrange multipliers for solving constrained optimization problems of the above form.

Definition 1. ***(Lagrangian)***

$$L(x, \lambda, \mu) := f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{i=1}^{p} \mu_i h_i(x), \tag{1}$$

where $\lambda_j \geq 0$, $\forall\ j \in [m]$ and $\mu_i \in \mathbb{R}$, $\forall\ i \in [p]$. λ_j and μ_i are called the Lagrangian multipliers.

Remark: x is called the primal variable and μ and λ are called the dual variables.

Property 1 of the Lagrangian:
Let $\Omega := \{x \in \mathbb{R}^d : f_j(x) \leq 0, \forall j \in [m]; h_i(x) = 0, \forall i \in [p]\}$. Consider $x \in \Omega$. Then the Lagrangian lower-bounds the function $f(x)$, that is

$$L(x, \lambda, \mu) \leq f(x), \tag{2}$$

where $\lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_m \end{bmatrix}$ and $\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$

Proof: By the definition of Lagrangian, we have

$$L(x, \lambda, \mu) := f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{i=1}^{p} \mu_i h_i(x).$$

If $x \in \Omega$, then $\sum_{j=1}^{m} \lambda_j f_j(x) \leq 0$, because $\lambda_j \geq 0, \forall j \in [m]$, and $f_j(x) \leq 0$ which results in individual $\lambda_j f_j(x) \leq 0$. If $x \in \Omega$, we also have $\sum_{i=1}^{p} \mu_i h_i(x) = 0$ since $h_i(x) = 0$.

Question: For any x that satisfies the functional constraints, i.e., $x \in \Omega$, when does $L(x, \lambda, \mu) = f(x)$?
Answer: When $\lambda = 0$, because in that case we have $L(x, 0, \mu) = f(x),\ \forall x \in \Omega$.

Property 2 of the Lagrangian:
Let $\Omega := \{x \in \mathbb{R}^d : f_j(x) \leq 0, \forall j \in [m]; h_i(x) = 0, \forall i \in [p]\}$. If $x \in \Omega$, then

$$\sup_{\lambda \geq 0; \mu} L(x, \lambda, \mu) = \begin{cases} f(x) & \text{if } x \in \Omega \\ \infty & \text{otherwise} \end{cases} \tag{3}$$

Proof: Property 2 follows from the definition of Lagrangian (1). If $x \notin \Omega, \exists\ j \in [m]$ or $i \in [p]$ such that $f_j(x) > 0$ or $h_i(x) \neq 0$, then we can choose appropriate λ_j or μ_i such that $L(x, \lambda, \mu) = \infty$. When $x \in \Omega$, then $L(x, \lambda, \mu)$ is upper bounded by $f(x)$ from Property 1.

Remark : Based on Property 1 and Property 2, we have the following result:

$$\inf_{x \in \mathbb{R}^d} \sup_{\lambda \geq 0; \mu} L(x, \lambda, \mu) = \inf_{x \in \Omega} \sup_{\lambda \geq 0; \mu} L(x, \lambda, \mu) = \inf_{x \in \Omega} f(x) \tag{4}$$

3 Dual function

The dual function is defined as

$$g(\lambda, \mu) := \inf_{x \in \mathbb{R}^d} L(x, \lambda, \mu) \tag{5}$$

Note that for some λ, μ the Lagrangian may be unbounded from below in x. In that case, the dual function $g(\lambda, \mu)$ takes on the value $-\infty$.

Remark: Observe that the dual function $g(\lambda, \mu)$ only depends on dual variables.

3.1 Weak and Strong Duality

Theorem 2. ***(Weak duality):*** *The primal value* $\inf_{x\in\Omega} f(x)$ *and the dual value* $\sup_{\lambda\geq 0;\mu} g(\lambda,\mu)$ *are related as*

$$\sup_{\lambda\geq 0;\mu} g(\lambda,\mu) \leq \inf_{x\in\Omega} f(x),$$

that is, the dual value is not greater than the primal value.

Proof. From the equality in (4), we have

$$\inf_{x\in\mathbb{R}^d} \sup_{\lambda\geq 0;\mu} L(x,\lambda,\mu) = \inf_{x\in\Omega} f(x) \tag{6}$$

$$\Rightarrow \inf_{x\in\mathbb{R}^d} L(x,\lambda,\mu) \leq \inf_{x\in\Omega} f(x). \tag{7}$$

Notice that (7) is true for any $\lambda \geq 0, \mu$, therefore we have

$$\inf_{x\in\mathbb{R}^d} L(x,\lambda,\mu) \leq \inf_{x\in\Omega} f(x) \tag{8}$$

$$\Leftrightarrow g(\lambda,\mu) \leq \inf_{x\in\Omega} f(x) \tag{9}$$

$$\Leftrightarrow \sup_{\lambda\geq 0;\mu} g(\lambda,\mu) \leq \inf_{x\in\Omega} f(x) \tag{10}$$

This completes the proof for weak duality. □

Definition 2. ***(Strong duality):*** *When the primal value and dual value are equal, we say strong duality is satisfied, that is*

$$\sup_{\lambda\geq 0;\mu} g(\lambda,\mu) = \inf_{x\in\Omega} f(x). \tag{11}$$

In the next lecture, we will learn the sufficient condition for strong duality, and how these conditions are necessary and sufficient when the optimization problem is convex.

4 Example

Consider the following primal problem, which is an example of a generic linear programming problem:

$$\min_{x\in\mathbb{R}^d} \langle c, x\rangle$$
$$\text{s.t } Ax \geq b,$$

where $A \in \mathbb{R}^{m \times d}$, and $b \in \mathbb{R}^m$. The above problem can be rewritten as

$$\begin{aligned} &\min_{x \in \mathbb{R}^d} \langle c, x \rangle \\ &\text{s.t } b - Ax \leq 0 \end{aligned}$$

Notice that the constraint $Ax \geq b$ implies m element-wise inequality constraints, and therefore is equivalent to

$$b[i] - (Ax)[i] \leq 0 \ , \ \forall i \in [m].$$

Let $\lambda \in \mathbb{R}^m_+$, where $\mathbb{R}^m_+$ denotes the set of m-dimensional vectors with non-negative elements.

Step 1: Get the Lagrangian:
The Lagrangian of the above problem is given as

$$L(x, \lambda) = \langle c, x \rangle + \langle \lambda, b - Ax \rangle.$$

After grouping terms we get the following

$$\begin{aligned} L(x, \lambda) &= \langle c, x \rangle + \langle \lambda, -Ax \rangle + \langle \lambda, b \rangle \\ &= \langle c, x \rangle + \langle A^\top \lambda, x \rangle + \langle \lambda, b \rangle \\ &= \langle c - A^\top \lambda, x \rangle + \langle \lambda, b \rangle. \end{aligned}$$

Step 2: Construct the dual function:
Next, we construct the dual function by minimizing the Lagrangian over the primal variable x, i.e.

$$g(\lambda) = \inf_{x \in \mathbb{R}^d} L(x, \lambda) = \inf_{x \in \mathbb{R}^d} \langle c - A^\top \lambda, x \rangle + \langle b, \lambda \rangle.$$

Minimizing over the values of x we have the following

$$g(\lambda) = \inf_x L(x, \lambda) = \begin{cases} b^\top \lambda & \text{, if } c = A^\top \lambda \\ -\infty & \text{, otherwise} \end{cases}$$

Step 3: Get the dual problem:
The last step is to get the dual problem, i.e.

$$\begin{aligned} &\sup_{\lambda \geq 0} g(\lambda) \\ &\text{s.t. } c = A^\top \lambda \end{aligned}$$

Bibliographic notes

More information can be found in Chapter 5 of Convex Optimization by Stephen Boyd and Lieven Vandenberghe [Boyd and Vandenberghe (2004)], and Chapter 5 of Algorithms for Convex Optimization. Nisheeth K. Vishnoi.[Vishnoi (2021)].

References

[Boyd and Vandenberghe (2004)] Boyd and Vandenberghe. Convex Optimization 2004.

[Vishnoi (2021)] Nisheeth K. Vishnoi. Algorithms for Convex Optimization. Cambridge University Press, 2021