ECE 273 Convex Optimization and Applications Instructor: Jun-Kun Wang
Scribe: Yilin Bao, Nicolas Schluep May 7, 2024
Editor/TA: Marialena Sfyraki

Lecture 11: Duality Theory Part II: KKT Conditions and Part III: Conjugate Function and Dual Problem

1 Review: Key Definitions

We begin be reviewing some key definitions from Lecture 10.

Definition 1. *(**Lagrangian**) Suppose we have an optimization problem with functional constraints:*

$$\inf_{x\in\mathbb{R}^d} f(x) \tag{1}$$

$$s.t.\ f_j(x) \le 0, \quad j = 1,\dots,m. \tag{2}$$

$$s.t.\ \ \textbf{\textit{affine}}\ \ h_i(x) = 0, \quad i = 1,\dots,p. \tag{3}$$

The Lagrangian for this optimization problem is:

$$L(x,\lambda,\mu) := f(x) + \sum_{j=1}^{m} \lambda_j f_j(x) + \sum_{i=1}^{p} \mu_i h_i(x), \tag{4}$$

where $\lambda \ge 0$.

Definition 2. *(**Dual Function**)* $g(\lambda,\mu) := \inf_{x\in\mathbb{R}^d} L(x,\lambda,\mu)$

Definition 3. *(**Dual Problem**)* $\sup_{\lambda\ge 0,\mu} g(\lambda,\mu)$

Definition 4. *(**KKT Conditions**) We say the primal variables $x_* \in \mathbb{R}^d$ and the dual variables $\lambda_* \in \mathbb{R}^m$, $\mu_* \in \mathbb{R}^p$ satisfy KKT conditions if*

- *(Primal feasibility) $\forall j \in [m] : f_j(x_*) \le 0$ and $\forall i \in [p] : h_i(x_*) = 0$.*
- *(Dual feasibility) $\lambda_* \ge 0$.*
- *(Stationarity) $\partial_x L(x_*,\lambda_*,\mu_*) = 0$.*
- *(Complementary slackness) $\forall j \in [m] : \lambda_j^* f_j(x_*) = 0$.*

2 Optimization problem and Dual problem

The goal of the optimization problem is to find the minimum value of a function $f(x)$ under given constraints. Here x is a d-dimensional real vector representing the variables of the optimization problem. The function $f(x)$ is called the objective function, and we aim to find the value of x that minimizes $f(x)$.

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

The constraints given can be classified into two groups, the inequality constraints and equality constraints. There are m inequality constraints, written as $f_j(x) \leq 0$, where $j = 1, \ldots, m$. There are p equality constraints, written as $h_i(x) = 0$, where $i = 1, \ldots, p$.

The Lagrangian function $L(x, \lambda, \mu)$ is a method to incorporate the constraints of the original optimization problem into the objective function, by introducing Lagrange multipliers (λ and μ) to account for these constraints. The Lagrangian function is defined as:

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

where $\lambda_j \geq 0$ are the Lagrange multipliers associated with the j-th inequality constraint, and μ_i are the multipliers associated with the i-th equality constraint.

In order to gain deeper insights into the problem's structure and potential solution strategies, we introduce the concept of dual problem to provide an alternative perspective to the primal optimization problem. The dual function is a crucial component in forming the dual problem. It is defined using the Lagrangian of the primal problem. Given a Lagrangian $L(x, \lambda, \mu)$, the dual function $g(\lambda, \mu)$ is expressed as:

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

This definition means that for fixed values of λ and μ, the dual function is the infimum (or greatest lower bound) of the Lagrangian over all possible x. In simpler terms, you evaluate the Lagrangian by minimizing it with respect to x while treating the Lagrange multipliers λ (associated with inequality constraints) and μ (associated with equality constraints) as constants.

Once the dual function is defined, the dual problem can be formulated as:

$$\sup_{\lambda \geq 0, \mu} g(\lambda, \mu)$$

This problem seeks to maximize the dual function over the space of all feasible Lagrange multipliers λ and μ. The maximization of the dual function and the minimization of the primal function satisfies the following:

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

We call it "weak duality". This always holds for any primal-dual problems. However, under certain conditions, we can find a stronger relationship between the primal and dual problems called "strong duality".

Strong duality is defined as the case where:

$$g\left(\lambda_*, \mu_*\right) = f\left(x_*\right),$$

where x_* is primal optimal; λ_* and μ_* are dual optimal.

And when will this special case appear? That's determined by the Karush-Kuhn-Tucker (KKT) conditions, which are necessary for deriving strong duality and will be discussed in the following section.

3 Strong Duality and the KKT Conditions

The Karush-Kuhn-Tucker (KKT) conditions are necessary conditions for a solution in certain optimization problems to be optimal, especially when the problem involves constraints. We say the primal variables $x_* \in \mathbb{R}^d$ and the dual variables $\lambda_* \in \mathbb{R}^m$, $\mu_* \in \mathbb{R}^p$ satisfy KKT conditions if

1. (Primal feasibility) $\forall j \in [m] : f_j\left(x_*\right) \leq 0$ and $\forall i \in [p] : h_i\left(x_*\right) = 0$.
2. (Dual feasibility) $\lambda_* \geq 0$.
3. (Stationarity) $\partial_x L\left(x_*, \lambda_*, \mu_*\right) = 0$.
4. (Complementary slackness) $\forall j \in [m] : \lambda_i^* f_j\left(x_*\right) = 0$.

Primal feasibility ensures that the solution x_* satisfies all the constraints of the original optimization problem: $f_j\left(x_*\right) \leq 0$ and $h_i\left(x_*\right) = 0$ must hold.

Dual feasibility requires that the dual variables associated with the inequality constraints be non-negative.

Stationarity $\partial_x L(x_*, \lambda_*, \mu_*) = 0$ involves the subgradient(s) (if not differentiable) or gradient (if differentiable) of the Lagrangian function with respect to the primal variables, stating that they must be zero at the optimal points.

Complementary slackness bridges the gap between the primal and dual problems, linking the dual variables and their corresponding primal constraints. This means that for each constraint, either the constraint is active (i.e., $f_j(x_*) = 0$) and the corresponding dual variable λ_j^* can be positive, or the constraint is inactive (i.e., $f_j(x_*) < 0$) and the corresponding dual variable must be zero.

4 Applications of the KKT Conditions: Projection

In the projection problem,

$$\text{Proj}_C(y) := \arg\min_{x \in C} \|y - x\|_2^2,$$

we choose a set C in the space, and for each point y, we try to find the closest point $x \in C$. Some typical examples are related to norm balls. For projecting onto the l_2-norm ball we have

$$C := \left\{ x \in \mathbb{R}^d : \|x\|_2 \leq 1 \right\},$$

The solution to this projection problem is derived in Lecture 10. We have that the solution is either $x = y$, when y is already in the ball, or $x = y/\|y\|$, when y is located outside. Hence, we have

$$\text{Proj}_C(y) = \frac{y}{\max\left\{1, \|y\|_2\right\}}$$

A more tricky example is the case where we want to project onto the l_1-norm ball

$$C := \left\{ x \in \mathbb{R}^d : \|x\|_1 \leq 1 \right\}.$$

If we denote $x = \text{Proj}_c(y)$, then $x = y$ if $\|y\|_1 \leq 1$; otherwise,

$$x[i] = \text{sign}(y[i])(|y[i]| - \lambda)_+, \forall i \in [d]$$

where λ is a number such that $\sum_{i=1}^d (|y[i]| - \lambda)_+ = 1$ and $(z)_+ := \max\{0, z\}$.

This derivation was mostly completed in Lecture 10. Here we will add the final parts. By the complementary slackness of KKT conditions, we have the inequality constraint (only one)

$$f_1(x) = \|x\|_1 - 1 \leq 0,$$

so complementary slackness gives

$$\lambda \left(\|x\|_1 - 1 \right) = 0.$$

Then there are two possible cases. The first case is if the constraint is inactive and dual variable is zero $\lambda = 0$, then we can say the point y is already in the norm ball. Writing this mathematically

$$\|x\|_1 < 1 \Rightarrow \lambda = 0 \Rightarrow x[i] = y[i].$$

The second case is if the constraint is active $\lambda \neq 0$. In order to satisfy the complementary slackness condition, we have that

$$\|x\|_1 = \sum_{i=1}^{d} |x[i]| = 1,$$

so together with the previously derived primal feasibility condition,

$$\sum_{i=1}^{d} \left| \text{sign}(y[i])(|y[i]| - \lambda)_+ \right| = 1,$$

which is equivalent to

$$\sum_{i=1}^{d} (|y[i]| - \lambda)_+ = 1.$$

Hence, our solution in this case will be of the form

$$x[i] = \text{sign}(y[i])(|y[i]| - \lambda)_+, \forall i \in [d],$$

where λ must satisfy

$$h(\lambda) := \sum_{i=1}^{d} (|y[i]| - \lambda)_+ - 1,$$

Notice that to find a solution in this case, we need to find the roots of $h(\lambda)$.

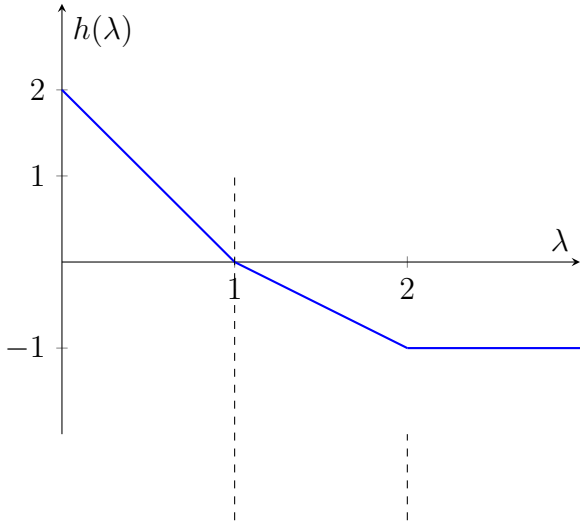
The example used in lecture for illustration is, when

$$y = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

then the piece-wise linear function $h(\lambda)$ is

$$h(\lambda) = (1-\lambda)_+ - 1 + (2-\lambda)_+ = \begin{cases} -1, & \text{if } \lambda \geq 2 \\ 1-\lambda, & \text{if } \lambda \in [1,2] \\ 2-2\lambda, & \text{if } \lambda \leq 1 \end{cases}$$

Observe that $h(\lambda)$ is a piece-wise, continuous and decreasing function. The following figure is a visualization of this piece-wise linear function $h(\lambda)$.



Notice that in this example, only $\lambda = 1$ is a root of $h(\lambda)$. This means that our solution $x[i] = \text{sign}(y[i])(|y[i]| - \lambda)_+, \forall i \in [d]$ for $\|y\|_1 > 1$ will be unique. This is explained by the fact that $\|y - x\|_2^2$ is a strongly convex function and $\|x\|_1 \leq 1$ is a closed convex set, which means that we must have a unique solution.

Remark: The reason why the KKT conditions are useful for solving projection problems is because of the KKT theorem, which we will prove in the next section. We can use the KKT theorem for projection problems such as $\min_{x \in C} \|y - x\|_2^2$, s.t. $\|x\|_1 \leq 1$ since both $f(x) = \|y - x\|_2^2$ and $f_1(x) = \|x\|_1 - 1$ are convex functions. Hence, by the KKT theorem, the KKT conditions being satisfied implies strong duality. Therefore, the x_* and the λ_*, μ_* that satisfy the KKT conditions will be primal and dual optimal, respectively.

5 KKT Theorem Proof

Theorem 1. ***(Strong Duality and KKT Conditions) Strong duality****, i.e.,*

$$f(x_*) = g(\lambda_*, \mu_*)$$

implies that x_*, λ_*, μ_* *satisfy* ***KKT conditions****. Furthermore, if* $f(\cdot), f_1(\cdot), \ldots, f_m(\cdot)$ *are* ***convex*** *and* $h_1(\cdot)$, $h_2(\cdot)$, $\ldots, h_p(\cdot)$ *are affine, then the converse is true:* ***KKT conditions implies the strong duality****.*

Proof. To begin, we want to prove that strong duality implies that x_*, λ_*, μ_* satisfy the KKT conditions. By the definition of strong duality we have that:

$$
\begin{aligned}
f(x_*) &= g(\lambda_*, \mu_*) && (5)\\
&= \inf_x L(x, \lambda_*, \mu_*) && \text{, by definition of the dual function} \quad (6)\\
&= \inf_x \left(f(x) + \sum_{j=1}^{m} \lambda_j^* f_j(x) + \sum_{i=1}^{p} \mu_i^* h_i(x) \right) && \text{, by definition of the Lagrangian} \quad (7)\\
&\leq f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) && \text{, by definition of the infimum} \quad (8)
\end{aligned}
$$

Now, since x_*, λ_*, μ_* are assumed to be feasible points, we have that $f_j(x_*) \leq 0, \lambda_j^* \geq 0, \forall j \in [m]$ and $h_i(x_*) = 0, \forall i \in [p]$. Hence, by primal and dual feasibility, we have the inequality:

$$
f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) \leq f(x_*) \tag{9}
$$

The result $f(x_*) \leq f(x_*)$ indicates that these should all be equalities. Hence:

$$
\begin{aligned}
f(x_*) &= g(\lambda_*, \mu_*) && (10)\\
&= \inf_x L(x, \lambda_*, \mu_*) && (11)\\
&= \inf_x \left(f(x) + \sum_{j=1}^{m} \lambda_j^* f_j(x) + \sum_{i=1}^{p} \mu_i^* h_i(x) \right) && (12)\\
&= f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) && (13)\\
&= f(x_*). && (14)
\end{aligned}
$$

From the equality:

$$
f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) = f(x_*),
$$

and since we assume that $h_i(x_*) = 0, \forall i \in [p]$, we have that:

$$\sum_{j=1}^{m} \lambda_j^* f_j(x_*) = 0, \quad \sum_{i=1}^{p} \mu_i^* h_i(x_*) = 0. \tag{15}$$

Since we know that $f_j(x_*) \leq 0, \lambda_j^* \geq 0, \forall j \in [m]$, we have that $\lambda_j^* f_j(x_*) \leq 0, \forall j \in [m]$. This along with (15) implies:

$$\lambda_j^* f_j(x_*) = 0, \forall j \in [m]. \tag{16}$$

Hence, complementary slackness is satisfied.
Now, the equality:

$$\inf_x \left(f(x) + \sum_{j=1}^{m} \lambda_j^* f_j(x) + \sum_{i=1}^{p} \mu_i^* h_i(x) \right) = f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*).$$

implies that x_* is a stationary point of $L(x, \lambda_*, \mu_*)$. Therefore:

$$\nabla_x L(x_*, \lambda_*, \mu_*) = 0. \tag{17}$$

Hence, the stationarity condition is satisfied. The primal feasibility and dual feasibility conditions are satisfied since they were assumed to hold throughout this proof. This means that the KKT conditions are satisfied.

We now want to show that if x_*, λ_*, μ_* satisfy the KKT conditions, and if $f(\cdot), f_1(\cdot), ..., f_m(\cdot)$ are convex, and if $h_1(\cdot), ..., h_p(\cdot)$ are affine, then strong duality holds. By the definition of the dual function we have:

$$g(\lambda_*, \mu_*) = \inf_x L(x, \lambda_*, \mu_*). \tag{18}$$

Now, since the conical combination (i.e., linear combination with non-negative coefficients) of convex functions is convex, we know that the Lagrangian is convex with respect to x. Furthermore, by the stationarity KKT condition $\partial_x L(x_*, \lambda_*, \mu_*) = 0$. Hence, we know that x_* must be a global minima of the Lagrangian, resulting in the equality:

$$\inf_x L(x, \lambda_*, \mu_*) = L(x_*, \lambda_*, \mu_*) \tag{19}$$

By the definition of the Lagrangian in this case:

$$L(x_*, \lambda_*, \mu_*) = f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) \tag{20}$$

By the primal feasibility (KKT condition) $h_i(x_*) = 0, \forall i \in [p]$:

$$f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) + \sum_{i=1}^{p} \mu_i^* h_i(x_*) = f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) \tag{21}$$

By the complementary slackness (KKT condition) we have that $\forall j \in [m], \lambda_j^* f_j(x_*) = 0$. Hence:

$$f(x_*) + \sum_{j=1}^{m} \lambda_j^* f_j(x_*) = f(x_*). \tag{22}$$

Finally, by the equalities above, we can conclude that:

$$g(\lambda_*, \mu_*) = f(x_*). \tag{23}$$

Therefore, strong duality holds. □

6 Fenchel Conjugate

Definition 5. *(**Fenchel Conjugate**) Consider a function $f(\cdot)$, then the Fenchel Conjugate is defined to be*

$$f^*(y) = \sup_{x \in \text{dom}(f)} \left(y^\top x - f(x) \right).$$

Theorem 2. *The conjugate function $f^*(y)$ is always convex, even if $f(\cdot)$ is non-convex.*

Proof. Denote $q_x(y) = y^T x - f(x)$. Notice that $q_x(y)$ is an affine function. Now suppose that we have points $y_1, y_2 \in dom(f)$. For $\alpha \in [0, 1]$, we have:

$$f^*((1-\alpha)y_1 + \alpha y_2) = \sup_{x \in dom(f)} q_x((1-\alpha)y_1 + \alpha y_2) \tag{24}$$

$$= \sup_{x \in dom(f)} (1-\alpha) q_x(y_1) + \alpha q_x(y_2), \tag{25}$$

where the last equality follows by definition of the affine function. Now notice that $q_x(y_1)$ and $q_x(y_2)$ may have different maximizing x values. Hence, the sum of their individual supremums may be greater than the supremum of their sum, which requires them to have the same x value. Therefore:

$$\sup_{x \in dom(f)} (1-\alpha) q_x(y_1) + \alpha q_x(y_2) \leq (1-\alpha) \sup_{x \in dom(f)} q_x(y_1) + \alpha \sup_{x \in dom(f)} q_x(y_2). \tag{26}$$

Hence:

$$f^*((1-\alpha)y_1 + \alpha y_2) \leq (1-\alpha) f^*(y_1) + \alpha f^*(y_2). \tag{27}$$

□

7 Conjugate Function Application: Dual Formulation of Empirical Risk Minimization (ERM)

Example: One application of the conjugate function is the dual formulation of empirical risk minimization (ERM). Suppose that we have a dataset $\{(z_i, y_i)\}_{i=1}^n$, where $z_i \in \mathbb{R}^d$ is a feature, $y_i \in \mathbb{R}$ is a label and n is the number of datapoints. The primal problem is:

$$\min_{x \in \mathbb{R}^d} F(x), \quad \text{where } F(x) := \frac{1}{n} \sum_{i=1}^n f_i(x^\top z_i) + \frac{\lambda}{2} \|x\|_2^2. \tag{28}$$

We want to show that the corresponding dual problem is:

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha), \quad \text{where } D(\alpha) := \frac{1}{n} \sum_{i=1}^n -f_i^*(-\alpha_i) - \frac{\lambda}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i z_i \right\|_2^2. \tag{29}$$

To begin, consider the following constrained optimization problem:

$$\begin{aligned} &\min_{x \in \mathbb{R}^d; (\theta_i)_{i=1}^n} \sum_{i=1}^n f_i(\theta_i) + \frac{\lambda n}{2} \|x\|_2^2 \\ &\text{subject to } \forall i, \theta_i = z_i^\top x \end{aligned} \tag{30}$$

Step 1: Construct the Lagrangian. Denote $\{\theta_i\}_{i=1}^n = \theta$ and $\{\alpha_i\}_{i=1}^n = \alpha$. Hence:

$$L(x, \theta, \alpha) = \sum_{i=1}^n f_i(\theta_i) + \frac{\lambda n}{2} \|x\|_2^2 + \sum_{i=1}^n \alpha_i (\theta_i - z_i^T x). \tag{31}$$

Step 2: Optimize over primal variables to get the dual function. By definition, the dual function is $g(\alpha) = \inf_{x,\theta} L(x, \theta, \alpha)$. Hence, we have:

$$\min_{x, \theta_1, \ldots, \theta_n} \sum_{i=1}^n (f_i(\theta_i) + \alpha_i \theta_i) + \frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x \tag{32}$$

$$\Leftrightarrow \min_x \min_{\theta_1, \ldots, \theta_n} \left(\sum_{i=1}^n \big(f_i(\theta_i) + \alpha_i \theta_i \big) \right) + \frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x \tag{33}$$

$$\Leftrightarrow \min_x \left(\sum_{i=1}^n \left(\min_{\theta_i} f_i(\theta_i) + \alpha_i \theta_i \right) \right) + \frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x. \tag{34}$$

Now, we want to rewrite this expression so that we can substitute in the conjugate function. The conjugate function is defined as $f^*(y) = \sup_{x \in \text{dom}(f)} \big(y^\top x - f(x)\big)$. Hence, we want to write the inner minimizations as maximizations. We know that:

$$\left(\min_{\theta_i} f_i(\theta_i) + \alpha_i \theta_i \right) = - \max_{\theta_i} - \big(f_i(\theta_i) + \alpha_i \theta_i \big) . \tag{35}$$

Plugging this into (34), we have:

$$\min_x \sum_{i=1}^n (- \max_{\theta_i} - \big(f_i(\theta_i) + \alpha_i \theta_i \big)) + \frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x. \tag{36}$$

By the definition of the conjugate function, we have $f_i^*(-\alpha_i) = \max_{\theta_i}(-\alpha_i \theta_i - f_i(\theta_i))$. Hence, we have equivalently:

$$- \sum_{i=1}^n f_i^*(-\alpha_i) + \min_x (\frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x). \tag{37}$$

Now define $\Phi(x) = \frac{\lambda n}{2} \|x\|_2^2 - \sum_{i=1}^n \alpha_i z_i^\top x$. Notice that $\min_x \Phi(x)$ is an unconstrained optimization problem for a convex function with respect to x. Hence, we can find a closed-form solution to this optimization problem by finding its stationary point:

$$\nabla \Phi(x) = 0 \tag{38}$$

$$\Leftrightarrow \lambda n x - \sum_{i=1}^n \alpha_i z_i = 0 \tag{39}$$

$$\Leftrightarrow x = \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i z_i. \tag{40}$$

Hence, our closed-form solution is:

$$\min_x \Phi(x) = \frac{\lambda n}{2} \| \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i z_i \|_2^2 - \langle \sum_{i=1}^n \alpha_i z_i, \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i z_i \rangle \tag{41}$$

$$= \frac{-1}{2 \lambda n} \| \sum_{i=1}^n \alpha_i z_i \|_2^2. \tag{42}$$

Finally, we can plug this back into our original optimization problem over the primal variables to get the dual function:

$$g(\alpha) = - \sum_{i=1}^n f_i^*(-\alpha_i) - \frac{\lambda n}{2} \left\| \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i z_i \right\|_2^2 . \tag{43}$$

Now, since $g(\alpha)$ was found for the original function $\sum_{i=1}^n f_i(x^\top z_i) + \frac{\lambda n}{2}\|x\|_2^2$, but in the problem statement we have that $F(x) := \frac{1}{n}\sum_{i=1}^n f_i(x^\top z_i) + \frac{\lambda}{2}\|x\|_2^2$, we need to multiply this result by $\frac{1}{n}$ to get the desired dual function:

$$D(\alpha) = \frac{1}{n}g(\alpha) = \frac{1}{n}\sum_{i=1}^n -f_i^*(-\alpha_i) - \frac{\lambda}{2}\left\|\frac{1}{\lambda n}\sum_{i=1}^n \alpha_i z_i\right\|_2^2. \tag{44}$$

8 Duality Gap

Definition 6. *(**Duality Gap**) Suppose that we have the primal problem* $\min_x F(x)$ *and the dual problem* $\max_\alpha D(\alpha)$*. We define:*

$$\textit{Duality gap} := F(x(\alpha)) - D(\alpha).$$

Theorem 3. *The primal optimality gap* $F(x(\alpha)) - F_*$ *is upper-bounded by the duality gap* $:= F(x(\alpha)) - D(\alpha)$*.*

Proof. By weak duality, we know that:

$$D(\alpha) \leq \min_x F(x) = F_*.$$

Hence:

$$F(x(\alpha)) - D(\alpha) \geq F(x(\alpha)) - F_*.$$

□

Remark: Notice that $F(x(\alpha)) - D(\alpha)$ is easily computable. Hence, if we design algorithms in the dual space (e.g., updating α instead of x directly), we can compute the upper-bound for the current primal optimality gap at each iteration. We can then use this upper-bound to stop the algorithm when we have achieved an at most ϵ primal optimality gap. An example of an algorithm in the dual space is Stochastic Dual Coordinate Ascent (SDCA), which is introduced in the next section.

9 Stochastic Dual Coordinate Ascent (SDCA)

To begin, the basic coordinate descent algorithm is given by:

Algorithm 1 COORDINATE DESCENT

1: **for** $k = 1, 2, \ldots$ **do**
2: Randomly pick a coordinate $i_k \in [d]$.
3: $x_{k+1}[i_k] = x_k[i_k] - \eta \nabla f(x_k)[i_k]$. only the i_k element is updated at a time.
4: **end for**

Remark: In general, the iteration cost of coordinate descent will be less than that of gradient descent. However, the iteration complexity to achieve an ϵ gap for coordinate descent is greater than or equal to that of gradient descent [Wright (2015)].

Question: Is the iteration cost of coordinate descent always $\frac{1}{d}$ times that of gradient descent?
Answer: Not necessarily. For $f(x) = \frac{1}{2}x^T Ax - b^T x$, the iteration cost of coordinate descent is $\frac{1}{d}$ times that of gradient descent. However, this is not the case for all functions $f(\cdot)$.

We can now turn our attention to the Stochastic Dual Coordinate Ascent (SDCA) algorithm when applied to the empirical risk minimization (ERM) problem discussed previously. The main idea behind this algorithm is that instead of solving the primal minimization problem, we solve the dual maximization problem:

$$\max_{\alpha \in \mathbb{R}^n} D(\alpha), \qquad \text{where } D(\alpha) := \frac{1}{n}\sum_{i=1}^{n} -f_i^*(-\alpha_i) - \frac{\lambda}{2}\left\| \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i z_i \right\|_2^2, \tag{45}$$

Consider updating a dual variable $\alpha_i \in \mathbb{R}^n$ at a time. That is, at the k-th iteration, we pick $i_k \in [n]$. Then, we have

$$\begin{aligned}
&\max_{\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\alpha_{i_k}\right) - \frac{\lambda}{2}\left\| \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i z_i \right\|_2^2 \\
&\Leftrightarrow \max_{\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\alpha_{i_k}\right) - \frac{\lambda}{2}\left\| \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i^{(k-1)} z_i + \frac{1}{\lambda n}\Delta\alpha_{i_k} z_{i_k} \right\|_2^2 \\
&\Leftrightarrow \max_{\Delta\alpha_{i_k}} -\frac{1}{n} f_{i_k}^*\left(-\left(\alpha_{i_k}^{(k-1)} + \Delta\alpha_{i_k}\right)\right) - \frac{\lambda}{2}\left\| x^{(k-1)} + \frac{1}{\lambda n}\Delta\alpha_{i_k} z_{i_k} \right\|_2^2,
\end{aligned}$$

where $\alpha_{i_k} = \underbrace{\alpha_{i_k}^{(k-1)}}_{\text{fixed}} + \underbrace{\Delta\alpha_{i_k}}_{\text{variable}}$ and $x^{(k-1)} = \frac{1}{\lambda n}\sum_{i=1}^{n} \alpha_i^{(k-1)} z_i$.

The algorithm proposed in [Shalev-Shwartz & Zhang (2013)] to solve this optimization problem is:

Algorithm 2 SDCA FOR ERM

1: Initialize dual variables $\alpha^{(1)} \in \mathbb{R}^n$.
2: **for** $k = 1, 2, \ldots, K$ **do**
3: Randomly pick a dual coordinate $i_k \in [n]$.
4: Maximizes the dual problem by updating the dual variable i_k while fixing the others

$$\max_{\Delta\alpha_{i_k}} -\frac{1}{n} f_{i_k}^* \left(-\left(\alpha_{i_k}^{(k-1)} + \Delta\alpha_{i_k} \right) \right) - \frac{\lambda}{2} \left\| x^{(k-1)} + \frac{1}{\lambda n} \Delta\alpha_{i_k} z_{i_k} \right\|_2^2 . \tag{46}$$

5: $\alpha^{(k)} = \alpha^{(k-1)} + \Delta\alpha_{i_k} e_{i_k} \in \mathbb{R}^n$.
6: $x^{(k)} = x^{(k-1)} + \frac{1}{\lambda n} \Delta\alpha_{i_k} z_{i_k} \in \mathbb{R}^d$.
7: **end for**
8: Output: $x(\alpha^{(K)}) := \frac{1}{\lambda n} \sum_{i=1}^n \alpha_i^{(K)} z_i$.

Some results from applying the SDCA algorithm to different optimization problems are shown in Figure 1 [Johnson & Zhang (2013)].

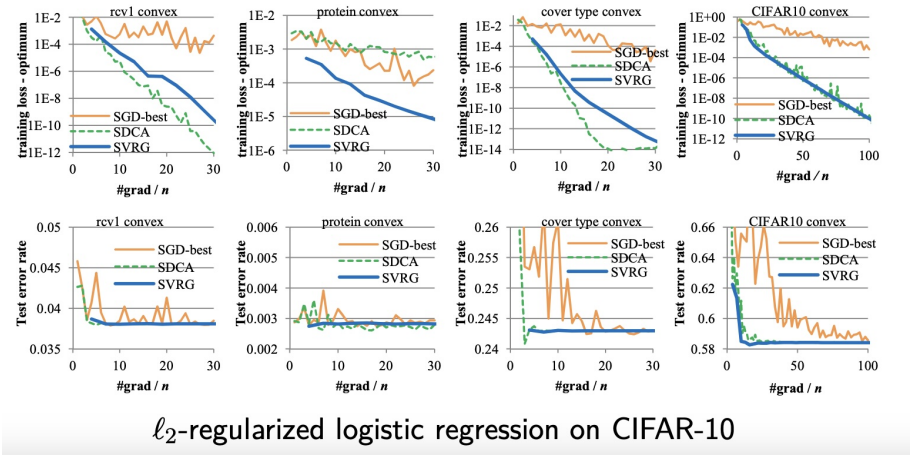


Figure 1: Performance comparison of the SGD-best, SDCA, and SVRG algorithms on various optimization problems.

Example: We want to determine the update $\Delta\alpha_{i_k}$ of the SDCA for ERM algorithm when using hinge loss. Hence, let us consider $f_i(\theta) := \max\{0, 1 - y_i\theta\}$ being the hinge

loss, where $y_i \in \{+1, -1\}$. Its conjugate function is

$$f_i{}^*(a) = \begin{cases} ay_i & \text{, if } ay_i \in [-1, 0] \\ \infty & \text{, otherwise} \end{cases} .$$

It can be shown that the update of SDCA for the hinge loss is

$$\Delta\alpha_{i_k} = y_{i_k} \max\left(0, \max\left(1, \frac{1 - z_{i_k}^\top x^{(k-1)} y_{i_k}}{\|z_{i_k}\|_2^2/\lambda n} + \alpha_{i_k}^{(k-1)} y_{i_k}\right)\right) - \alpha_{i_k}^{(k-1)}. \qquad (47)$$

Bibliographic Notes

More information about duality theory can be found in Chapter 5 of [Boyd & Vandenberghe (2004)] and Chapter 5 of [Vishnoi (2021)].

References

[Wright (2015)] Stephen J. Wright. Coordinate Descent Algorithms. 2015.

[Shalev-Shwartz & Zhang (2013)] Shai Shalev-Shwartz and Tong Zhang. Stochastic Dual Coordinate Ascent Methods for Regularized Loss Minimization. JMLR 2013.

[Johnson & Zhang (2013)] Rie Johnson and Tong Zhang. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. NeurIPS 2013.

[Boyd & Vandenberghe (2004)] Stephen Boyd and Lieven Vandenberghe. Convex Optimization. Cambridge University Press, 2004.

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