



Karash-Kuhn-Tucker conditions

MFDS Team



BITS Pilani

Pilani | Dubai | Goa | Hyderabad



- ▶ We will look at constrained optimization and Lagrange multipliers.
- ▶ We will look at primal and dual problems and how their solutions are related.
- ▶ We will set up Karash-Kuhn-Tucker conditions.



- ▶ Consider the following problem: $\min_{\mathbf{x}} f(\mathbf{x}), f: \mathbb{R}^D \rightarrow \mathbb{R}$, subject to additional constraints - so we are looking at a minimization problem except that the set of all \mathbf{x} over which minimization is performed is not all of \mathbb{R}^D .
- ▶ The constrained problem becomes $\min_{\mathbf{x}} f(\mathbf{x})$ subject to $g_i(\mathbf{x}) \leq 0 \forall i=1, 2, \dots, m$.
- ▶ Since we have a method of finding a solution to the unconstrained optimization problem, one way to proceed now is to convert the given constrained optimization problem into an unconstrained one.
- ▶ We construct $J(\mathbf{x}) = f(\mathbf{x}) + \sum_{i=1}^m \mathbf{1}(g_i(\mathbf{x}))$, where $\mathbf{1}(z) = 0$ for $z \leq 0$ and $\mathbf{1}(z) = \infty$ for $z > 0$.



- ▶ The formulation of $J(\mathbf{x})$ in the previous slide ensures that its value is infinity if any one of the constraints $g_i(\mathbf{x})$ is not satisfied. This ensures that the optimal solution to the unconstrained problem is the same as the constrained problem.
- ▶ The step-function is also difficult to optimize, and our solution is to replace the step-function by a linear function using Lagrange multipliers.
- ▶ We create the Lagrangian of the given constrained optimization problem as follows:
$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{i=m} \lambda_i g_i(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{g}(\mathbf{x}), \text{ where } \lambda_i \geq 0 \text{ for all } i.$$



- ▶ The primal problem is $\min f(\mathbf{x})$ subject to $g_i(\mathbf{x}) \leq 0, 1 \leq i \leq m$. Optimization is performed over the primal variables \mathbf{x} .
- ▶ The associated Lagrangian dual problem is $\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \mathfrak{D}(\boldsymbol{\lambda})$ subject to $\boldsymbol{\lambda} \geq 0$ where $\boldsymbol{\lambda}$ are dual variables.
- ▶ $\mathfrak{D}(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$.
- ▶ The following minimax inequality holds over two arguments \mathbf{x}, \mathbf{y} : $\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) \leq \min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y})$.



- ▶ Why is this inequality true?
- ▶ Assume that \mathbf{x}, \mathbf{y} : $\max_{\mathbf{y}} \min_{\mathbf{x}} \phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}_A, \mathbf{y}_A)$ and $\min_{\mathbf{x}} \max_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}_B, \mathbf{y}_B)$.
- ▶ Fixing \mathbf{y} at \mathbf{y}_A we see that the inner operation on the left hand side of the minimax inequality is a min operation over \mathbf{x} and returns \mathbf{x}_A . Thus we have $\phi(\mathbf{x}_A, \mathbf{y}_A) \leq \phi(\mathbf{x}_B, \mathbf{y}_A)$.
- ▶ Fixing \mathbf{x} at \mathbf{x}_B we see that the inner operation on the right hand side of the minimax inequality is a max operation over \mathbf{y} and returns \mathbf{y}_B . Thus we have $\phi(\mathbf{x}_B, \mathbf{y}_B) \geq \phi(\mathbf{x}_B, \mathbf{y}_A)$.
- ▶ From the above we conclude that $\phi(\mathbf{x}_B, \mathbf{y}_B) \geq \phi(\mathbf{x}_A, \mathbf{y}_A)$.



- ▶ The difference between $J(\mathbf{x})$ and the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is that the indicator function is relaxed to a linear function.
- ▶ When $\boldsymbol{\lambda} \geq 0$, the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is a lower bound on $J(\mathbf{x})$.
- ▶ The maximum of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ is $J(\mathbf{x})$ - if the point \mathbf{x} satisfies all the constraints $g_i(\mathbf{x}) \leq 0$, then the maximum of the Lagrangian is obtained at $\boldsymbol{\lambda} = 0$ and it is equal to $J(\mathbf{x})$. If one or more constraints is violated such that $g_i(\mathbf{x}) > 0$, then the associated Lagrangian coefficient λ_i can be taken to be ∞ so as to equal $J(\mathbf{x})$.



- ▶ From the previous slide, we have $J(\mathbf{x}) = \max_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda)$.
- ▶ Our original constrained optimization problem boiled down to minimizing $J(\mathbf{x})$, in other words we are looking at $\min_{\mathbf{x} \in \mathbb{R}^d} \max_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda)$
- ▶ Using the minimax inequality we see that $\min_{\mathbf{x} \in \mathbb{R}^d} \max_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda) \geq \max_{\lambda \geq 0} \min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda)$.
- ▶ This is known as weak duality. The inner part of the right hand side of the inequality is $\mathcal{D}(\lambda)$, and the inequality above is the reason for setting up the associated Lagrangian dual problem for the original constrained optimization problem.



- ▶ In contrast to the original formulation $\mathfrak{D}(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda)$ is an unconstrained optimization problem for a given value of λ .
- ▶ We observe that $\mathfrak{D}(\lambda) = \min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \lambda)$ is a point-wise minimum of affine functions and hence $\mathfrak{D}(\lambda)$ is concave even though $f()$ and $g()$ may be nonconvex.
- ▶ We have obtained a Lagrangian formulation for a constrained optimization problem where the constraints are inequalities. What happens when some constraints are equalities?



- ▶ Suppose the problem is $\min_{\mathbf{x}} f(\mathbf{x})$ subject to $g_i(\mathbf{x}) \leq 0$ for all $1 \leq i \leq m$ and $h_j(\mathbf{x}) = 0$ for $1 \leq j \leq n$.
- ▶ We model the equality constraint $h_j(\mathbf{x}) = 0$ with two inequality constraints $h_j(\mathbf{x}) \geq 0$ and $h_j(\mathbf{x}) \leq 0$.
- ▶ The resulting Lagrange multipliers are then unconstrained.
- ▶ The Lagrange multipliers for the original inequality constraints are non-negative while those corresponding to the equality constraints are unconstrained.



- ▶ We are interested in a class of optimization problems where we can guarantee global optimality.
- ▶ When $f()$, the objective function, is a convex function and $g()$ and $h()$ are convex functions, we have a convex optimization problem.
- ▶ In this setting we have strong duality - the optimal solution of the primal problem is equal to the optimal solution of the dual problem.
- ▶ What is a convex function?



- ▶ First we need to know what is a convex set. A set C is a convex set if for any $x, y \in C$, $\theta x + (1 - \theta)y \in C$ where $0 \leq \theta \leq 1$.
- ▶ For any two points lying in the convex set, a line joining them lies entirely in the convex set.
- ▶ Let a function $f : \mathbb{R}^d \rightarrow R$ be a function whose domain is a convex set C .
- ▶ The function is a convex function if for any $\mathbf{x}, \mathbf{y} \in C$,
$$f(\theta \mathbf{x} + (1 - \theta) \mathbf{y}) \leq \theta f(\mathbf{x}) + (1 - \theta) f(\mathbf{y})$$
- ▶ Another way of looking at a convex function is to use the gradient: for any two points \mathbf{x} and \mathbf{y} , we have
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla_{\mathbf{x}} f(\mathbf{x})(\mathbf{y} - \mathbf{x}).$$

- ▶ The negative entropy, a useful function in Machine Learning, is convex: $f(x) = x \log_2 x$ for $x > 0$.
- ▶ First let us check if $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$. Take $x = 2$, $y = 4$, and $\theta = 0.5$ to get
 $f(0.5 * 2 + 0.5 * 4) = f(3) = 3 \log_2 3 \approx 4.75$. Then
 $\theta f(2) + (1 - \theta)f(4) = 0.5 * 2 \log_2 2 + 0.5 * 4 \log_2 4 = \log_2 32 = 5$.
Therefore the convexity criterion is satisfied for these two points.
- ▶ Let us now use the gradient criterion. We have
 $\nabla f(x) = \log_2 x + x \frac{1}{x \log_e 2}$. Calculating $f(2) + \nabla f(2) * (4 - 2)$
gives $2 \log_2 2 + (\log_2 2 + \frac{1}{\log_e 2} * 2) \approx 6.9$. We see that
 $f(4) = 4 \log_2 4 = 8$ which shows that the gradient criterion is also satisfied.



- ▶ Let us look at a convex optimization problem where the objective function and constraints are all linear.
- ▶ Such a convex optimization problem is called a linear programming problem.
- ▶ We can express a linear programming problem as $\min_{\mathbf{x}} \mathbf{c}^T \mathbf{x}$ subject to $\mathbf{Ax} \leq \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{m \times d}$ and $\mathbf{b} \in \mathbb{R}^{m \times 1}$.
- ▶ The Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is given by $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}^T (\mathbf{Ax} - \mathbf{b})$ where $\boldsymbol{\lambda} \in \mathbb{R}^m$ is the vector of non-negative Lagrangian multipliers.
- ▶ We can rewrite the Lagrangian as $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b}$.



- ▶ Taking the derivative of the Lagrangian with respect to \mathbf{x} and setting it to zero we get $\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} = 0$.
- ▶ Since $\mathcal{D}(\boldsymbol{\lambda}) = \min_{\mathbf{x} \in \mathbb{R}^d} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$, plugging in the above equation gives $\mathcal{D}(\boldsymbol{\lambda}) = -\boldsymbol{\lambda}^T \mathbf{b}$.
- ▶ We would like to maximize $\mathcal{D}(\boldsymbol{\lambda})$, subject to the constraint $\boldsymbol{\lambda} \geq 0$.
- ▶ Thus we end up with the following problem:

$$\begin{aligned} \max_{\boldsymbol{\lambda} \in \mathbb{R}^m} & -\boldsymbol{\lambda}^T \mathbf{b} \\ \text{subject to} & \mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda} = 0 \\ & \boldsymbol{\lambda} \geq 0 \end{aligned}$$



- ▶ We can solve the original primal linear program or the dual one - the optimum in each case is the same.
- ▶ The primal linear program is in d variables but the dual is in m variables, where m is the number of constraints in the original primal program.
- ▶ We choose to solve the primal or dual based on which of m or d is smaller.



- ▶ We now consider the case of a quadratic objective function subject to affine constraints:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \frac{1}{2} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{c}^T \mathbf{x} \text{ subject to}$$
$$\mathbf{A} \mathbf{x} \leq \mathbf{b}$$

- ▶ Here $\mathbf{A} \in \mathbb{R}^{m \times d}$, $\mathbf{b} \in \mathbb{R}^m$, $\mathbf{c} \in \mathbb{R}^d$



- ▶ The Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ is given by $\frac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + \mathbf{c}^T \mathbf{x} + \boldsymbol{\lambda}^T (\mathbf{A}\mathbf{x} - \mathbf{b})$.
- ▶ Rearranging the above we have $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^T \mathbf{Q}\mathbf{x} + (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{x} - \boldsymbol{\lambda}^T \mathbf{b}$
- ▶ Taking the derivative of $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ and setting it equal to zero gives $\mathbf{Q}\mathbf{x} + (\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda}) = \mathbf{0}$.
- ▶ If we take \mathbf{Q} to be invertible, we have $\mathbf{x} = \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})$.
- ▶ Plugging this value of \mathbf{x} into $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda})$ gives us $\mathcal{D}(\boldsymbol{\lambda}) = -\frac{1}{2}(\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbf{b}$.
- ▶ This gives us the dual optimization problem:
 $\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} -\frac{1}{2}(\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda})^T \mathbf{Q}^{-1}(\mathbf{c} + \mathbf{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \mathbf{b}$ subject to $\boldsymbol{\lambda} \geq \mathbf{0}$.

- ▶ The minimax inequality establishes weak duality which states that the optimal solution of the primal problem is greater than or equal to that of the dual problem.
- ▶ When equality holds, this becomes strong duality.
- ▶ Strong duality is useful in that one can solve the dual problem to get the same solution as solving the primal problem.
- ▶ Solving the dual problem may be easier.
- ▶ When does strong duality hold?



We shall work with the following optimization problem:

$$\begin{aligned} &\text{minimize } f(\mathbf{x}) \text{ subject to} \\ &g_i(\mathbf{x}) \leq 0 \quad \forall i \in [m] \\ &h_j(\mathbf{x}) = 0 \quad \forall j \in [p] \end{aligned}$$

The Lagrangian associated with this optimization problem is

$$\text{minimize } f(\mathbf{x}) + \sum_{i=1}^{i=m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{j=p} \nu_j h_j(\mathbf{x})$$

The λ_i s and h_j s are called Lagrange multipliers.





Given a Lagrangian $L(\mathbf{x}, \lambda, \nu)$ over some optimization domain D , the Lagrangian dual is the function $F(\lambda, \nu) = \inf_{\mathbf{x} \in D} L(\mathbf{x}, \lambda, \nu)$. The dual optimization problem is

$$\begin{aligned} & \max F(\lambda, \nu) \\ & \text{subject to } \lambda \geq 0 \end{aligned}$$

- ▶ For a primal optimization problem we say that it obeys Slater's condition if the objective function f is convex, the constraints g_i are all convex and the constraint functions h_j are all linear and there exists a point \bar{x} in the interior of the region, i.e. $g_i(\bar{x}) < 0$ for all $i \in [m]$ and $h_j(\bar{x}) = 0$ for all $j \in [p]$.
- ▶ **Theorem: Suppose Slater's condition holds and the region has a non-empty interior. Then we have strong duality.**



Let us define two sets $A = \{(\mathbf{u}, \mathbf{v}, t) | \exists \mathbf{x} \in D \text{ such that } g_i(\mathbf{x}) \leq u_i, i = 1 \dots m, h_i(\mathbf{x}) = v_i, i = 1 \dots p, f(\mathbf{x}) \leq t\}$ and $B = \{(\mathbf{0}, \mathbf{0}, s) \in \mathbb{R}^m \times \mathbb{R}^p \times \mathbb{R} | s < p^*\}$, where p^* is the optimal value to the primal problem.

We can show that the sets A and B are convex sets and are disjoint. This means according to the separating hyperplane theorem, there exists a separating hyperplane which separates two disjoint convex sets.





We can define the separating hyperplane as follows:

$$(\mathbf{u}, \mathbf{v}, t) \in A \implies \tilde{\lambda}^T \mathbf{u} + \tilde{\nu}^T \mathbf{v} + \mu t \geq \alpha$$

$$(\mathbf{0}, \mathbf{0}, t) \in B \implies \tilde{\lambda}^T \mathbf{u} + \tilde{\nu}^T \mathbf{v} + \mu t \leq \alpha$$

We can see from the above that $\tilde{\lambda} \geq 0$ and $\mu > 0$. This is because if $(\mathbf{u}, \mathbf{v}, t) \in A$ then $(k\mathbf{u}, \mathbf{v}, kt), k > 1 \in A$, and a negative $\tilde{\lambda}, \mu$ will make the left hand-side of the inequality arbitrarily small, so that it cannot be lower-bounded by α .

The bottom condition means that $\mu t \leq \alpha$ for $t < p^*$ which means that $\mu p^* \leq \alpha$.

For any $\mathbf{x} \in D$

$$\sum_{i=1}^{i=m} \tilde{\lambda}_i g_i(\mathbf{x}) + \nu^T (A\mathbf{x} - b) + \mu f(\mathbf{x}) \geq \alpha \geq \mu p^*$$



There are now two cases: $\mu > 0$ and $\mu = 0$. When $\mu > 0$, we can divide the left and right-hand sides to get

$$L(\mathbf{x}, \tilde{\lambda}/\mu, \tilde{\nu}/\mu) \geq p^*$$

for all $\mathbf{x} \in D$. Defining $\lambda = \tilde{\lambda}/\mu$ and $\nu = \tilde{\nu}/\mu$, we can set $g(\lambda, \nu) = \inf_{\mathbf{x}} L(\mathbf{x}, \lambda, \nu)$. We can see that $g(\lambda, \nu) \geq p^*$.



- ▶ By weak duality we know that $p^* \geq g(\lambda, \nu)$. From the previous slide we have $g(\lambda, \nu) = p^*$.
- ▶ Let us now consider the case $\mu = 0$.
- ▶ Then, for all $\mathbf{x} \in D$, we have $\sum_{i=1}^{i=m} \tilde{\lambda}_i g_i(\mathbf{x}) + \nu^T (A\mathbf{x} - b) \geq 0$.
- ▶ For the point $\tilde{\mathbf{x}}$ that satisfies Slater's condition (which is $g_i(\tilde{\mathbf{x}}) < 0$ and $A\tilde{\mathbf{x}} = b$), we have $\sum_{i=1}^{i=m} \tilde{\lambda}_i g_i(\tilde{\mathbf{x}}) \geq 0$.



- ▶ From $g_i(\tilde{\mathbf{x}}) < 0$ and $\tilde{\lambda}_i \geq 0$, we conclude that $\tilde{\lambda}_i = 0$.
- ▶ From $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$, and $(\tilde{\lambda}, \mu) = 0$ we conclude that $\tilde{\nu} \neq 0$.
- ▶ Then from $\sum_{i=1}^m \tilde{\lambda}_i g_i(\mathbf{x}) + \nu^T (A\mathbf{x} - b) \geq 0$, we have $\tilde{\nu}^T (A\mathbf{x} - b) \geq 0$.
- ▶ We already know that $\tilde{\mathbf{x}}$ is such that $A\tilde{\mathbf{x}} - b = 0$. Since $\tilde{\mathbf{x}} \in D$, we can think of a point $\tilde{\mathbf{x}} + \epsilon \in D$ such that $\tilde{\nu}^T (A(\tilde{\mathbf{x}} + \epsilon) - b) < 0$ unless $\tilde{\nu}^T A = 0$.
- ▶ But if there exists non-zero $\tilde{\nu}$ such that $\tilde{\nu}^T A = 0$, then it means A does not have rank p which is a contradiction. Thus $\tilde{\nu} = 0$, but this contradicts $(\tilde{\lambda}, \tilde{\nu}, \mu) \neq 0$. Therefore μ cannot be zero.

- ▶ In some cases computing the optimum solution for the dual problem is easier than computing the optimal solution for the primal problem.
- ▶ Let α^* denote the optimal solution to the primal problem and β^* denote the optimal solution to the dual problem. From weak duality we know that $\alpha^* \geq \beta^*$.
- ▶ Any feasible solution to the dual problem is a lower bound on the optimal solution to the primal problem.
- ▶ We have $f(\mathbf{x}) - \alpha^* \leq f(\mathbf{x}) - F(\lambda, \nu)$. If we know that $f(\mathbf{x}) - F(\lambda, \nu) < \epsilon$, then we know that \mathbf{x} is at most ϵ away from the true optimal solution. $f(\mathbf{x}) - F(\lambda, \nu)$ is called the duality gap.





We make the following claim: Claim 1: Let $\mathbf{x}^* \in \mathbb{R}^n$ be primal optimal and $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ be dual optimal. Then

- ▶ $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}} L(\mathbf{x}, \lambda^*, \nu^*)$
- ▶ $\lambda_i^* g_i(\mathbf{x}^*) = 0 \forall i \in [m]$

Proof of complementary slackness



We have $f(\mathbf{x}^*) = F(\lambda^*, \nu^*)$ because of strong duality. Then we can write

$$\begin{aligned} f(\mathbf{x}^*) &= F(\lambda^*, \nu^*) \\ &= \inf_{\mathbf{x}} (f(\mathbf{x}) + \sum_{i \in [m]} \lambda_i^* g_i(\mathbf{x}) + \sum_{i \in [p]} \nu_i^* h_i(\mathbf{x})) \\ &\leq f(\mathbf{x}^*) + \sum_{i \in [m]} \lambda_i^* g_i(\mathbf{x}^*) + \sum_{i \in [p]} \nu_i^* h_i(\mathbf{x}^*) \\ &\leq f(\mathbf{x}^*) \end{aligned}$$

Proof of complementary slackness



- ▶ The first line of the preceding set of equations is due to strong duality.
- ▶ The second line states shows how the optimal dual solution is defined.
- ▶ The third line is simply the definition of the Lagrangian dual.
- ▶ The fourth and final line comes about because we know that the primal feasibility of \mathbf{x}^* gives $g_i(\mathbf{x}^*) \leq 0$, $h_i(\mathbf{x}^*) = 0$ and the dual feasibility of (λ^*, ν^*) gives $\lambda_i^* \geq 0$.



- ▶ Our chain of inequalities started with $f(\mathbf{x}^*)$ and ended with $f(\mathbf{x}^*)$. Thus the inequalities are actually equalities. In particular, there is an equality between the third and fourth line which means $\sum_{i \in [m]} \lambda_i^* g_i(\mathbf{x}^*) = 0$. Each term in the summation $\sum_{i \in [m]} \lambda_i^* g_i(\mathbf{x}^*)$ has the same sign which means that the sum can be zero only when each term is zero. Thus we have $\lambda_i^* g_i(\mathbf{x}^*) = 0 \forall i \in [m]$. This is known as complementary slackness.



Given a primal optimization problem, we say that \mathbf{x}^* and $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ respect the Karash-Kuhn-Tucker conditions if:

- ▶ $g_i(\mathbf{x}^*) \leq 0 \forall i \in [m]$.
- ▶ $h_i(\mathbf{x}^*) = 0 \forall i \in [p]$.
- ▶ $\lambda_i^* \geq 0 \forall i \in [m]$.
- ▶ $\lambda_i^* g_i(\mathbf{x}^*) = 0 \forall i \in [m]$.
- ▶ $\nabla f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(\mathbf{x}^*) = 0$.



Theorem: For any optimization problem, if strong duality holds then any primal optimal solution \mathbf{x}^* and dual optimal solution $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ respect the KKT conditions. Conversely if f and g_i are convex for all $i \in [m]$ and h_i are affine for all $i \in [p]$ then the KKT conditions are sufficient for strong duality. Therefore the KKT conditions are both necessary and sufficient for strong duality. We will show the proof in the next slides.


KKT conditions for strong duality



Assume that strong duality holds and \mathbf{x}^* and $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ are primal and dual optimal solutions. Since \mathbf{x}^* is feasible, we see that the first two KKT conditions are true: $g_i(\mathbf{x}^*) \leq 0 \forall i \in [m]$ and $h_i(\mathbf{x}^*) = 0 \forall i \in [p]$. Since (λ^*, ν^*) is dual feasible, we see that the third KKT condition is true: $\lambda_i^* \geq 0$.

The previous claim we proved establishes that for the primal and dual feasible solutions, the fourth KKT condition must hold, i.e. $\lambda_i^* g_i(\mathbf{x}^*) = 0 \forall i \in [m]$. The previous claim also establishes that $\mathbf{x}^* \in \operatorname{argmin}_{\mathbf{x}} L(\mathbf{x}, \lambda^*, \nu^*)$, which means that the gradient of L must vanish at \mathbf{x}^* . Thus the last KKT condition must hold true.

Now we will show that if we assume the KKT conditions and the problem is convex, we have strong duality. The first two conditions

 $g_i(\mathbf{x}^*) \leq 0 \forall i \in [m] \text{ and } h_i(\mathbf{x}^*) = 0 \forall i \in [p]$

The condition $\lambda_i^* \geq 0 \forall i \in [m]$ together with the information that f and the constraints g_i are convex and the constraints h_i are affine enable us to establish that

$L(\mathbf{x}, \lambda^*, \nu^*) = f(\mathbf{x}) + \sum_{i=1}^m \lambda_i^* g_i(\mathbf{x}) + \sum_{i=1}^m \nu_i^* h_i(\mathbf{x})$ is a convex function.

By the last condition we see that the gradient of this convex function vanishes at \mathbf{x}^* which means \mathbf{x}^* is a local and global minimum.





Thus we have

$$\begin{aligned} F(\lambda^*, \nu^*) &= L(\mathbf{x}^*, \lambda^*, \nu^*) \\ &= f(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* g_i(\mathbf{x}^*) + \sum_{i=1}^p \nu_i^* h_i(\mathbf{x}^*) \\ &= f(\mathbf{x}^*) \end{aligned}$$