Mathematical Foundations of Data Science

KKT conditions, duality and multiplier

Yue Xie

Nov 18, 2025

Constrained optimization

$$
\begin{aligned}
\min \quad & f(x) \\
\text{s.t.} \quad & c_i(x) = 0, \ i \in \mathcal{E}, \\
& c_i(x) \geq 0, \ i \in \mathcal{I},
\end{aligned} \tag{1}
$$

- $f : \mathbb{R}^n \to \mathbb{R}$, $c_i : \mathbb{R}^n \to \mathbb{R}$, $i \in \mathcal{E} \cup \mathcal{I}$ are continuously differentiable real-valued functions.
- $\mathcal{I}$ and $\mathcal{E}$ are two finite sets of indices.
- f is the objective function
- $c_i, i \in \mathcal{E}$ are the equality constraints
- $c_i, i \in \mathcal{I}$ are the inequality constraints

Denote

$$\Omega := \{x \mid c_i(x) = 0, \ i \in \mathcal{E}; \ c_i(x) \geq 0, \ i \in \mathcal{I}\}.$$

Local solution and active set

Definition 1

A vector x^* is a (strict) local solution/minimal point of the problem (1) if $x^* \in \Omega$ and there is a neighbourhood $\mathcal{N}$ of x^* such that $f(x) \geq f(x^*)$ ($f(x) > f(x^*)$) for $x \in \mathcal{N} \cap \Omega$ with $x \neq x^*$.

Definition 2 (Active set)

The active set $\mathcal{A}(x)$ at any feasible x consists of the equality constraint indices from $\mathcal{E}$ together with the indices of the inequality constraints i for which $c_i(x) = 0$; that is,

$$\mathcal{A}(x) := \mathcal{E} \cup \{i \in \mathcal{I} \mid c_i(x) = 0\}.$$

At a feasible point x, the inequality constraint $i \in \mathcal{I}$, is said to be *active* if $c_i(x) = 0$ and *inactive* if the strict inequality $c_i(x) > 0$ is satisfied.

Constraint qualification

Definition 3 (LICQ)

Given the point x and the active set $\mathcal{A}(x)$ defined in Definition 2, we say that the linear independence constraint qualification (LICQ) holds if the set of active constraint gradients $\{\nabla c^i(x), i \in \mathcal{A}(x)\}$ is linearly independent.

Karush-Kuhn-Tucker conditions

Define the Lagrangian function for (1):

$$\mathcal{L}(x,\lambda) := f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x). \tag{2}$$

Theorem 4

Suppose that x^* is a local solution of (1), that the functions f and c_i in (1) are continuously differentiable, and the LICQ holds at x^*. Then there is a Lagrange multiplier vector λ^*, with components λ_i^*, $i \in \mathcal{E} \cup \mathcal{I}$, such that the following conditions are satisfied at (x^*, λ^*),

$$\begin{aligned}
\nabla_x \mathcal{L}(x^*, \lambda^*) &= 0, && && \text{(3a)}\\
c_i(x^*) &= 0, && \text{for all } i \in \mathcal{E}, && \text{(3b)}\\
c_i(x^*) &\geq 0, && \text{for all } i \in \mathcal{I}, && \text{(3c)}\\
\lambda_i^* &\geq 0, && \text{for all } i \in \mathcal{I}, && \text{(3d)}\\
\lambda_i^* c_i(x^*) &= 0, && \text{for all } i \in \mathcal{E} \cup \mathcal{I}. && \text{(3e)}
\end{aligned}$$

Karush-Kuhn-Tucker conditions

- We can omit the terms for indices $i \notin \mathcal{A}(x^*)$ from (3a) (since (3e) implies $\lambda_i = 0$, $\forall i \notin \mathcal{A}(x^*)$) and rewrite this conditions as

$$\nabla_x \mathcal{L}(x^*, \lambda^*) = \nabla f(x^*) - \sum_{i \in \mathcal{A}(x^*)} \lambda_i^* \nabla c_i(x^*) = 0.$$

- For a given problem (1) and solution point x^*, there may be many vectors λ^* for which (3) are satisfied. When the LICQ holds, however, the optimal λ^* is unique.

Strict Complementarity

Definition 5 (Strict Complementarity)

Given a local solution x^* of (1) and a vector λ^* satisfying (3), we say that the strict complementarity condition holds if exactly one of λ_i^* and $c_i(x^*)$ is zero for each index $i \in \mathcal{I}$. In other words, we have that $\lambda_i^* > 0$ for each $i \in \mathcal{I} \cap \mathcal{A}(x^*)$.

Satisfaction of the strict complementarity property usually makes it easier for algorithms to determine the active set $\mathcal{A}(x^*)$ and converge rapidly to the solution x^*.

Example

Consider the problem:

$$\begin{aligned} \min_{x} \quad & \left(x_1 - \frac{3}{2}\right)^2 + \left(x_2 - \frac{1}{2}\right)^4 \\ \text{s.t.} \quad & \begin{bmatrix} 1 - x_1 - x_2 \\ 1 - x_1 + x_2 \\ 1 + x_1 - x_2 \\ 1 + x_1 + x_2 \end{bmatrix} \geq 0, \end{aligned} \tag{4}$$

Find the solution and check KKT conditions.
Question. Is strict complementarity satisfied?

Discussion.

Draw feasible region and contour of the problem. It is fairly easy to see that the solution is $x^* = (1, 0)^T$. The first and second constraints in (4) is active at this point. Denoting them by c_1 and c_2 (and the inactive contraints by c_3 and c_4), we have

$$\nabla f(x^*) = \begin{pmatrix} -1 \\ -\frac{1}{2} \end{pmatrix}, \quad \nabla c_1(x^*) = \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \quad \nabla c_2(x^*) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Therefore, the KKT conditions (3a)-(3e) are satisfied when we set

$$\lambda^* = \left(\frac{3}{4}, \frac{1}{4}, 0, 0\right)^T. \tag{5}$$

Duality

- Duality theory shows how we can construct an alternative problem from the functions and data that define the original optimization problem.
- In some cases, the dual problem is easier to solve computationally than the original problem.
- In other cases, the dual can be used to obtain easily a lower bound on the optimal value of the objective for the primal problem.
- The dual has also been used to design algorithms for solving the primal problem.

Problem of interest

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c(x) \geq 0, \end{aligned} \tag{6}$$

where the $c(x)$ is a vector function:

$$c(x) \triangleq (c_1(x), c_2(x), \ldots, c_m(x))^T.$$

f and $-c_i$ are all convex real-valued functions.

Dual Problem

- For (6) the Lagrangian function with Lagrange multiplier vector $\lambda \in \mathbb{R}^m$ is

$$\mathcal{L}(x, \lambda) = f(x) - \lambda^T c(x).$$

- Define the dual objective function $q : \mathbb{R}^n \to \mathbb{R}$ as follows:

$$q(\lambda) \triangleq \inf_x \mathcal{L}(x, \lambda). \tag{7}$$

- In many problems, this infimum is $-\infty$ for some values of λ. We define the domain of q as the set of λ values for which q is finite, that is,

$$\mathcal{D} \triangleq \{\lambda \mid q(\lambda) > -\infty\}. \tag{8}$$

- When f and $-c_i$ are convex functions and $\lambda \geq 0$, the function $\mathcal{L}(\cdot, \lambda)$ is also convex. In this situation, all local minimizers are global minimizers, so computation of $q(\lambda)$ in (7) is more practical.
- The dual problem to (6) is defined as follows:

$$\max_{\lambda \in \mathbb{R}^m} \quad q(\lambda) \qquad \text{s.t.} \qquad \lambda \geq 0. \tag{9}$$

Example

Consider the problem

$$\begin{aligned} \min_{(x_1,x_2)} \quad & x_1^2 + x_2^2 \\ \text{s.t.} \quad & x_1 - 2 \geq 0. \end{aligned} \tag{10}$$

Find out its dual problem.

Discussion.

The Lagrangian is

$$\mathcal{L}(x_1, x_2, \lambda) = x_1^2 + x_2^2 - \lambda_1(x_1 - 2).$$

If we hold λ_1 fixed, this is a convex function of $(x_1, x_2)^T$. Therefore, the infimum with respect to $(x_1, x_2)^T$ is achived when the partial derivatives with respect to x_1 and x_2 are zero, that is,

$$2x_1 - \lambda_1 = 0, \quad x_2 = 0.$$

By substituting these infimal values into $\mathcal{L}(x_1, x_2, \lambda_1)$ we obtain the dual objective (7):

$$q(\lambda_1) = \lambda_1^2/4 + 0 - \lambda_1(\lambda_1/2 - 2) = -\lambda_1^2/4 + 2\lambda_1.$$

Hence, the dual problem (9) is

$$\max_{\lambda_1 \geq 0} \quad -\lambda_1^2/4 + 2\lambda_1,$$

which has the solution $\lambda_1 = 4$.

Property of dual problem

Theorem 6

The function q defined by (7) is concave ($-q$ is convex) and its domain $\mathcal{D}$ is convex.

Theorem 7 (Weak duality)

For any $\bar{x}$ feasible for (6) and any $\bar{\lambda} \geq 0$, we have $q(\bar{\lambda}) \leq f(\bar{x})$.

The optimal value of the dual problem (9) gives a lower bound on the optimal objective value for the primal problem.

Proof.

(Theorem 6) For any λ^0 and λ^1 in $\mathbb{R}^m$, any $x \in \mathbb{R}^n$, and any $\alpha \in [0, 1]$, we have

$$\mathcal{L}(x, (1-\alpha)\lambda^0 + \alpha\lambda^2) = (1-\alpha)\mathcal{L}(x, \lambda^0) + \alpha\mathcal{L}(x, \lambda^1).$$

By taking the infimum of both sides of this expression, using the definition (7), and the results that the infimum of a sum is greater than or equal to the sum of infimums, we obtain

$$q((1-\alpha)\lambda^0 + \alpha\lambda^1) \geq (1-\alpha)q(\lambda^0) + \alpha q(\lambda^1),$$

confirming concavity of q. If both λ^0 and λ^1 belong to $\mathcal{D}$, this inequality implies that $q((1-\alpha)\lambda^0 + \alpha\lambda^1) > -\infty$ also, and therefore $(1-\alpha)\lambda^0 + \alpha\lambda^1 \in \mathcal{D}$, verifying convexity of $\mathcal{D}$. ∎

Proof.

(Theorem 7)

$$q(\bar{\lambda}) = \inf_x f(x) - \bar{\lambda}^T c(x) \le f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) \le f(\bar{x}),$$

where the final inequality follows from $\bar{\lambda} \ge 0$ and $c(\bar{x}) \ge 0$. ∎

KKT conditions and dual

KKT conditions specialized to (6) are as follows:

$$\nabla f(\bar{x}) - \nabla c(\bar{x})\bar{\lambda} = 0, \tag{11a}$$

$$c(\bar{x}) \geq 0, \tag{11b}$$

$$\bar{\lambda} \geq 0, \tag{11c}$$

$$\bar{\lambda}_i c_i(\bar{x}) = 0, i = 1, 2, \ldots, m, \tag{11d}$$

where $\nabla c(x)$ is the $n \times m$ matrix defined by
$\nabla c(x) = [\nabla c_1(x), \nabla c_2(x), \ldots, \nabla c_m(x)]$.

Theorem 8

Suppose that $\bar{x}$ is a solution of (6) and that f and $-c_i$, $i = 1, 2, \ldots, m$ are convex functions on $\mathbb{R}^m$ that are differentiable at $\bar{x}$. Then any $\bar{\lambda}$ for which $(\bar{x}, \bar{\lambda})$ satisfies the KKT conditions (11) is a solution of (9).

Proof.

Suppose that $(\bar{x}, \bar{\lambda})$ satisfies (11). We have from $\bar{\lambda} \geq 0$ that $\mathcal{L}(\cdot, \bar{\lambda})$ is a convex and differentiable function. Hence, for any x, we have

$$\mathcal{L}(x, \bar{\lambda}) \geq \mathcal{L}(\bar{x}, \bar{\lambda}) + \nabla_x \mathcal{L}(\bar{x}, \bar{\lambda})^T (x - \bar{x}) = \mathcal{L}(\bar{x}, \bar{\lambda}),$$

where the last equality follows from (11a). Therefore, we have

$$q(\bar{\lambda}) = \inf_x \mathcal{L}(x, \bar{\lambda}) = \mathcal{L}(\bar{x}, \bar{\lambda}) = f(\bar{x}) - \bar{\lambda}^T c(\bar{x}) = f(\bar{x}),$$

where the last equality follows from (11d). Since from Theorem 7, we have $q(\bar{\lambda}) \leq f(\bar{x})$ for all $\lambda \geq 0$ it follows immediately from $q(\lambda) = f(\bar{x})$ that $\bar{\lambda}$ is a solution of (9). ∎

Strong duality

Definition 9

Denote $p^*(d^*)$ as the optimal value of the primal(dual) problem. If the equality

$$d^* = p^*$$

holds, i.e., the optimal duality gap is zero, then we say that strong duality holds.

Example

Find out the dual problem of the following linear programming problem

$$\min \quad c^T x \qquad \text{s.t.} \qquad Ax - b \geq 0, \tag{12}$$

Solution. The dual objective is

$$q(\lambda) = \inf_x [c^T x - \lambda^T (Ax - b)] = \inf_x [(c - A^T \lambda)^T x + b^T \lambda].$$

If $c - A^T \lambda \neq 0$, the infimum is clearly $-\infty$. When $c - A^T \lambda = 0$, on the other hand, the dual objective is simply $b^T \lambda$. In maximizing q, we can exclude λ for which $c - A^T \lambda \neq 0$ from consideration. Hence, we can write the dual problem (9) as follows:

$$\max_\lambda \quad b^T \lambda \qquad \text{s.t.} \qquad A^T \lambda = c, \ \lambda \geq 0.$$

More general case

Consider the following problem:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{s.t.} \quad & c(x) \geq 0, \quad h(x) = 0. \end{aligned} \tag{13}$$

where $c(x)$ and $h(x)$ are all vector functions:

$$\begin{aligned} c(x) &\triangleq (c_1(x), c_2(x), \ldots, c_m(x))^T, \\ h(x) &\triangleq (c_1(x), c_2(x), \ldots, c_p(x))^T. \end{aligned}$$

f, c_i, h_i are real-valued functions.

The dual problem of (13) is defined as

$$\begin{aligned} \max_{\lambda,\mu} \quad & q(\lambda, \mu) \\ \text{s.t.} \quad & \lambda \in \mathbb{R}^p, \quad \mu \in \mathbb{R}^m_+, \end{aligned} \tag{14}$$

where $q(\lambda, \mu) = \min_x \mathcal{L}(x, \lambda, \mu) = \min_x f(x) - \lambda^T h(x) - \mu^T c(x)$.

Theorem 10

Suppose that f and $-c_i$ are convex and continuously differentiable, h_j is convex affine (linear plus a constant). The following statements hold:

(i) *Any local minimum point of* (13) *us a global minimum point;*

(ii) $-q(\lambda, \mu)$ *with* $\mu \geq 0$ *is convex.*

(iii) *(Weak duality) For any feasible* $\bar{x}$ *for* (13) *and any feasible* $\bar{\lambda}, \bar{\mu} \geq 0$ *for* (14), *we have* $f(\bar{x}) \geq q(\bar{\lambda}, \bar{\mu})$.

(iv) *If **LICQ** holds at* x^* *and* x^* *is a global minimum point of* (13) *with Lagrangian multipliers* $\mu^* \geq 0$ *and* λ^*, *then **strong duality** holds and* (14) *has a global maximum w.r.t.* $\mu \geq 0$ *and* λ *at* μ^* *and* λ^*.

Proof.

Omitted. ∎

Algorithm: Augmented Lagrangian Method (ALM)

First we introduce the equality constrained problem as a special case of (1):

$$\min \quad f(x) \qquad \text{s.t.} \quad c(x) = 0,\ x \in \Omega. \tag{ECP}$$

- $c(x) = (c_1(x), \ldots, c_m(x))^T$, $f : \mathbb{R}^n \to \mathbb{R}$, $c_i : \mathbb{R}^n \to \mathbb{R}$, and Ω is a closed set in $\mathbb{R}^n$.
- f and c_i, $i = 1, \ldots, m$ are smooth.

We can reformulate (ECP) as follows:

$$\min_{x \in \Omega} \left\{ \max_{\lambda} \mathcal{L}(x, \lambda) := f(x) - \lambda^T c(x) \right\} \tag{D}$$

The inner problem is infinite when $c(x) \neq 0$. We could introduce a proximal penalty term to penalize deviation from a previous guess $\bar{\lambda}$:

$$\begin{aligned}
&\min_{x \in \Omega} \left\{ \max_{\lambda} f(x) - \lambda^T c(x) - \frac{1}{2\rho} \|\lambda - \bar{\lambda}\|^2 \right\} \\
=& \min_{x \in \Omega} \left\{ f(x) - (\bar{\lambda} - \rho c(x))^T c(x) - \frac{1}{2\rho} \|\bar{\lambda} - \rho c(x) - \bar{\lambda}\|^2 \right\} \\
=& \min_{x \in \Omega} \left\{ f(x) - \bar{\lambda}^T c(x) + \frac{\rho}{2} \|c(x)\|^2 \right\},
\end{aligned}$$

We denote the *augmented Lagrangian function* as

$$\mathcal{L}_\rho(x,\lambda) := f(x) - \lambda^T c(x) + \frac{\rho}{2}\|c(x)\|^2.$$

It is a summation of the ordinary Lagrangian function and a quadratic penalty term that penalizes violation of the equality constraint $c(x) = 0$.

We denote the *augmented Lagrangian function* as

$$\mathcal{L}_\rho(x, \lambda) := f(x) - \lambda^T c(x) + \frac{\rho}{2}\|c(x)\|^2.$$

It is a summation of the ordinary Lagrangian function and a quadratic penalty term that penalizes violation of the equality constraint $c(x) = 0$. Consider the following algorithm to solve (ECP):

$$\begin{aligned} x^{k+1} &\in \arg\min_{x \in \Omega} \mathcal{L}_{\rho_k}(x, \lambda^k) \\ \lambda^{k+1} &= \lambda^k - \rho_k c(x^{k+1}) \end{aligned} \tag{ALM}$$

Historically, this algorithm was referred to as the *method of multipliers* in the optimization literature. More recently, it has been known as the *augmented Lagrangian method* (ALM).

Implementation of ALM

Now let us discuss the implementation of ALM for (ECP) when $\Omega = \mathbb{R}^n$.

Augmented Lagrangian method for (ECP)

0: Choose $\epsilon_k > 0$, $k = 1, 2, \ldots$, $x^0 \in \mathbb{R}^n$, $\lambda^0 \in \mathbb{R}^m$, $\rho_0 > 0$, $\tau \in (0, 1)$, $\gamma > 1$; Set $k := 0$;

1: Find x^{k+1} such that $\|\nabla_x \mathcal{L}_{\rho_k}(x^{k+1}, \lambda^k)\| \leq \epsilon_k$ (may start the subproblem solver from last iterate x^k);

2: STOP when certain stopping criterion holds and output x^{k+1} as the approximate solution;

3: $\lambda^{k+1} := \lambda^k - \rho_k c(x^{k+1})$;

4: Update ρ_k: if $\|c(x^{k+1})\| > \tau \|c(x^k)\|$, let $\rho_{k+1} = \gamma \rho_k$; otherwise, $\rho_{k+1} = \rho_k$;

5: Set $k := k + 1$ and return to Step 1.

Alternating direction method of multiplier (ADMM)

ADMM is an algorithm based on ALM and efficient in solving problems in traditional machine learning and image processing.

$$\min_{x,y} \quad f(x) + g(y) \qquad \text{s.t.} \quad Ax + By = b, \tag{SP}$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, $g : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ are extended real valued functions. $A \in \mathbb{R}^{p \times n}$, $B \in \mathbb{R}^{p \times m}$ and $b \in \mathbb{R}^p$.

Write the Augmented Lagrangian of (SP) as

$$\mathcal{L}_\rho(x, y, \lambda) = f(x) + g(y) - \lambda^T(Ax + By - b) + \frac{\rho}{2}\|Ax + By - b\|_2^2.$$

Main subproblem in implementing ALM is (suppose that $\rho_k \equiv \rho$)

$$\min_{x,y} \mathcal{L}_\rho(x, y, \lambda^k)$$

but there is coupling between x and y via the penalty term $\|Ax + By - b\|_2^2$. In ADMM, we minimize w.r.t. x and y separately and sequentially:

$$\begin{aligned}
x^{k+1} &:= \arg\min_x \mathcal{L}_\rho(x, y^k, \lambda^k), \\
y^{k+1} &:= \arg\min_y \mathcal{L}_\rho(x^{k+1}, y, \lambda^k), \\
\lambda^{k+1} &:= \lambda_k - \rho(Ax^{k+1} + By^{k+1} - b).
\end{aligned}$$

This approach makes sense when it is much easier to minimize $f(x)+$ (convex quadratic in x and $g(y)+$ (convex quadratic in y) than to minimize $f(x) + g(y)+$ (convex quadratic in (x, y)).

Example

Lasso (ℓ_1 regularized linear regression):

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Cx - d\|^2 + \gamma\|x\|_1, \tag{15}$$

where $C \in \mathbb{R}^{l \times n}$, $d \in \mathbb{R}^l$.

Discussion

For $\alpha > 0$, denote

$$S_\alpha^e(u) := \arg\min_{z \in \mathbb{R}^m} \frac{1}{2}\|z - u\|_2^2 + \alpha\|z\|_1.$$

Then $S_\alpha^e : \mathbb{R}^m \to \mathbb{R}^m$ has closed form:

$$i\text{th component of } S_\alpha^e(u) = S_\alpha(u^i) = \begin{cases} u^i - \alpha, & \text{if } u^i > \alpha, \\ 0, & \text{if } -\alpha \le u^i \le \alpha, \\ u^i + \alpha, & \text{if } u^i < -\alpha. \end{cases}$$

S_α is named soft thresholding operator and S_α^e apply soft thresholding elementwisely. Another formula of $S_\alpha(u^i)$ is

$$S_\alpha(u^i) = (1 - \alpha/|u^i|)_+ u^i.$$

This reflects the shrinkage property of the operator S_α.

The problem (15) can be reformulated as:

$$\min_{x,y\in\mathbb{R}^n} \frac{1}{2}\|Cx-d\|^2+\gamma\|y\|_1 \ \text{ s.t. } \ x=y.$$

Aug Lagr is

$$\mathcal{L}_\rho(x,y,\lambda)=\frac{1}{2}\|Cx-d\|^2+\gamma\|y\|_1-\lambda^T(x-y)+\frac{\rho}{2}\|x-y\|_2^2.$$

ADMM steps are:

$$\begin{aligned}
x_{k+1} &:= \arg\min_x \frac{1}{2}\|Cx-d\|^2-(\lambda_k)^Tx+\frac{\rho}{2}\|x-y_k\|^2\\
&= (C^TC+\rho I)^{-1}(C^Td+\lambda_k+\rho y_k)\\
y_{k+1} &:= \arg\min_y \frac{\rho}{2}\|y-x_{k+1}\|^2+\lambda_k^Ty+\gamma\|y\|_1\\
&= S^e_{\gamma/\rho}(x_{k+1}-\lambda_k/\rho)\\
\lambda_{k+1} &:= \lambda_k-\rho(x_{k+1}-y_{k+1})
\end{aligned}$$

We have closed-form solution for all updates! This justifies the purpose of using ADMM.