Reference:

THEOREM (TAYLOR’S THEOREM)

Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $p \in \mathbb{R}^n$. Then for some $t \in (0, 1)$ we have that

$$f(x + p) = f(x) + \nabla f(x + tp)'p$$

Moreover, if $f$ is twice continuously differentiable, for some $t \in (0, 1)$ we have that

$$f(x + p) = f(x) + \nabla f(x)'p + \frac{1}{2}p'\nabla^2 f(x + tp)p$$
THEOREM (FIRST-ORDER NECESSARY CONDITIONS)

Let \( f : \mathbb{R}^n \to \mathbb{R} \) be continuously differentiable and \( x^* \) a local optimizer. Then

\[
\nabla f(x^*) = 0
\]

Proof:

• Assume by contradiction that \( p = -\nabla f(x^*) \neq 0 \). Let \( g(t) = p' \nabla f(x^* + tp) \).
  Then \( g(0) = p' \nabla f(x^*) = -\|\nabla f(x^*)\|^2 < 0 \)

• \( \nabla f \) is continuous around \( x^* \), so \( g \) is also continuous wrt \( t \) in \( t = 0 \), and therefore
  \( \exists T > 0 \) such that \( g(t) < 0 \) for all \( t \in [0, T] \)

• For any \( \bar{t} \in (0, T] \) by Taylor’s theorem we have that for some \( t \in (0, \bar{t}) \)

\[
f(x^* + \bar{t}p) = f(x^*) + \bar{t}p' \nabla f(x^* + tp) = f(x^*) + g(t) \bar{t} < f(x^*), \ \forall \bar{t} \in (0, T]\)

• Then \( x^* \) is not a local minimizer, which is a contradiction. \( \square \)
THEOREM (SECOND-ORDER NECESSARY CONDITIONS)

Let the Hessian matrix function $\nabla^2 f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ exist and be continuous in an open neighborhood of a local optimizer $x^*$. Then

$$\nabla f(x^*) = 0, \quad \nabla f^2(x^*) \succeq 0$$

Proof:

• Assume by contradiction that $\nabla^2 f(x^*) \not\succeq 0$. Then there exist $p$ such that $p' \nabla^2 f(x^*) p < 0$.

• Since $\nabla^2 f(x)$ is continuous around $x^*$, $\exists T > 0$ such that $p' \nabla^2 f(x^* + tp) p < 0$ for all $t \in [0, T]$.

• By doing a Taylor expansion around $x^*$, $\forall \bar{t} \in (0, T]$ there exists $t \in (0, \bar{t})$ such that

$$f(x^* + \bar{t}p) = f(x^*) + \bar{t}p' \nabla f(x^*) + \frac{1}{2} \bar{t}^2 p' \nabla^2 f(x^* + tp)p < f(x^*)$$

• Then $x^*$ is not a local minimizer, which is a contradiction.
Optimality Conditions

Theorem (Second-order Sufficient Conditions)

Let $\nabla^2 f : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ exist and be continuous in an open neighborhood of $x^*$. Let $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) \succ 0$. Then $x^*$ is a strict local minimizer of $f$.

Proof:

• Since the Hessian function $\nabla^2 f(x)$ is continuous at $x^*$ and $\nabla^2 f(x^*) \succ 0$, $\nabla^2 f(x) \succ 0$ for all $x$ in an open ball $B(x^*, r)$ for some scalar $r > 0$

• For any $p$ such that $\|p\|_2 < r$ we have that $x^* + p \in B(x^*, r)$ and hence

$$f(x^* + p) = f(x^*) + p' \nabla f(x^*) + \frac{1}{2} p' \nabla^2 f(x^* + tp)p = f(x^*) + \frac{1}{2} p' \nabla^2 f(x^* + tp)p$$

for some $t \in (0, 1)$.

• Since $x^* + tp \in B(x^*, r)$, $p' \nabla^2 f(x^* + tp)p > 0$, and therefore

$$f(x^* + p) > f(x^*), \forall p \in B(0, r).$$

\[1\] For a positive scalar $r > 0$, the Euclidean ball $B(x_0, r)$ is the set \( \{ x : \|x - x_0\|_2 \leq r \} \).
• Consider the constrained optimization problem

\[
\begin{align*}
\min_x f(x) \\
\text{s.t.} & \quad g_i(x) \leq 0, \ i \in I \\
& \quad g_j(x) = 0, \ j \in E
\end{align*}
\]

with \( I \cup E = \{1, \ldots, m\} \).

• A vector \( x \) is **feasible** if \( g_i(x) \leq 0, \ \forall i \in I \), and \( g_j(x) = 0, \ \forall j \in E \)

• We say that the inequality constraint \( i \in I \) is **active** if \( g_i(x) = 0 \), **inactive** if \( g_i(x) < 0 \) (equality constraints \( g_j(x), j \in E \), are always active).
**Optimality Conditions - Constrained Case**

- The **active set** $A(x)$ at any feasible vector $x$ is the set of indexes
  $$A(x) = \{ i \in I : g_i(x) = 0 \} \cup E$$

- We say that the **linear independence constraint qualification** (LICQ) condition holds at $x$ if the vectors $\{ \nabla g_i(x) \}_{i \in A(x)}$ are linearly independent.

- The set $\mathcal{F}(x)$ of **linearized feasible directions** at a feasible $x$ is the cone
  $$\mathcal{F}(x) = \{ d : d' \nabla g_i(x) = 0, \forall i \in E, d' \nabla g_i(x) \leq 0, \forall i \in A(x), i \not\in E \}$$

Note that $g_i(x + d) \approx g_i(x) + \nabla g_i(x)'d$ for $d \to 0, \forall i \in A(x)$

- Linear case example:
  $$\begin{cases} A_1 x \leq b_1 \\ A_2 x \leq b_2 \end{cases} \quad \rightarrow \quad \begin{cases} A_1 d \leq 0 \\ A_2 d \leq 0 \end{cases}$$
**THEOREM**

If $x^*$ is a local minimum and the LICQ condition is satisfied then

$$\nabla f(x^*)'d \geq 0, \forall d \in F(x^*)$$

- Define the **Lagrangian function**

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x)$$

where $\lambda \in \mathbb{R}^m$ are the **Lagrange multipliers**, $I \cup E = \{1, \ldots, m\}$
**THEOREM (FIRST-ORDER NECESSARY CONDITIONS)**

Let \( f \) and \( g_i, i = 1, \ldots, m \), be continuously differentiable and \( x^* \) a local optimizer. Let the LICQ condition hold at \( x^* \). Then

\[ \exists \lambda^* \in \mathbb{R}^m \text{ such that} \]

\[ \nabla_x \mathcal{L}(x^*, \lambda^*) = 0 \]

\[ g_i(x^*) \leq 0 \quad \forall i \in I \]

\[ g_i(x^*) = 0 \quad \forall i \in E \]

\[ \lambda_i^* \geq 0 \quad \forall i \in I \]

\[ \lambda_i^* g_i(x^*) = 0 \quad \forall i = 1, \ldots, m \]

- \( \lambda_i^* g_i(x^*) = 0 \) is a **complementary slackness** condition
- **strict complementarity** holds if \( \lambda_i^* > 0 \) for all \( i \in A(x^*) \)
- \( \lambda^* \) is unique if the LICQ condition holds
\[ -\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*), \quad \lambda_i^* \geq 0, \quad E = \emptyset \]

\[ f(x^* + \epsilon d) \approx f(x^*) + \epsilon \nabla f(x^*)' d \]

- \( f \) decreases when \(-\nabla f(x^*)' d > 0\)

- if \(-\nabla f(x^*)' d = \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*)' d\) were positive then \(\nabla g_i(x^*)' d > 0\) for some \(i \in A(x^*)\) such that \(\lambda_i^* > 0\).

Hence \(f\) can only decrease at \(x^*\) if some active constraint \(g_i\) is violated, as
\[ g_i(x^* + \epsilon d) \approx g_i(x^*) + \epsilon \nabla g_i(x^*)' d = \epsilon \nabla g_i(x^*)' d > 0, \quad \epsilon > 0 \]

- Vice versa, if \(-\nabla f(x^*)\) does not belong to the convex cone one can move in a direction \(d\) such that \(d' \nabla f(x^*) < 0\) (that is, decrease \(f\)) while keeping \(g_i(x) \leq 0\).
KKT conditions for equality-constrained QP

- **Quadratic programming** problem subject to **equality** constraints:

\[
\begin{align*}
\text{min } & \quad \frac{1}{2} x' Q x + c' x \\
\text{s.t. } & \quad A x = b
\end{align*}
\]

\[Q = Q' \succ 0, \ A \text{ full row rank}\]

- Lagrangian function: \( \mathcal{L}(x, \lambda) = \frac{1}{2} x' Q x + c' x + \lambda' (A x - b) \)

- KKT conditions:

\[
\begin{align*}
Q x + c + A' \lambda &= 0 \\
A x &= b
\end{align*}
\]

\[\Rightarrow \quad x = -Q^{-1} (c + A' \lambda) \quad A Q^{-1} A' \lambda = -(b + A Q^{-1} c)\]

and therefore

\[
\begin{align*}
\lambda^* &= -(AQ^{-1} A')^{-1}(b + AQ^{-1} c) \\
x^* &= -Q^{-1} (c - A'(AQ^{-1} A')^{-1}(b + AQ^{-1} c))
\end{align*}
\]

- In this case, the KKT conditions are also **sufficient** for optimality (this is a convex optimization problem, see later ...)

©2023 A. Bemporad - Numerical Optimization
KKT CONDITIONS FOR QP

- **Quadratic programming** problem

  \[
  \begin{align*}
  \text{min} & \quad \frac{1}{2} x' Q x + c' x \\
  \text{s.t.} & \quad Ax \leq b \\
  & \quad Ex = f
  \end{align*}
  \]

- Lagrangian function: \( \mathcal{L}(x, \lambda, \nu) = \frac{1}{2} x' Q x + c' x + \lambda' (Ax - b) + \nu' (Ex - f) \)

- KKT conditions:

  \[
  \begin{align*}
  Q x + c + A' \lambda + E' \nu &= 0 \\
  Ex &= f \\
  Ax &\leq b \\
  \lambda &\geq 0 \\
  \lambda' (Ax - b) &= 0
  \end{align*}
  \]

  where we replaced \( \lambda_i(A_i x - b_i) = 0, \forall i \), with \( \sum_i \lambda_i(A_i x - b_i) = 0 \), having imposed \( \lambda_i \geq 0, A_i x \leq b_i, \forall i \)
Let $x^*, \lambda^*$ satisfy the KKT conditions. The critical cone $C(x^*, \lambda^*)$ is defined as

$$C(x^*, \lambda^*) = \left\{ \begin{array}{l} \nabla g_i(x^*)'w = 0, \forall i \in E \\ \nabla g_i(x^*)'w = 0, \forall i \in A(x^*) \cap I \text{ with } \lambda_i^* > 0 \\ \nabla g_i(x^*)'w \leq 0, \forall i \in A(x^*) \cap I \text{ with } \lambda_i^* = 0 \end{array} \right\}$$

The critical cone $C(x^*, \lambda^*)$ contains directions in $F(x^*)$ for which it is not clear from gradient information only whether $f$ will increase or decrease, as from the KKT conditions we have

$$w'\nabla f(x^*) = \sum_{i=1}^{m} \lambda_i^* w' \nabla g_i(x^*) = 0, \forall w \in C(x^*, \lambda^*)$$
THEOREM (2ND-ORDER NECESSARY CONDITIONS)
Assume $f$, $g$ be twice continuously differentiable. Let $x^*$ be a local minimum and the LICQ condition satisfied and $\lambda^*$ such that the KKT conditions are satisfied. Then

$$w' \nabla_{xx} \mathcal{L}(x^*, \lambda^*) w \geq 0, \forall w \in \mathcal{C}(x^*, \lambda^*)$$

THEOREM (2ND-ORDER SUFFICIENT CONDITIONS)
Assume $f$, $g$ be twice continuously differentiable. Let $x^*$, $\lambda^*$ satisfy the KKT conditions and assume that

$$w' \nabla_{xx} \mathcal{L}(x^*, \lambda^*) w > 0, \forall w \in \mathcal{C}(x^*, \lambda^*), w \neq 0$$

Then $x^*$ is a strict local minimum.
• **Question**: if we slightly perturb a constraint $g_i$ how much $f(x^*)$ will change?

• The Lagrange multipliers $\lambda^*$ answer such a sensitivity analysis question

• If $g_i(x^*) < 0 (\Rightarrow \lambda^*_i = 0)$, perturbing $g_i(x) \leq 0$ to $g_i(x) \leq -\epsilon$ does not change the solution, $\forall \epsilon < -g_i(x^*)$, as the same $x^*$, $\lambda^*$ satisfy the KKT

• Let us change one of the active constraints $g_i(x) \leq 0$ to $g_i(x) \leq -\epsilon$, $i \in A(x^*)$

• Let $x^*(\epsilon)$ be the perturbed optimal solution and assume $|\epsilon|$ small enough so that $A(x^*(\epsilon)) = A(x^*)$
• By taking the Taylor expansion of $g_j(x^*(\epsilon))$ around $\epsilon = 0$ we get

$$g_j(x^*(\epsilon)) - g_j(x^*) \approx \nabla g_j(x^*)'(x^*(\epsilon) - x^*), \quad j = 1, \ldots, m$$

• Since we assumed $A(x^*(\epsilon)) = A(x^*)$, then $g_i(x^*(\epsilon)) = -\epsilon$ and $g_j(x^*(\epsilon)) = 0$, $\forall j \in A(x^*) \setminus \{i\}$, in addition to $g_j(x^*) = 0$, $\forall j \in A(x^*)$

• By expanding $f(x^*(\epsilon))$ around $\epsilon = 0$ and using the KKT conditions

$$f(x^*(\epsilon)) - f(x^*) \approx \nabla f(x^*)'(x^*(\epsilon) - x^*) = \sum_{j \in A(x^*)} -\lambda^*_j \nabla g_j(x^*)'(x^*(\epsilon) - x^*)$$

$$= \sum_{j \in A(x^*)} -\lambda^*_j (g_j(x^*(\epsilon)) - g_j(x^*)) = \epsilon \lambda^*_i$$

• For $\epsilon \to 0$ we get

$$\frac{df(x^*(\epsilon))}{d\epsilon} = \lambda_i^*$$
Sensitivity Analysis

**Definition**
Let $i \in \mathcal{A}(x^*)$. An inequality constraint $g_i$ is *strongly active* if $\lambda_i^* > 0$, *weakly active* if $\lambda_i^* = 0$

- If a constraint is weakly active, modifying it slightly does not change the optimal value since 
  
  \[
  \frac{df(x^*(\epsilon))}{d\epsilon} = 0
  \]

- Let us scale the constraints to $\beta_i g_i(x) \leq 0$, $\beta_i > 0$. The KKT conditions are satisfied for $x^*$ and $\frac{\lambda_i^*}{\beta_i}$

- For the consistent perturbation of the constraint $\beta_i g_i(x) \leq -\beta_i \epsilon$ we get the same optimizer $x^*(\epsilon)$, and moreover the sensitivity at the solution is
  
  \[
  \frac{\lambda_i^*}{\beta_i} = \frac{df(x^*(\epsilon))}{d(\beta_i \epsilon)} = \frac{1}{\beta_i} \frac{df(x^*(\epsilon))}{d\epsilon} \quad \Rightarrow \quad \frac{df(x^*(\epsilon))}{d\epsilon} = \frac{\lambda_i^*}{\beta_i}
  \]
• Consider again the optimization problem

\[
\begin{align*}
\min_x & \quad f(x) \\
\text{s.t.} & \quad g_i(x) \leq 0, \ i \in I \\
& \quad g_j(x) = 0, \ j \in E
\end{align*}
\]

\[
I \cup E = \{1, \ldots, m\}
\]

• Define the dual function \( q : \mathbb{R}^m \rightarrow \mathbb{R} \cup \{-\infty\} \)

\[
q(\lambda) = \inf_x \mathcal{L}(x, \lambda) = \inf_x \left\{ f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) \right\}
\]

• The domain \( \mathcal{D} \) of \( q \) is the set of all \( \lambda \) for which \( q(\lambda) > -\infty \)

• A vector \( \lambda \in \mathcal{D} \) is dual feasible if \( \lambda_i \geq 0, \ \forall i \in I \)

• A vector is \( x \in \mathbb{R}^n \) primal feasible if \( g_i(x) \leq 0, \ \forall i \in I \) and \( g_j(x) = 0, \ \forall j \in E \)
**THEOREM (WEAK DUALITY)**

For any given **primal feasible** \( x \) and **dual feasible** \( \lambda \)

\[
q(\lambda) \leq f(x)
\]

In particular \( q(\lambda) \leq f(x^*) \).

**Proof:**

- Since \( x \) and \( \lambda \) are feasible, \( \lambda_i g_i(x) \leq 0, \forall i \in I \) and \( \lambda_j g_j(x) = 0, \forall j \in E \)

- Therefore

\[
f(x) \geq f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) = \mathcal{L}(x, \lambda) \geq \inf_x \mathcal{L}(x, \lambda) = q(\lambda)
\]

- Since the above relation holds for all feasible \( x \), in particular it holds for \( x^* \)

\[
f(x^*) \geq q(\lambda), \ \forall \lambda \text{ such that } \lambda_i \geq 0, \ i \in I
\]
The dual function $q(\lambda)$ is concave and its domain $\mathcal{D}$ is convex.

Proof:

- Take any $\lambda^1, \lambda^2 \in \mathcal{D}$, and $\alpha \in [0, 1]$. We want to verify that $\alpha \lambda^1 + (1 - \alpha) \lambda^2 \in \mathcal{D}$ and that Jensen’s inequality holds:

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

$$= \inf_x \left\{ f(x) + \sum_{i=1}^{m} (\alpha \lambda^1_i + (1 - \alpha) \lambda^2_i) g_i(x) \right\}$$

$$= \inf_x \left\{ (\alpha + 1 - \alpha) f(x) + \alpha \sum_{i=1}^{m} \lambda^1_i g_i(x) + (1 - \alpha) \sum_{i=1}^{m} \lambda^2_i g_i(x) \right\}$$

$$= \inf_x \left\{ \alpha (f(x) + \sum_{i=1}^{m} \lambda^1_i g_i(x)) + (1 - \alpha) (f(x) + \sum_{i=1}^{m} \lambda^2_i g_i(x)) \right\}$$

$$\geq \inf_{x_1} \left\{ \alpha (f(x_1) + \sum_{i=1}^{m} \lambda^1_i g_i(x_1)) \right\} + \inf_{x_2} \left\{ (1 - \alpha) (f(x_2) + \sum_{i=1}^{m} \lambda^2_i g_i(x_2)) \right\}$$
Finally, we get

\[ q(\alpha \lambda^1 + (1 - \alpha) \lambda^2) \geq \alpha q(\lambda^1) + (1 - \alpha) q(\lambda^2) > -\infty \]

which proves that \( q \) is concave and that \( \alpha \lambda^1 + (1 - \alpha) \lambda^2 \in D \)

Recall that the minimum of a finite number of affine functions is concave. \( q(\lambda) \) is the minimum of infinitely many affine functions (one for each \( x \)).
• We define dual problem of a given optimization problem the new problem

\[
\max_{\lambda} \quad q(\lambda) \\
\text{s.t.} \quad \lambda_i \geq 0, \forall i \in I \\
\lambda \in \mathcal{D}
\]

• The dual problem is always a convex programming problem, even if the primal problem is not convex

• Since \( f(x^*) \geq q(\lambda) \) for all dual feasible \( \lambda \), we also have that the optimum of the dual problem satisfies the weak duality condition

\[
q(\lambda^*) \leq f(x^*)
\]

• Strong duality holds when \( q(\lambda^*) = f(x^*) \)

• The difference \( f(x^*) - q(\lambda^*) \) is called duality gap
Let $x^*(\lambda) = \arg\min_x \mathcal{L}(x, \lambda)$. For all $\lambda \geq 0$, the gradient

$$\nabla_\lambda q(\lambda) = g(x^*(\lambda))$$

Proof:

$$\nabla_\lambda q(\lambda) = \nabla_\lambda (\inf_x \mathcal{L}(x, \lambda)) = \nabla_\lambda \mathcal{L}(x^*(\lambda), \lambda)$$

$$= \nabla_\lambda x^*(\lambda) \begin{pmatrix} \frac{\partial \mathcal{L}(x^*(\lambda), \lambda)}{\partial x} \\ \frac{\partial \mathcal{L}(x^*(\lambda), \lambda)}{\partial \lambda} \end{pmatrix} = 0 \text{ by optimality of } x^*(\lambda)$$

The first-order Taylor expansion of the dual function around $\lambda_0$ is

$$q(\lambda) \approx q(\lambda_0) + g(x^*(\lambda_0))' \lambda$$

Proof:

$$q(\lambda) \approx q(\lambda_0) + \nabla_\lambda q(\lambda_0)'(\lambda - \lambda_0) = q(\lambda_0) + g(x^*(\lambda_0))'(\lambda - \lambda_0)$$

$$= \inf_x \mathcal{L}(x, \lambda_0) + g(x^*(\lambda_0))'(\lambda - \lambda_0) = f(x^*(\lambda_0)) + g(x^*(\lambda_0))'\lambda_0$$

$$+ g(x^*(\lambda_0))'(\lambda - \lambda_0) = f(x^*(\lambda_0)) + g(x^*(\lambda_0))'\lambda$$
• Consider the **convex programming** problem

\[
\begin{align*}
\min_x & \quad f(x) \\
\text{s.t.} & \quad g_i(x) \leq 0, \ i \in I & I \cup E = \{1, \ldots, m\} \\
& \quad A_j x = b_j, \ j \in E
\end{align*}
\]

where are \( f, g_i \) are convex functions.

• We say that **Slater’s constraint qualification** is verified if the problem is strictly feasible:

\[
\exists x : g_i(x) < 0, \ \forall i \in I, \ A_j x = b_j, \ \forall j \in E
\]

• **Strong duality** always holds if Slater’s constraint qualification is satisfied

• Other types of constraint qualifications exist
**Theorem**

Let $x^*$ be the solution of a convex programming problem and $f, g_i$ differentiable at $x^*$. Any $\lambda^*$ satisfying the KKT conditions with $x^*$ solves the dual problem.

**Proof:**

- Assume $x^*, \lambda^*$ satisfy the KKT conditions and consider

$$
\mathcal{L}(x, \lambda^*) = f(x) + \sum_{i \in I} \lambda_i^* g_i(x) + \sum_{j \in E} \lambda_j^* (A_j x - b_j)
$$

- $\mathcal{L}(x, \lambda^*)$ is differentiable w.r.t. $x$ at $x^*$, and is also a convex function of $x$, as $\lambda_i^* \geq 0$ for all $i \in I$

- By convexity of $\mathcal{L}(x, \lambda^*)$ we obtain

$$
= 0 \text{ because of KKT}
$$

$$
\mathcal{L}(x, \lambda^*) \geq \mathcal{L}(x^*, \lambda^*) + \nabla_x \mathcal{L}(x^*, \lambda^*)' (x - x^*) = \mathcal{L}(x^*, \lambda^*)
$$
• Since $\mathcal{L}(x, \lambda^*) \geq \mathcal{L}(x^*, \lambda^*)$ for all $x$ we get
\[
q(\lambda^*) = \inf_x \mathcal{L}(x, \lambda^*) = \mathcal{L}(x^*, \lambda^*)
\]
\[
= f(x^*) + \sum_{i \in I} \lambda_i^* g_i(x^*) + \sum_{j \in E} \lambda_j^* (A_j x^* - b_j) = f(x^*)
\]

(\text{complementarity})

(\text{feasibility})

• Since $q(\lambda) \leq f(x^*)$ for all dual feasible $\lambda$, it follows that

\[
q(\lambda) \leq q(\lambda^*)
\]

• As $\lambda^*$ is dual feasible, it is therefore an optimizer of the dual problem.

□

• Note that we have also proved that the duality gap is zero, as $q(\lambda^*) = f(x^*)$

• In general, for $x_\lambda \in \arg \inf_x \mathcal{L}(x, \lambda)$ the duality gap is

\[
f(x_\lambda) - q(\lambda) = \sum_{i \in I} \lambda_i g_i(x_\lambda) + \sum_{j \in E} \lambda_j (A_j x_\lambda - b_j)
\]
Wolfe’s dual problem is defined as follows:

\[
\max_{x, \lambda} \quad \mathcal{L}(x, \lambda) \\
\text{s.t.} \quad \nabla_x \mathcal{L}(x, \lambda) = 0 \\
\lambda_i \geq 0, \quad \forall i \in I
\]

**Theorem**

Consider a convex programming problem with \( f, g_i \) differentiable on \( \mathbb{R}^n \). Let \( x^*, \lambda^* \) satisfy the KKT conditions and LICQ hold. Then \( x^*, \lambda^* \) is an optimizer of Wolfe’s dual problem.
Proof:

- Since \((x^*, \lambda^*)\) satisfies the KKT conditions it is a feasible point of Wolfe’s dual problem, and moreover \(\mathcal{L}(x^*, \lambda^*) = f(x^*)\)
- For any \((x, \lambda)\) satisfying \(\nabla_x \mathcal{L}(x, \lambda) = 0, \lambda_i \geq 0, \forall i \in I\), we get

\[
\mathcal{L}(x^*, \lambda^*) = f(x^*) \geq f(x^*) + \sum_{i \in I} \lambda_i g_i(x^*) + \sum_{j \in E} \lambda_j \left( A_j x^* - b_j \right) \leq 0
\]

\[
= \mathcal{L}(x^*, \lambda) \geq \mathcal{L}(x, \lambda) + \nabla_x \mathcal{L}(x, \lambda)'(x^* - x) = 0
\]

\[
= \mathcal{L}(x, \lambda)
\]

- Hence \(\mathcal{L}(x^*, \lambda^*) = f(x^*)\) is the maximum achievable value of \(\mathcal{L}(x, \lambda)\) under the constraints \(\nabla_x \mathcal{L}(x, \lambda) = 0, \lambda_i \geq 0, \forall i \in I\). \(\square\)
• Consider the linear program

\[
\begin{align*}
\min_x & \quad c'x \\
\text{s.t.} & \quad Ax \leq b
\end{align*}
\]

• The dual function is

\[
q(\lambda) = \inf_x \{c'x + \lambda'(Ax - b)\} = \inf_x \{(c + A'\lambda)'x - b'\lambda\}
\]

• \(q(\lambda) > -\infty\) only when \(c + A'\lambda = 0\), and \(q(\lambda) = -b'\lambda\)

• The dual problem is therefore

\[
\begin{align*}
\max_\lambda & \quad -b'\lambda \\
\text{s.t.} & \quad A'\lambda = -c \\
& \quad \lambda \geq 0
\end{align*}
\]

• It is easy to prove that the dual of the dual LP is the original LP (\(\min_{x,s} c'x\) s.t. \(Ax + s = b, s \geq 0\)). The original \(x =\) dual vector of constraint \(-A'\lambda + c = 0\), and \(s =\) dual vector of constraint \(\lambda \geq 0\).
THEOREM (THEOREM OF ALTERNATIVES)
For given \( A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m \), exactly one of the following two alternatives is true:

1. there exists \( x \) such that \( Ax \leq b \)
2. there exists \( y \) such that \( y \geq 0, A'y = 0, b'y < 0 \)

LEMMA (FARKAS' LEMMA)
For a given matrix \( A \) and vector \( b \), exactly one of the following two alternatives is true:

1. there exists \( x \) such that \( Ax = b, x \geq 0 \)
2. there exists \( y \) such that \( A'y \geq 0, b'y < 0 \)
Farkas’ lemma has the following geometric interpretation. Let $A_i$ be the $i$th column of $A, i = 1, \ldots, n$, $A = [A_1 A_2 \ldots A_n]$

- **1st alternative:**
  \[
  b = \sum_{i=1}^{n} x_i A_i, \quad x_i \geq 0, \quad i = 1, \ldots, n
  \]

  $b$ is in the convex cone generated by the columns of $A$

- **2nd alternative:**
  \[
  y' A_i \geq 0, \quad i = 1, \ldots, n
  \]
  \[
  y' b < 0
  \]

  Vector $b$ cannot be in the convex cone generated by the columns of $A$
THEOREM (STRONG LP DUALITY)

1. If either the primal or the dual LP has a finite solution, so does the other and $c'x^* = -b'\lambda^*$ (strong duality)

2. If one of the two is unbounded the other is infeasible

- To see that infeasibility of dual LP implies unboundedness of a feasible primal LP, apply Farkas’ Lemma with matrices $-A'$, $c$

  
  $-A'\lambda = c, \lambda \geq 0$ infeasible  \[\implies\]  $\exists d \in \mathbb{R}^n : -Ad \geq 0, c'd < 0$

- Take a feasible $x_0 \in \mathbb{R}^n$. Then $A(x_0 + \sigma d) = Ax_0 + \sigma Ad \leq b, \forall \sigma \geq 0$, and $c'(x_0 + \sigma d) = c'x_0 - \sigma |c'd|$

- As $\sigma$ can be arbitrarily large, the infimum of the primal LP is $-\infty$. 
• Consider the linear program

\[
\begin{align*}
\min_x & \quad c'x \\
\text{s.t.} & \quad Ax \geq b \\
& \quad x \geq 0
\end{align*}
\]

• The dual function is

\[
q(\lambda, \nu) = \inf_x \{c'x + \lambda'(b - Ax) - \nu'x\} = \inf_x \{(c - A'\lambda - \nu)'x + b'\lambda\} = b'\lambda
\]

for \(c - A'\lambda - \nu = 0, \nu \geq 0\), or equivalently \(A'\lambda \leq c\)

• The dual problem is therefore

\[
\begin{align*}
\max_\lambda & \quad b'\lambda \\
\text{s.t.} & \quad A'\lambda \leq c \\
& \quad \lambda \geq 0
\end{align*}
\]

• At optimality \(c'x^* = b'\lambda^*\)
A linear complementarity problem (LCP) is a feasibility problem of the form 
(Cottle, Pang, Stone, 2009)

\[
\begin{align*}
    w &= Mz + q \\
    w'z &= 0 \\
    w, z &\geq 0
\end{align*}
\]

By introducing the vector \( s \) of slack variables, \( s = Ax - b \geq 0 \), the KKT conditions for the following LP are

\[
\begin{align*}
    \min_x & \quad c'x \\
    \text{s.t.} & \quad Ax \geq b \\
    & \quad x \geq 0
\end{align*}
\]

\[
\begin{align*}
    c - A'\lambda - \nu &= 0 \\
    Ax - b - s &= 0 \\
    x, \lambda, \nu, s &\geq 0 \\
    x'\nu = \lambda's &= 0
\end{align*}
\]

Therefore, the original LP can be solved by solving the LCP

\[
\begin{bmatrix}
    \nu \\
    s
\end{bmatrix}
= \begin{bmatrix}
    0 & -A' \\
    A & 0
\end{bmatrix}
\begin{bmatrix}
    x \\
    \lambda
\end{bmatrix}
+ \begin{bmatrix}
    c \\
    -b
\end{bmatrix},
\begin{bmatrix}
    x \\
    \lambda
\end{bmatrix},
\begin{bmatrix}
    \nu \\
    s
\end{bmatrix} \geq 0,
\quad
\begin{bmatrix}
    x'\nu = \lambda's = 0
\end{bmatrix}
\]

\[\iff x'\nu + \lambda's = w'z = 0\]
• Consider the quadratic program

\[
\begin{align*}
\min_x & \quad \frac{1}{2} x' Q x + c' x \\
\text{s.t.} & \quad A x \leq b
\end{align*}
\]

\[Q = Q' \succ 0\]

• The dual function is

\[
q(\lambda) = \inf_x \left\{ \frac{1}{2} x' Q x + c' x + \lambda' (A x - b) \right\}
\]

• Since \(Q \succ 0\) the infimum is achieved when 

\[
0 = \nabla_x \mathcal{L}(x, \lambda) = Q x + c + A' \lambda,
\]

i.e., for \(x, \lambda = -Q^{-1} (c + A' \lambda)\).

• By substitution, Lagrange’s dual QP problem is therefore

\[
\max_{\lambda \geq 0} - \left( \frac{1}{2} \lambda' (A Q^{-1} A') \lambda + (b + A Q^{-1} c)' \lambda + \frac{1}{2} c' Q^{-1} c \right)
\]
Let \( Q \succ 0 \) and consider the dual QP problem

\[
\min_{\lambda} \quad \frac{1}{2} \lambda' (AQ^{-1}A') \lambda + (b + AQ^{-1}c)' \lambda \\
\text{s.t.} \quad \lambda \geq 0
\]

The KKT conditions for the dual QP are the LCP problem

\[
H \lambda + d = s \\
s' \lambda = 0 \\
s, \lambda \geq 0
\]

where \( H = AQ^{-1}A' \) is the dual Hessian and \( d = b + AQ^{-1}c \)

We can therefore solve the QP problem as an LCP to get the dual solution \( \lambda^* \) and then reconstruct the primal solution \( x^* = -Q^{-1}(c + A' \lambda^*) \)
LCP AND DUAL QP

- Vice versa, let $M = M' \succ 0$, $M \in \mathbb{R}^{n \times n}$, and consider the LCP

  \[
  x = My + d \\
  0 \leq x \perp y \geq 0
  \]

- Consider the QP problem

  \[
  \begin{align*}
  \min & \quad \frac{1}{2} y' My + d'y \\
  \text{s.t.} & \quad y \geq 0
  \end{align*}
  \]

- The corresponding KKT optimality conditions are

  \[
  My + d - x = 0 \\
  y \geq 0 \\
  x \geq 0 \\
  x_i y_i = 0, \quad i = 1, \ldots, n
  \]

  that are exactly the given LCP
WOLFE'S DUAL QP

• Consider now Wolfe's dual problem

\[
\begin{align*}
\max_{x,\lambda} & \quad \frac{1}{2} x' Q x + c' x + \lambda' (A x - b) \\
\text{s.t.} & \quad Q x + c + A' \lambda = 0, \quad \lambda \geq 0
\end{align*}
\]

• We can subtract \(0 = (Qx + c + A' \lambda)'x\) without changing the function and get the convex programming problem

\[
\begin{align*}
\max_{x,\lambda} & \quad -\frac{1}{2} x' Q x - \lambda' b \\
\text{s.t.} & \quad Q x + c + A' \lambda = 0 \\
& \quad \lambda \geq 0
\end{align*}
\]

• Note that Wolfe's dual QP only requires \(Q \succeq 0\).
• Consider again the LASSO problem

\[
\min_x \frac{1}{2} \|Ax - b\|_2^2 + \gamma \|x\|_1 \quad A \in \mathbb{R}^{m \times n}, \ b \in \mathbb{R}^m, \ \gamma > 0
\]

• With \( x = y - z \) and \( y, z \geq 0 \), LASSO becomes the positive semidefinite QP

\[
\min_{y, z \geq 0} \frac{1}{2} \|A(y - z) - b\|_2^2 + \gamma \Pi'(y + z)
\]

where \( \Pi' = [1 \ldots 1] \) (as \( \gamma > 0 \) at least one of \( y_i^*, z_i^* \) will be zero at optimality)

• The above QP is the dual of the following least distance programming (LDP) (constrained LS) problem (see next slide)

\[
\min_v \quad \frac{1}{2} \|v - b\|_2^2 - b'b \\
\text{s.t.} \quad \|A'v\|_\infty \leq \gamma
\]
• **Proof**: The constrained LS problem is equivalent to the following QP

\[
\min_v \; \frac{1}{2} v' v - b' v - \frac{1}{2} b'b \\
\text{s.t.} \; -\gamma I \leq A' v \leq \gamma I
\]

whose dual QP problem is exactly the original LASSO’s QP reformulation

\[
\min_{y,z \geq 0} \frac{1}{2} \begin{bmatrix} y \\ z \end{bmatrix} \begin{bmatrix} A' & \mathbb{I} \\ -\mathbb{I} & A' \end{bmatrix} \begin{bmatrix} A - A \\ \mathbb{I} \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} + (\gamma \begin{bmatrix} \mathbb{I} \\ \mathbb{I} \end{bmatrix} - \begin{bmatrix} A' \\ -A' \end{bmatrix} \begin{bmatrix} b \\ b \end{bmatrix})' \begin{bmatrix} y \\ z \end{bmatrix} + \frac{1}{2} b'b - \frac{1}{2} b'b
\]

\[\Box\]

• The LDP reformulation of LASSO is always a strictly convex QP with \(m\) variables, \(2n\) constraints, and Hessian = identity matrix

• The original QP formulation is only convex with \(2n\) variables and \(2n\) constraints
**Support Vector Regression**

(Smola, Schölkopf, 2004)

- We have a training set \((x_1, y_1), \ldots, (x_N, y_N)\), \(x_i \in \mathbb{R}^n, y \in \mathbb{R}\) and want to fit a linear function

  \[
  f(x) = w'x + b \quad w \in \mathbb{R}^n, b \in \mathbb{R}
  \]

  such that each \(|y_i - f(x_i)| \leq \epsilon\)

- Since such a function \(f\) may not exist, we want to penalize \(|y_i - f(x_i)| > \epsilon\)

\[
\min_{w, b, v, s} \quad \frac{1}{2} \|w\|_2^2 + C \sum_{i=1}^{N} (v_i + s_i)
\]

s.t.

- \(y_i - w'x_i - b \leq \epsilon + v_i\)
- \(y_i - w'x_i - b \geq -\epsilon - s_i\)
- \(v_i, s_i \geq 0, \quad i = 1, \ldots, N\)

\[s = \max(-\epsilon - t, 0)\]

\[v = \max(t - \epsilon, 0)\]

\[t = y - w'x - b\]
• By setting $X = [x_1 \ldots x_N], Y = [y_1 \ldots y_N]'$, we can rewrite in vector form

$$\begin{align*}
\min_{w,b,v,s} & \quad \frac{1}{2}w'w + C \mathbb{I}'(v + s) \\
\text{s.t.} & \quad Y - X'w - b \mathbb{I} \leq \epsilon \mathbb{I} + v \\
& \quad Y - X'w - b \mathbb{I} \geq -\epsilon \mathbb{I} - s \\
& \quad v, s \geq 0
\end{align*}$$

• Introduce the vectors of $\mathbb{R}^N$ of Lagrange multipliers $\alpha, \beta, \gamma, \delta \geq 0$

• The Lagrangian function is

$$\mathcal{L} \left( \begin{bmatrix} w \\ b \\ v \\ s \end{bmatrix}, \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} \right) = \frac{1}{2} w'w + C \mathbb{I}'(v + s) + \alpha' (Y - X'w - (b + \epsilon) \mathbb{I} - v)$$

$$+ \beta' (-Y + X'w + (b - \epsilon) \mathbb{I} - s) - \gamma'v - \delta' s$$

• The dual function $q(\alpha, \beta, \gamma, \delta) = \inf_{w,b,v,s} \mathcal{L}(w, b, v, s, \alpha, \beta, \gamma, \delta)$
Let us zero the partial derivatives of $\mathcal{L}$ with respect to $w, b, v, s$:

\begin{align*}
0 &= \frac{\partial \mathcal{L}}{\partial w} = w - X\alpha + X\beta \quad \Rightarrow \quad w = X(\alpha - \beta) \\
0 &= \frac{\partial \mathcal{L}}{\partial b} = -\alpha' \mathbb{I} + \beta' \mathbb{I} \quad \Rightarrow \quad \mathbb{I}'(\alpha - \beta) = 0 \\
0 &= \frac{\partial \mathcal{L}}{\partial v} = C \mathbb{I} - \alpha - \gamma \quad \Rightarrow \quad \gamma = C \mathbb{I} - \alpha \geq 0 \\
0 &= \frac{\partial \mathcal{L}}{\partial s} = C \mathbb{I} - \beta - \delta \quad \Rightarrow \quad \delta = C \mathbb{I} - \beta \geq 0
\end{align*}

By substituting the above expressions in the Lagrangian we get

\[ q(\alpha, \beta, \gamma, \delta) = \frac{1}{2}w'w + (Y - X'w)'(\alpha - \beta) - \epsilon \mathbb{I}'(\alpha + \beta) \]
\[ = -\frac{1}{2}(\alpha - \beta)'X'X(\alpha - \beta) + Y'(\alpha - \beta) - \epsilon \mathbb{I}'(\alpha + \beta) \]

The dual problem is therefore the following QP

\[ \min_{\alpha, \beta} \quad \frac{1}{2}(\alpha - \beta)'X'X(\alpha - \beta) - Y'(\alpha - \beta) + \epsilon \mathbb{I}'(\alpha + \beta) \]
\[ \text{s.t.} \quad 0 \leq \alpha \leq C \mathbb{I}, \quad 0 \leq \beta \leq C \mathbb{I}, \quad \mathbb{I}'(\alpha - \beta) = 0 \]
After solving the dual QP problem we can retrieve

\[ w = X(\alpha^* - \beta^*) = \sum_{i=1}^{N}(\alpha_i^* - \beta_i^*)x_i \]

\[ f(x) = w'x + b = (\alpha^* - \beta^*)'X'x + b = \sum_{i=1}^{N}(\alpha_i^* - \beta_i^*)x_i'x + b \]

(see next slide for how to reconstruct \( b \))

- \( f(x) \) is defined by a linear combination of the training vectors \( x_i \)
- The vectors \( x_i \) for which \( \alpha_i^* - \beta_i^* \neq 0 \) are called support vectors
- Note that the QP is also equivalent to the \( \ell_1 \)-regularized problem

\[
\begin{align*}
\min_z & \quad \frac{1}{2}z'X'Xz - Y'z + \epsilon \|z\|_1 \\
\text{s.t.} & \quad |z_i| \leq C, \quad \sum_{i=1}^{N} z_i = 0
\end{align*}
\]
The scalar \( b \) can be retrieved from the complementarity slackness conditions:

\[
\begin{align*}
0 &= \alpha_i (y_i - x_i'w - (b + \epsilon) - v_i), \quad i = 1, \ldots, N \\
0 &= \beta_i (-y_i + x_i'w + (b - \epsilon) - s_i) \\
0 &= \gamma_i v_i = (C - \alpha_i) v_i \\
0 &= \delta_i s_i = (C - \beta_i) s_i
\end{align*}
\]

- if any \( \alpha_i^* \in (0, C) \) then \( v_i^* = 0 \Rightarrow b^* = y_i - x_i'w^* - \epsilon \)
- if any \( \beta_i^* \in (0, C) \) then \( s_i^* = 0 \Rightarrow b^* = y_i - x_i'w^* + \epsilon \)
• Otherwise, consider the case all $\alpha_i^*, \beta_i^* \in \{0, C\}$

• $\alpha_i^*, \beta_i^*$ cannot be positive at the same time, as they refer to bilateral constraints ($y_i - w'x_i - b$ cannot be both positive and negative)

\[
\begin{align*}
\alpha_i = 0 & \quad \Rightarrow \quad v_i = 0 \quad \Rightarrow \quad y_i - x_i'w - (b + \epsilon) \leq 0 \\
\beta_i = 0 & \quad \Rightarrow \quad s_i = 0 \quad \Rightarrow \quad -y_i + x_i'w + (b - \epsilon) \leq 0 \\
\alpha_i = C & \quad \Rightarrow \quad \beta_i = 0 \quad \Rightarrow \quad s_i = 0, \quad -y_i + x_i'w + (b - \epsilon) \leq 0 \\
\beta_i = C & \quad \Rightarrow \quad \alpha_i = 0 \quad \Rightarrow \quad v_i = 0, \quad y_i - x_i'w - (b + \epsilon) \leq 0
\end{align*}
\]

• Let $\mathcal{I} = \{i : \alpha_i^* = 0 \text{ or } \beta_i^* = C\}$ and $\mathcal{J} = \{i : \alpha_i^* = C \text{ or } \beta_i^* = 0\}$. Then

\[
\begin{align*}
b^* & \geq y_i - x_i'w^* - \epsilon, \quad \forall i \in \mathcal{I} \\
b^* & \leq y_i - x_i'w^* + \epsilon, \quad \forall i \in \mathcal{J}
\end{align*}
\]

• Therefore, any $b^* \in \left[\max_{i \in \mathcal{I}} \{y_i - x_i'w^* - \epsilon\}, \min_{i \in \mathcal{J}} \{y_i - x_i'w^* + \epsilon\}\right]$ is optimal
Support Vector Regression

- **Kernel trick**: if we generalize $x_i$ to an arbitrary nonlinear basis $\phi(x_i)$ we get

$$f(x) = \sum_{i=1}^{N} (\alpha_i^* - \beta_i^*) k(x_i, x) + b$$

where $k(x, y) = \phi'(x)\phi(y)$ is a kernel function, $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$

- Example: $x \in \mathbb{R}^2$, $\phi(x) = [x_1^2 \quad \sqrt{2}x_1x_2 \quad x_2^2]'$, $k(x, y) = (x'y)^2$

- The $(i, j)$th term $x'_i x_j$ of the dual Hessian gets replaced by $k(x_i, x_j)$

- $b$ depends on $x'_i w = x'_i X (\alpha - \beta)$ that gets replaced by $k(x_i, X)(\alpha^* - \beta^*)$

- Therefore $\phi, w$ are not required, and can have arbitrary dimensions!

- Example: **Gaussian radial basis function kernel** $k(x, y) = e^{-\frac{1}{2} \|x-y\|^2 / \sigma^2}$ (RBF) the corresponding $\phi$ is infinite dimensional
• Generate $N = 100$ random samples of the course-logo function

$$f(x_1, x_2) = -e^{-(x_1^2 + x_2^2)} + 0.3 \sin \left( \frac{1}{10} x_1^3 + x_2^2 \right) + 1.2$$

• Solve SVR problem with $C = 100$, $\epsilon = 0.01$, Gaussian kernel with $\sigma = 1$