Master-MIRI
Topics on Optimization and Machine Learning (TOML)

José M. Barceló Ordinas
Departament d'Arquitectura de Computadors
(UPC)

The Lagrangian

- Let an optimization problem (not necessarily convex) be expressed in its standard form:

minimize	$f_0(x)$
subject to	$f_i(x) \le 0 \quad i=1,\dots,m$
	$h_i(x) = 0 \quad i=1,\dots,p$

where $x \in R^n$, and $X = f_0 \cap_{i=1\dots m} \text{dom } f_i \cap_{i=1\dots p} \text{dom } h_i$

We define the **Lagrangian** L: $R^n x R^m x R^p \longrightarrow R$ as

$$L(x,\lambda,\nu) = f_0(x) + \Sigma_{i=1,..m} \lambda_i f_i(x) + \Sigma_{i=1,..p} \nu_i h_i(x)$$

with domain $L = X x R^m x R^p$.

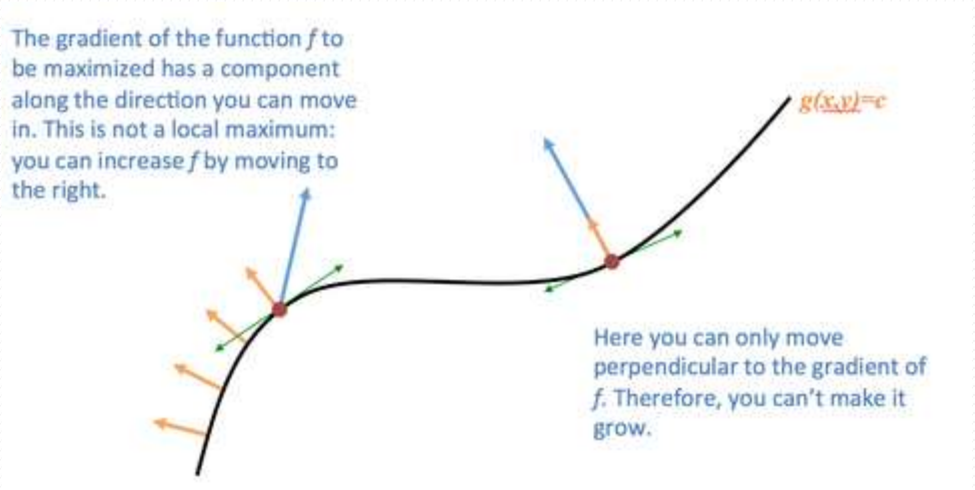
We refer to λ_i as **Lagrange multiplier** associated to the inequality constraints and ν_i as **Lagrange multiplier** associated to the equality constraints.

The Lagrangian multiplier

maximize $f(x,y)$
subject to $g(x,y)=c$

At any point (x,y), the gradient $\nabla f(x,y)$ is a vector that tells you where to head if you want to increase f as efficiently as possible. As long as you can walk in that exact direction, you're going to "go up" (increase f). If you can't go exactly in the direction of the gradient but you can go in a direction that has a non-trivial component along the gradient, you're still going to go up f, albeit more slowly. But if you can't - namely, if you can only move in a direction orthogonal to the gradient - then you're not able to increase f any more: you've reached a local maximum. Why would you be unable to move along the gradient? Well, because you have to stay on the constraint set $g(x,y)=c$. In other words, the allowed directions for you to move in are along the tangents to this constraint curve. → f and g are tangent if their gradients are parallel, however, although the two gradient vectors are parallel, the magnitudes of the gradient vectors are generally not equal:

$$\nabla f(x,y) = \lambda \nabla g(x,y)$$



- ## The Lagrange Dual function

The **Lagrange Dual function** q: $R^m x R^p \longrightarrow R$ is defined as the minimum of the Lagrangian over $\lambda \in R^m$ and $\nu \in R^p$,

$$\mathbf{q(\lambda,\nu) = inf_{x\in X}\, L(x,\lambda,\nu) = inf_{x\in X}\{\, f_0(x) + \Sigma_{i=1,..m}\, \lambda_i f_i(x) + \Sigma_{i=1,..p}\, \nu_i h_i(x)\, \}}$$

  - Note that since $q(\lambda,\nu)$ is an **infimum** of a family of affine functions, then it is a **concave** function.

- ## Lower Bounds on optimal value
  - For $\lambda \geq 0$ and any $\nu \rightarrow q(\lambda,\nu) \leq p^*$ (easy to proof, check on a feasible point x and you will see that $q(\lambda,\nu) \leq f_0(x)$ if x is feasible)
  - The pair (λ,ν) is called Dual Feasible

Some Examples

- **Least-squares solution of linear equations**:

minimize	$x^T x$
subject to	$Ax=b$

Then, $\mathbf{L(x,\nu) = x^T x + \nu^T (Ax-b)}$, is the Lagrangian

And, $\mathbf{q(\nu) = \inf_{x\in X} L(x,\nu) = \inf_{x\in X} \{ x^T x + \nu^T (Ax-b) \}}$

Since $L(x,\nu)$ is a convex quadratic function we can find the optimum:

$$\nabla_x L(x,\nu) = 2x + A^T\nu = 0 \rightarrow x=-(1/2)\, A^T\nu$$

Then,

$$q(\nu)= L(x,\nu) = L(-(1/2)\, A^T\nu,\nu)=-(1/4)\, \nu^T A A^T \nu - b^T\nu,$$

The Lagrange dual function vs the Conjugate function

- The Lagrange dual function and the conjugate function are closely related.
- Consider the following optimization problem:

minimize	$f_0(x)$
subject to	$Ax \leq b$
	$Cx = d$

The Lagrange Dual Function $q(\lambda,\nu)$ is:

$$q(\lambda,\nu) = \inf_{x\in X} L(x,\lambda,\nu) = \inf_{x\in X}\{ f_0(x)+\lambda^T(Ax-b)+\nu^T(Cx-d) \}$$

$$= -b^T\lambda - d^T\nu + \inf_{x\in X}\{ (A^T\lambda+C^T\nu)^T x + f_o(x)\}$$

$$= -b^T\lambda - d^T\nu - \sup_{x\in X}\{ (-A^T\lambda-C^T\nu)^T x - f_o(x)\}$$

$$= -b^T\lambda - d^T\nu - f_0^*(-A^T\lambda-C^T\nu)$$

Take care with inf and sup:

$$g(y) = \inf_{x\in X}\{ f_0(x) + y^T x\} = -\sup_{x\in X}\{(-y)^T x - f_o(x)\} = -f^*(-y)$$

The Lagrange Dual Problem

- For each pair (λ,ν) with $\lambda\geq 0$, $q(\lambda,\nu)\leq p^*$. Then we can question what is the best lower bound that can be obtained with the Lagrange dual function ?,

It is to say, Let us define the **Lagrange Dual Problem** as:

$$\begin{array}{ll} \textbf{maximize} & q(\lambda,\nu) \\ \textbf{subject to} & \lambda\geq 0 \end{array}$$

This problem is a COP (convex) since the objective is concave and the constraint is convex.

Let be (λ^*,ν^*) the optimal pair that solves this optimization problem → **optimal Lagrange multipliers**

Let be **d*** the solution of the Lagrange Dual Problem,

$$d^*=\sup \{ q(\lambda,\nu) \mid \lambda\geq 0\} < \infty$$

Some Examples

- **Standard form of a LP:**

minimize $c^T x$
subject to $Ax=b;\ x \geq 0$ (or $-x \leq 0$)

Then, $L(x,\lambda,\nu) = c^T x - \lambda^T x + \nu^T (Ax-b) = -b^T \nu + (c + A^T \nu - \lambda)^T x$

And, $q(\lambda,\nu) = \inf_{x \in D} L(x,\lambda,\nu) = -b^T \nu + \inf_{x \in D}\{(c + A^T \nu - \lambda)^T x\}$

then,

$$q(\lambda,\upsilon) = \begin{cases} -b^T \upsilon & A^T \upsilon - \lambda + c = 0 \\ -\infty & \text{otherwise} \end{cases}$$

The **Dual Problem** will be:

maximize $-b^T \nu$
subject to $c + A^T \nu - \lambda = 0$ and $\lambda \geq 0$

that can be expressed as

maximize $-b^T \nu$
subject to $c + A^T \nu \geq 0$

Some Examples

- **Entropy optimization:**

minimize	$f_0(x)=\Sigma_{i=1..m}\, x_i \log(x_i)$
subject to	$Ax<b$
	$1^T x=1$

The conjugate function of $f_0(x)=\Sigma_{i=1..m}\, x_i \log(x_i)$ is:

$$f^*_0(y)=\Sigma_{i=1..m} \exp(y_i-1), \qquad \text{with dom } f^*_0=R^n,$$

and then the dual function is:

$$\begin{aligned} q(\lambda,\nu) &= -b^T\lambda-d^T\nu - f_0^*(-A^T\lambda-C^T\nu) = \\ &= -b^T\lambda-1^T\nu - \Sigma_{i=1..m} \exp(-a_i^T\lambda-\nu-1) = \\ &= -b^T\lambda-1^T\nu - e^{-\nu-1}\, \Sigma_{i=1..m} \exp(-a_i^T\lambda) \end{aligned}$$

Some Examples

- **Entropy Optimization:**

And the dual Problem is:

maximize $\quad -b^T\lambda - 1^T\nu - e^{-\nu-1}\,\Sigma_{i=1..m}\exp(-a_i^T\lambda)$

subject to $\quad \lambda \geq 0$

Maximizing over ν and fixing λ:

$$\nu^* = \log\left(\Sigma_{i=1..m}\exp(-a_i^T\lambda)\right) - 1$$

And substituting the optimal value of ν in the dual problem:

maximize $\quad -b^T\lambda - \log\left(\Sigma_{i=1..m}\exp(-a_i^T\lambda)\right)$

subject to $\quad \lambda \geq 0$

that is geometric problem in its convex form

Some Examples

- **Unconstrained geometric program:**

minimize $f_0(x)= \log(\Sigma_{i=1..m} \exp(a_i^T x+b_i))$

We introduce new variables:

minimize $f_0(x)= \log(\Sigma_{i=1..m} \exp(y_i))$
subject to $a^T x+b=y$

And calculate the conjugate of $f_0(x)$ (remember from class 2),

$$f_0^*(y)= \Sigma_{i=1..m} y_i \log (y_i) \text{ with } 1^T y =1 \text{ and } y\geq 0$$

Then, the dual problem can be reformulated as:

maximize $-b^T\nu - \Sigma_{i=1..m} \nu_i \log (\nu_i)$
subject to $1^T\nu =1$
$A^T\nu =0$
$\nu \geq 0$

Which is an entropy maximization problem

Weak Duality

- By definition, the Lagrange dual function is a lower bound of p*:

$$\mathbf{q}(\lambda,\nu) \leq p^*$$

and specifically

$$\mathbf{q}(\lambda^*,\nu^*) \leq p^* \rightarrow d^* \leq p^*$$

We call the value d*-p* the **optimal duality gap** and <u>gives the best lower bound that can be obtained from the Lagrange dual function.</u>

- What conditions make d*=p* ?

- **Duality gap (geometric interpretation)**

Primal problem

minimize $f(x)$

s.t. $g(x) \leq 0$

Dual Problem

maximize $q(\mu) = \min_{x \in X} \{f(x) + \mu\, g(x)\}$

s.t. $\mu \geq 0$

Let us consider the following set $V=\{ (g(x),f(x)) \mid x \in X \}$.

- The **primal optimal f*** corresponds to the minimum vertical axis value of all points on the left half plane, i.e., all points of the form $\{ g(x) \leq 0 \mid x \in X \}$.
- The **dual value $q(\mu)$ for a feasible $\mu \geq 0$** corresponds to the vertical intercept value of all hyperplanes with normal $(\mu, 1)$ and support the set V from below. The **dual optimal q*** corresponds to the maximum intercept value of such hyperplanes over all $\mu \geq 0$.

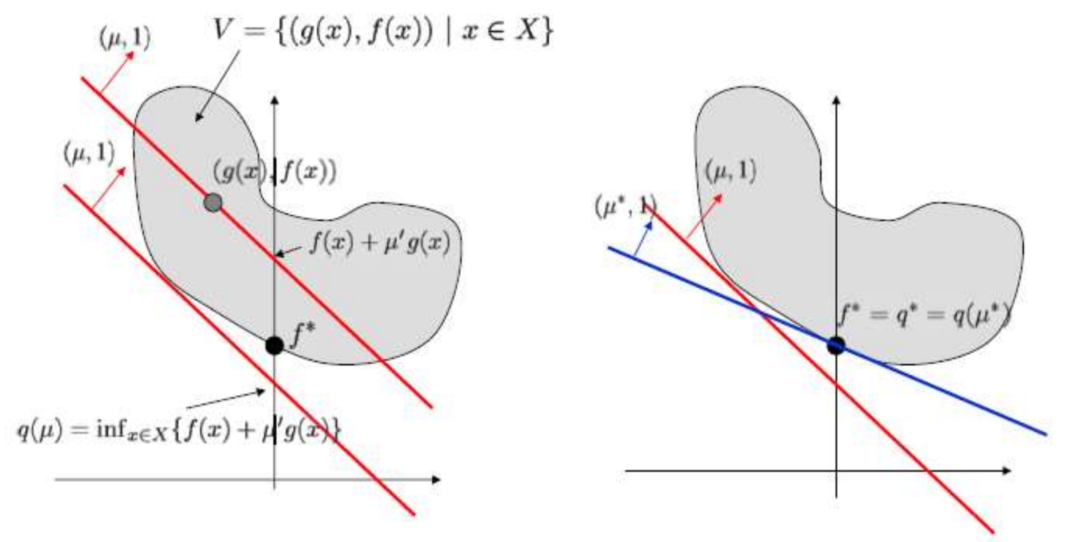


Figure 1.2 Illustration of the primal and the dual problem.

- **Strong Duality**
  - **Strong duality holds when $d^*=p^*$.**
  - If the problem is a COP many times (not always) strong duality holds.

Slater's Condition ensures strong duality:

If the primal problem is convex and there exists an $x \in \text{relint } X$, such that:

$$f_i(x)<0, \ i=1,\ldots,m \text{ and}$$
$$Ax=b,$$

where the relint X is defined as the "relative interior" of a set, then strong duality holds → $d^*=p^*$.
The **vector x** that fulfil this condition is called **Slater vector**.

(the relative interior of a set contains all points which are not on the "edge" of the set, relative to the smallest subspace in which this set lies.)

- **Duality gap – Slater Condition (geometric interpretation)**

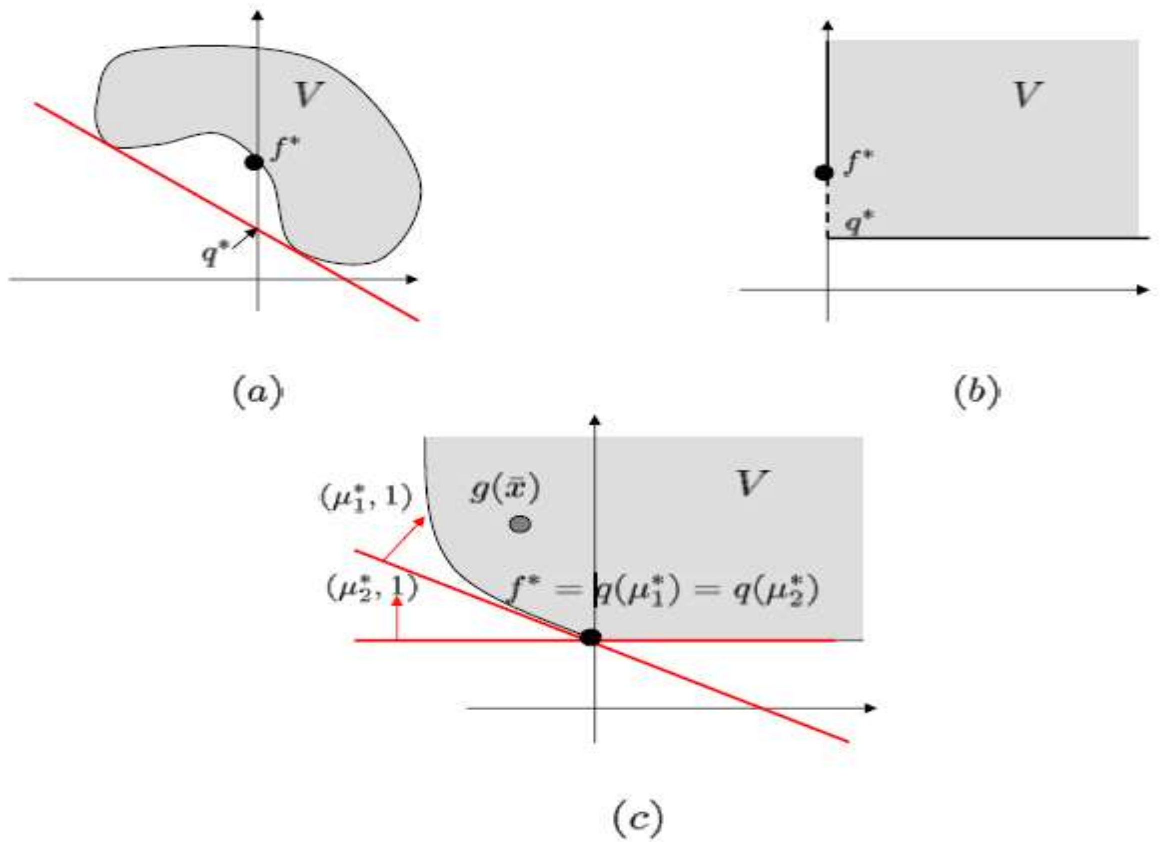


Figure 1.3 Parts (a) and (b) provide two examples where there is a duality gap [due to lack of convexity in (a) and lack of "continuity around origin" in (b)]. Part (c) illustrates the role of the Slater condition in establishing no duality gap and boundedness of the dual optimal solutions. Note that dual optimal solutions correspond to the normal vectors of the (nonvertical) hyperplanes supporting set V from below at the point $(0, q^*)$.

Reminder of saddle-points

Saddle-Point: is a point in the domain of a function that is a stationary point (**first derivative equal to zero**) but not a local extremum. There is a relative minimum between peaks (axial direction) and a relative maximum at crossing axis.

Testing whether a point is a saddle-point:

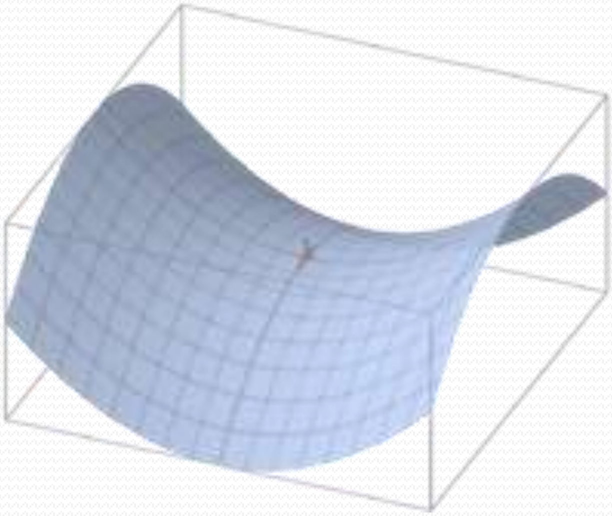
i) compute the Hessian

ii) if the Hessian is indefinite (have positive and negative eigenvalues) then the point is a saddle-point

- Example: $f(x,y)=x^2-y^2$ has Hessian

$$H=\nabla_{x,y} f(x,y)=\begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$$

is indefinite at saddle-point (0,0)



Saddle-point interpretation of Lagrange Duality

- For simplicity, let us assume that there only are inequality constraints

$$\sup_{\lambda \geq 0} L(x,\lambda) = \sup_{\lambda \geq 0} \left(f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) \right)$$

$$= \begin{cases} f_0(x) & f_i(x) \leq 0, i = 1,\ldots,m \\ \infty & otherwise \end{cases}$$

From the primal problem:

$$\mathbf{p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda)}$$

From the dual function:

$$\mathbf{d^* = \sup_{\lambda \geq 0} \inf_x L(x, \lambda)}$$

- **Saddle-point interpretation of Lagrange Duality**

From the weak duality:

$$\sup_{\lambda\geq 0} \inf_x L(x,\lambda) = d^* \leq p^* = \inf_x \sup_{\lambda\geq 0} L(x,\lambda)$$

From the strong duality:

$$\sup_{\lambda\geq 0} \inf_x L(x,\lambda) = d^* = p^* = \inf_x \sup_{\lambda\geq 0} L(x,\lambda)$$

Strong Max-min property or Saddle-point property: if for any $f: R^n x R^m \longrightarrow R$ with $W \subseteq R^n$, $Z \subseteq R^m$, the next equality is satisfied:

$$\sup_{z\in Z} \inf_{w\in W} f(w,z) = \inf_{w\in W} \sup_{z\in Z} f(w,z)$$

i. Then (w,z) is a saddle-point and w solves the primal and z solves the dual.
ii. On the other hand, if w solves the primal and satisfies the Slater Condition, there exists a z such that (w,z) is a saddle-point.

<u>Saddle-point Interpretation</u>: for any saddle-point (w^*,z^*)

$$f(w^*,z) \leq f(w^*,z^*) \leq f(w,z^*) \quad \text{for all } z\in Z \text{ and } w\in W$$

Thus, w^* minimizes $f(w,z^*)$ and z^* maximizes $f(w^*,z)$

Complementary Slackness

Let us assume strong duality → p*=d* and x* is the optimal value of the primal problem and (λ*,ν*) the optimal value of the dual problem.

$$f_0(x^*) = q(\lambda^*, \upsilon^*) = \inf_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} \upsilon_i^* h_i(x)\right)$$

$$\leq f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} \upsilon_i^* h_i(x^*) \leq f_0(x^*)$$

We conclude that:

x* minimizes L(x,λ*,ν*) → **Gradient of Lagrangian vanishes**

$\nabla_x L(x,\lambda^*,\nu^*) = 0$

Complementary slackness property holds for any primal optimal x* and any dual optimal (λ*,ν*), when strong duality holds, then $\lambda_i^* f_i(x^*)=0$ for i=1,...,m.

$\lambda_i^*>0 \quad \Rightarrow \quad f_i(x^*)=0$ for i=1,...,m

or

$f_i(x^*)<0 \quad \Rightarrow \quad \lambda_i^*=0$ for i=1,...,m

Karush-Kuhn-Tucker (KKT) optimality conditions

Let us assume tha x^* is the optimal value of the primal problem and (λ^*,ν^*) the optimal value of the dual problem.

The **KKT conditions** are:

i. **Primal constraints:** $f_i(x^*) \leq 0,$ $i=1,\ldots,m$
ii. **Primal constraints:** $h_i(x^*) = 0,$ $i=1,\ldots,p$
iii. **Dual constraints:** $\lambda_i^* \geq 0$ $i=1,\ldots,m$
iv. **Complementary slackness:** $\lambda_i^* f_i(x^*)=0$ $i=1,\ldots,m$
v. **Gradient of Lagrangian vanishes:**

$$\nabla_x L(x,\lambda^*,\nu^*) = \nabla f_0(x^*) + \Sigma_{i=1,..m} \lambda^*_i \nabla f_i(x^*) + \Sigma_{i=1,..p} \nu^*_i \nabla h_i(x^*) = 0$$

Non convex optimization problems:

Strong-duality $\Rightarrow$ KKT conditions

Convex optimization problems:

Strong-duality $\Leftrightarrow$ KKT conditions

- **Example:**

minimize $(1/2)x^TPx+q^Tx+r$

subject to $Ax=b$

The KKT conditions are:

i. $Ax^*=b$
ii. $Px^*+q+A^T\nu^*=0$

Which can be re-written in matrix form:

$$\begin{bmatrix} P & A^T \\ A & 0 \end{bmatrix}\begin{bmatrix} x^* \\ \upsilon^* \end{bmatrix}=\begin{bmatrix} -q \\ b \end{bmatrix}$$

that is a set of m+n equations whose solution gives the optimal primal and dual.