# Clase 6: Multiplicadores de Lagrange II

Ignacio Ramírez

31 de agosto de 2016

*Estos apuntes son preliminares y en buena parte están incompletos, pero sirven de base para seguir el contenido esencial de la clase.*

## 1. Condiciones de desigualdad

Consideremos ahora el caso en que algunas de las restricciones del problema son de la forma $g_j(x) \leq 0$,

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \dots, m \\ & g_j(x) \leq 0,\ j = 1, \dots, r \end{array} \tag{1}$$

Una forma de ver este problema es reducirlo al caso de igualdad de la siguiente manera.

Definamos el *conjunto activo* $A(x) = \{j : g_j(x) = 0\}$, es decir $A(x)$ incluye los índices $j$ de aquellas restricciones $g_j(x)$ que están *activas*.

Supongamos ahora que $x^*$ es un mínimo local del problema (1). Resulta obvio que $x^*$ es también mínimo local de un problema para el cual las restricciones no activas $j \notin A(x^*)$ son descartadas. Por otro lado, las restricciones *activas* en $x^*$ pueden ser tratadas como restricciones de igualdad. Entonces, $x^*$ cumple las condiciones necesarias del teorema de Lagrange,[1]

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla g_j(x^*) = 0.$$

Si ahora definimos $\mu_j = 0$ para $j \notin A(x^*)$ llegamos a

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0, \tag{2}$$

$$\mu_j^* = 0, \quad j \notin A(x^*) \tag{3}$$

Las condiciones anteriores, junto a alguna cosita más, constituyen las llamadas condiciones Karush-Kuhn-Tucker (KKT). Un dato adicional de éstas condiciones es que los $\mu_j^*$ son no negativos. Esto es fácil de ver siguiendo el argumento de sensibilidad visto anteriormente (pendiente:dibujo). El enunciado formal es el siguiente.

**Teorema 1** (Condiciones necesarias de Karush-Kuhn-Tucker)**.** *Sea $x^*$ mínimo local del problema*

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \dots, m \\ & g_j(x) \leq 0,\ j = 1, \dots, r, \end{array} \tag{4}$$

[1] Sujeto al tecnicismo de que $x^*$ sea *regular* para ese problema, es decir, que $\{\nabla h_i(x^*)\} \bigcup \{\nabla g_j(x^*) : j \in A(x^*)\}$ sea un conjunto l.i. Este tecnicismo se puede evitar, como veremos más adelante, pero es necesario para el teorema clásico, las KKT.

*donde $f(x)$, $h_i(x)$ y $g_j(x)$ son continuamente diferenciables y $x^*$ es regular. Entonces existen vectores multiplicadores de Lagrange $\lambda^*$ y $\mu^*$ tales que*

$$\nabla_x L(x, \lambda^*, \mu^*) = 0, \tag{5}$$
$$\mu_j^* \geq 0, \quad j = 1, \ldots, r \tag{6}$$
$$\mu_j^* = 0, \quad j \in A(x^*), \tag{7}$$

*donde $A(x^*)$ es el conjunto de restricciones activas de $x^*$. Si además $f(x)$, $h_i(x)$ y $g_j(x)$ son doblemente diferenciables entonces*

$$d^T \nabla_{xx}^2 L(x, \lambda^*, \mu^*) d \geq 0$$

*para $\{d : \nabla h_i(x^*)^T d = 0\,, \nabla g_j(x^*) = 0, \forall\, j \in A(x^*)\}$.*

No vamos a demostrarlo aquí, si bien la demostración no es difícil (ver Bertsekas p. 316). Cabe notar que todo se extiende naturalmente a este caso, incluso el concepto de que la hessiana de $f(x)$ sea semidefinida positiva en el "subespacio de direcciones factibles", ahora incluyendo las restricciones de desigualdad activas.

También, al igual que en los otros casos, se obtienen las condiciones suficientes (para funciones doblemente diferenciables) al exigir que $\nabla_x^2 xL(x^*, \lambda^*, \mu^*)$ sea definida positiva sobre el subespacio de direcciones factibles; en este caso se agrega la condición adicional de que $\mu_j^* > 0$ para las restricciones activas (de nuevo evidenciando el crecimiento estricto de la función al momento de llegar al borde del conjunto factible).

## 2. Condiciones de Fritz John

Todas las condiciones necesarias que vimos anteriormente (no las suficientes) exigen que $x^*$ sea *regular*, algo que en principio no es del todo intuitivo. Para el caso de restricciones de igualdad esto es exigir que los $\{\nabla h_i\}$ sean l.i., de modo que siempre pueda haber una combinación lineal que produzca $-\nabla f$:

$$\nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) = 0.$$

Las condiciones de Fritz John son una versión más moderna de las KKT, y por ende del teorema de Lagrange, que resuelven el problema agregando un multiplicador a $\nabla f(x^*)$,

$$\mu_0^* \nabla f(x^*) + \sum_i \lambda_i^* \nabla h_i(x^*) + \sum_j \mu_j^* g_j(x^*) = 0,$$

con el requerimiento adicional de que $\mu_0^*, \lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*$ no son todos cero a la vez.

## 3. Conversión a igualdad

Es fácil convertir problemas con desigualdades en igualdades y viceversa. En el primer caso, podemos convertir

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \ldots, m \\ & g_j(x) \leq 0,\ j = 1, \ldots, r \end{array} \tag{8}$$

a un problema equivalente con igualdades, agregando variables auxiliares $z_j$ de la siguiente manera:

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \ldots, m \\ & g_j(x) + z_j^2 = 0,\ j = 1, \ldots, r \end{array} \tag{9}$$

Para reescribirlo de la forma canónica basta definir $g'_j(x, z_j) = g_j(x) + z_j^2$.

El caso inverso es más fácil todavía. Para convertir un problema de igualdades en desigualdades basta reescribir,

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & h_i(x) = 0,\ i = 1, \dots, m \end{array} \tag{10}$$

como

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & -h_i(x) \leq 0,\ i = 1, \dots, m \\ & h_i(x) \leq 0,\ i = 1, \dots, m \end{array} \tag{11}$$

### 3.1. Ejemplos/ejercicios

**Restricciones de igualdad** Bertsekas 3.1.3, 3.1.6 (media geométrica vs. aritmética), 3.1.7 (centroide restringido), 3.1.8 (angulos de un triangulo).

**Desigualdad** 3.3.1 (sencillito para ir calentando), 3.3.2, codigo de Shannon,

## 4. Métodos basados en el Lagrangeano

En esta parte vamos a ver cómo resolver problemas con restricciones en base a secuencias de problemas sin restricciones. Hay dos grandes familias de métodos de este tipo.

Una es la de los *métodos de punto interior*, pensados originalmente para problemas lineales. Éstos métodos revolucionaron la optimización en la década de los '80, con una serie de trabajos de Kamarkar.

La otra se basa en el Lagrangeano aumentado. Mientras que los primeros son algo maduro y estándar, los segundos están, al momento de escribir estas notas, en el frente de la optimización moderna, siendo adaptables a problemas muy diversos, en particular a problemas modernos para los cuales utilizar métodos tipo punto interior puede no ser una buena idea.

### 4.1. Breve reseña sobre métodos de punto interior o barrera

Consideermos el problema

$$\begin{array}{ll} \text{mín} & f(x) \\ \text{sujeto a} & g(x) \leq 0. \end{array} \tag{12}$$

La idea es resolver en su lugar el problema

$$\text{mín}\, f(x) + \epsilon^k B(x) \tag{13}$$

para una secuencia *decreciente* de valores $\epsilon^k$. La función $B(x)$ es acotada en el *interior* del conjunto factible $S = \{x : g(x) < 0\}$ y tiende a infinito rápidamente al acercarse al borde. Ejemplos típicos de $B(x)$ son $\sum_j \log -g_j(x)$ o $-\sum_j \frac{1}{g_j(x)}$.

De esta manera, todas las soluciones $x^k$ de (13) son puntos factibles del problema original. Luego puede verse (muy fácil de ver geométricamente) que a medida que $\epsilon^k$ decrece, la penalización se vuelve inconsecuente en $S$ y por ende $x^k$ tiende a la solución $x^*$ del problema original (12). Es más, si la solución $x^*$ está en el borde, puede verse que $x^k$ se acerca arbitrariamente cerca a $x^*$ a medida que aumenta $k$.

## 5. Lagrangeano aumentado

Recordemos la forma del lagrangeano aumentado para resolver el problema mín $f(x)$ sujeto a $h(x) = 0$,

$$L_c(x, \lambda) = f(x) + \lambda h(x) + \frac{c}{2}\|h(x)\|^2, \tag{14}$$

La idea es aproximar la solución del problema original mediante la solución de (14). Hay dos formas de lograr una buena aproximación: a) usando una buena estimación de $\lambda^*$, b) con un $\lambda$ arbitrario, haciendo tender $c$ a infinito. Veamos el Ejemplo 3.2.1 del Bertsekas como ilustración.

## 5.1. El método de penalización cuadrática

Veamos primero el caso más simple de implementar, que es ir aumentando $c$. Históricamente éste es el primer método que se utilizó, incluso con $\lambda = 0$. Las siguientes proposiciones, ambas fáciles de demostrar, establecen la convergencia de este tipo de métodos.

**Proposición 1** (Convergencia del método de penalización cuadrático exacto)**.** *Supóngase que $f$ y $h$ son continuas, que $\{x : h(x) = 0\}$ es un conjunto no nulo, y que para cada $k = 1, \ldots$ se puede calcular el mínimo global $x^k$ de,*

$$\min L_{c^k}(x, \lambda^k), \tag{15}$$

*donde $\{\lambda^k\}$ es acotada y $0 < c^k < c^{k+1}$ para todo $k$ con $c^k \to \infty$. Entonces todo punto límite de la secuencia $\{x^k\}$ es un mínimo global del problema original.*

En la práctica uno no obtiene el mínimo global exactamente, sino que se contenta con un punto cercano, por ejemplo uno que cumpla

$$\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k,$$

para $\epsilon^k$ pequeño. La siguiente proposición muestra que aún en este caso se puede llegar al mínimo global, aunque bajo condiciones sobre $\epsilon^k$, la vieja y querida condición de regularidad.

En realidad en este caso se muestra que se llega a un punto que cumple las condiciones necesarias de optimalidad, no necesariamente el óptimo. Por otro lado, el método de demostración de la siguiente proposición nos da explícitamente el valor en el límite de $\lambda^*$, y como veremos adelante, una forma de actualizarlo.

**Proposición 2** (Convergencia del método de penalización cuadrático exacto)**.** *Supóngase que $f$ y $h$ son continuamente diferenciables, que $\{x : h(x) = 0\}$ es un conjunto no nulo, y que para cada $k = 1, \ldots$ se puede obtener un punto $x^k$ tal que,*

$$\|\nabla_x L_{c^k}(x^k, \lambda^k)\| \leq \epsilon^k, \tag{16}$$

*donde $\{\lambda^k\}$ es acotada, $\epsilon^k \geq 0$, $\epsilon^k \to 0$, $0 < c^k < c^{k+1}$, $c^k \to \infty$. Además se asume que una subsecuencia de $x^k$, $\{x^k\}_K$ converge a un $x^*$ tal que $\nabla h(x^*)$ tiene rango completo $m$. Entonces,*

$$\{\lambda^k + ch(x^k)\}_K \to \lambda^*,$$

*donde $(x^*, \lambda^*)$ satisfacen las condiciones necesarias de primer orden,*

$$\nabla f(x^*) + \lambda^* \nabla h(x^*) = 0\,, \quad h(x^*) = 0.$$

*Demostración.* (pendiente, la hacemos en el pizarron por ahora). □

## 5.2. Consideraciones prácticas

Ver ejemplo 4.2.2

El método anterior tiene como problema principal el mal condicionamiento al que eventualmente se llegua para valores grandes de $c^k$. Una posibilidad es utilizar un escalado tipo Newton (común o diagonalizado); eso ayuda mucho. Lo otro que ayuda mucho es utilizar *warm restarts*, es decir, al resolver el problema para $c^{k+1}$, utilizar la solución del problema anterior $x^k$ como punto de partida. Aquí surge una relación de compromiso: si $c^k$ y $c^{k+1}$ son muy distintos, entonces $x^k$ puede ser muy distinto de $x^{k+1}$ y no ayudar demasiado en acelerar la resolución del nuevo problema. Por otro lado, la convergencia del método en su totalidad se acelera mucho si $c^k$ crece rápidamente.

### 5.3. Manejo de desigualdades

Podemos trabajar desigualdades en el marco del método anterior convirtiendo las desigualdades en igualdad como vimos en (8) y (9). Luego las proposiciones anteriores se aplican sin cambios al problema convertido. Además, resulta que es posible optimizar separadamente los $x^k$ y los $z_j^k$, y que la actualización de éstos últimos tiene una forma cerrada muy sencilla, por lo que el costo computacional adicional de agregar esas variables es modesto.

(ver libro, capaz que lo vemos en el pizarrón)

El único problema es que el la Hessiana del lagrangeano aumentado puede ser discontinua, lo cual motiva la existencia de otros métodos que luego veremos.

## 6. Método de los multiplicadores de Lagrange

Consideremos ahora la posibilidad de actualizar la secuencia $\{\lambda^k\}$ de modo que tienda al vector de multiplicadores de Lagrange $\lambda^*$. Bajo ciertas condiciones, esto evita la necesidad de llevar $c^k$ a $\infty$, lo cual reduce el mal condicionamiento de los subproblemas. Como ventaja adicional, la tasa de convergencia de estos métodos es mejor que la de los que sólo miran $c^k$.

### 6.1. Método exacto

La idea es aplicar el método visto anteriormente, actualizando

$$\lambda^{k+1} = \lambda^k + ch(x^k).$$

La idea sale de la proposición 2, donde se tenía que $\{\lambda^k + c^k h(x^k)\} \to \lambda^*$ cuando $x^k$ se aproxima al óptimo.

Ver ejemplo 4.2.4

### 6.2. Método inexacto

De la misma manera que para el caso del método de la penalización, se puede mostrar convergencia si en lugar de exigir que $x^k$ sea la solución exacta de cada subproblema, se exige sólamente que el gradiente del lagrangeano aumentado en $x^k$ tenga norma decreciente y tendiendo a cero.

### 6.3. Convergencia, tasa de convergencia, consideraciones prácticas

Como se dijo, puede demostrarse que los métodos anteriores convergen a un mínimo local (global si el problema es convexo). También puede demostrarse que la convergencia de estos métodos es lineal, y es *superlineal* si $c^k \to \infty$. Claro que la linealidad aquí se aplica al paso "grande" $k$; si el problema se vuelve mal condicionado a medida que aumenta $k$, los problemas pueden ser cada vez más difíciles de resolver y en total arruinar la eficiencia del algoritmo.

En definitiva, hay que probar con distintas recetas. Lo que se sugiere en el Bertsekas es utilizar una ley exponencial para $c^k$, $c^{k+1} = \beta c^k$, con $\beta > 1$. También se sugiere utilizar valores de $\beta$ de hasta 10, siempre y cuando se utilice un método robusto frente al mal condicionamiento para resolver los subproblemas, por ejemplo Newton.

### 6.4. Penalizaciones no cuadráticas para problemas con desigualdades

En problema con la penalización cuadrática es que la Hessiana es no diferenciable incluso si $f$ y $h$ lo son. En este caso puede sustituirse la penalización cuadrática por otra que cumpla el mismo rol y que garantice que la Hessiana sea siempre diferenciable y definida positiva. Un ejemplo es la función $B(x) = e^x - 1$.