UNIVERSIDAD CENTROCCIDENTAL
"LISANDRO ALVARADO"

Decanato de Ciencias y Tecnología
Departamento de Matemáticas.

Una mejora del método de Nesterov

Trabajo Especial de Grado presentado por

Br. Hibelmar Pastora Mendoza Riera.

Como requisito final
para obtener el título de Licenciado
en Ciencias Matemáticas

Área de Conocimiento: Matemática aplicada.
Tutor: Dr. Javier Hernández Benítez.

Barquisimeto - Venezuela
1 de julio de 2010

A Dios que es mi fortaleza y
a mi padre, este logro también es tuyo papá

Agradecimientos

Quiero dar las gracias por este logro primeramente a Dios quien me guía en cada paso que doy.

A mi padre José Mendoza quien ha sido el pilar fundamental en este logro, por su compañía, abnegación, sus gestos, porque ha sabido compartir esta felicidad conmigo, te quiero papá.

A mi madre Nancy de Mendoza por sus atenciones, sus consejos, sus cuidados, son tantas las cosas que debo agradecerte mamá.

A mi novio Danilo López por compartir todos estos años conmigo, por cada alegría, tristeza, consejo, porque has sido tu quien ha vivido cada momento difícil en la carrera, por llenarme de valor y por tantos desvelos estudiando. Te amo mi amor.

A mi tía Mariela Riera porque ha sido para mi una hermana, por su cariño, sus atenciones, su ayuda, sus cuidados, porque eres para mi un ejemplo de lucha, sacrificio y dedicación, te admiro.

A mi tía Rafaela Mendoza por estar siempre pendiente de mi, llena de risas y ocurrencias haciendo que los momentos tristes se conviertan en alegres. Gracias por todo tía.

A mi tío Edilio Mendoza, porque se que desde donde estas compartes conmigo este momento. Siempre te llevaré en mis recuerdos.

Al Dr. Javier Hernández por todos los conocimientos brindados, por su disposición y apoyo e inculcar en mi la motivación por la optimización y la programación además de ayudarme en el desarrollo de mi trabajo especial de grado. Mil gracias profe.

A mi hermano Augusto Mendoza por su compañía.

A mis amigos, Yesika Valera, Katherina Bastida y Jesus Freitez por su compañerismo.

A todos, Gracias.

Resumen

Se hará una disertación del algoritmo propuesto por Gonzaga y Karas [3] el cual consiste en una modificación del algoritmo de primer orden para programación convexa, propuesto por Nesterov [8]. El Algoritmo resultante mantiene la complejidad obtenida por Nesterov sin necesidad de conocer la constante de Lipschitz para el gradiente de la función objetivo.

Índice general

Índice de figuras

Introducción

El trabajo de Gonzaga y Karas [3] muestra un algoritmo que ofrece una solución optima al problema de programación no lineal

$$\begin{aligned} \text{mín} \quad & f(x) \\ s.\,a. \quad & x \in \mathbb{R}^n \end{aligned} \tag{1}$$

donde $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es convexa y continuamente diferenciable, con constante de Lipschitz $L > 0$ para el gradiente de $f(\cdot)$ y con parámetro de convexidad $\mu \geq 0$. Los algoritmos descritos en el mismo usan el parámetro de convexidad μ pero sin la suposición de que sea positivo. Se supone que el problema tiene una solución óptima y no se supone que la constante de Lipschitz L es conocida aunque su conocimiento puede ser muy útil, de hecho no es necesario que $L < \infty$, a menos que se requiera para el análisis de complejidad. Por lo tanto se puede usar, por ejemplo, en problemas donde participen barreras logarítmicas.

El método más conocido para resolver el problema (1) es el algoritmo de máximo descenso ideado por Cauchy en el siglo XIX. Este construye una sucesión (x^k) dada por $x^{k+1} = x^k - \nu_k \bigtriangledown f(x^k)$. El tamaño de paso ν_k debe producir un gran decrecimiento de $f(\cdot)$. Existen varios métodos de búsqueda lineal en la literatura (ver, por ejemplo [1], [11]). Si se conoce L entonces el tamaño de paso $\nu_k = 1/L$ asegura la convergencia global del algoritmo.

En la segunda mitad del siglo pasado hubo mucha actividad en el desarrollo de algoritmos Cuasi - Newton, con la construcción iterativa de matrices que de cierto modo fueran una buena aproximación de la matriz hessiana $\bigtriangledown^2 f(x^k)$ con el fin de lograr una convergencia superlineal. Todos estos métodos calculan únicamente primeras derivadas, así que cada iteración está basada en la acumulación

de información de primer orden de las previas iteraciones. El objetivo principal de estos métodos es obtener altas velocidades de convergencia asintótica.

El último cuarto del siglo XX fue testigo de la llegada de la teoría de la complejidad computacional que envuelve la optimización convexa continua. La programación lineal y cuadrática se revolucionó con la llegada de los métodos de punto interior, y por primera vez los métodos para la optimización diferencial los cuales eran eficientes en la practica.

Los resultados de complejidad para la programación no lineal son limitados para problemas convexos, pero son impresionantes (ver [6], [10]): ningún método para resolver el problema (1) basado en la acumulación de información de primer orden puede lograr una iteración con cota de complejidad por debajo de $O(1/\sqrt{\epsilon})$ donde $\epsilon > 0$ es la precisión absoluta del valor de la función objetivo final.

El método de máximo descenso no puede lograr una mejor complejidad que $O(1/\epsilon)$. Estos resultados y muchos más son explicados en el libro de Yurii Nesterov [8].

El trabajo de Gonzaga y Karas se basa en el ingenioso estudio de máximo descenso de Nesterov, primero publicado en 1983 [7]. El demuestra que con un incremento en el esfuerzo de cálculo por iteración, la complejidad del método de máximo descenso se reduce hasta el valor óptimo $O(\frac{1}{\sqrt{\epsilon}})$. El método tiene la habilidad de resultados de complejidad. La teoría siguió latente por muchos años y hasta ahora llamando la atención de la comunidad de optimización continua.

Gonzaga y Karas hacen una mejora del método de Nesterov el cual se basa en gran medida en el conocimiento de la constante de Lipschitz L y logra la complejidad óptima con un posible menor tiempo computacional por iteración: sólo un cálculo del gradiente y no hay evaluaciones de la función (sin contar las utilizadas en la búsqueda lineal). Se tratará esta economía en las evaluaciones de la función para una búsqueda lineal inexacta extra eliminando la necesidad de conocer L o μ y probando la eficiencia en cada iteración, manteniendo al mismo tiempo la complejidad óptima en el número de iteraciones.

Capítulo 1

Preliminares

En este capítulo se presentan algunos aspectos necesarios para el desarrollo de los capítulos siguientes.

La norma $\|\cdot\|$ utilizada en este trabajo es la conocida norma euclidea, definida como,

$$\|x\| = \sqrt{\sum_{i=1}^{n} x_i^2},$$

donde $x = (x_1, x_2, ..., x_n)^T$.

Recordemos que en $\mathbb{R}^n$ tenemos una propiedad muy importante conocida como la desigualdad de Cauchy Schwarz la cual nos dice que para todo $x, y \in \mathbb{R}^n$ se cumple

$$y^T x \leq \|y\|\|x\|.$$

Definición 1.1. Una función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es lipshitziana en $\mathbb{R}^n$ si existe una constante $L > 0$ tal que para todo $x, y \in \mathbb{R}^n$ se cumple,

$$\|f(x) - f(y)\| \leq L\|x - y\|,$$

L se llama constante de Lipschitz.

Definición 1.2. Sea C un conjunto convexo no vacío en $\mathbb{R}^n$. Una función $f : C \longrightarrow \mathbb{R}$ es convexa en C cuando para todo $x, x' \in C$ y todo $\alpha \in [0,1]$ se cumple que,

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha)f(x').$$

Diremos que $f(\cdot)$ es estrictamente convexa en C cuando la desigualdad anterior se cumple de manera estricta siempre que $x \neq x'$.

Definición 1.3. Sea C un conjunto convexo no vacío en $\mathbb{R}^n$. Diremos que una función $f : C \longrightarrow \mathbb{R}$ es fuertemente convexa en C si existe $c > 0$ tal que para todo $x, x' \in C$ y todo $\alpha \in [0,1]$ se cumple la desigualdad,

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha)f(x') - \frac{1}{2}c\alpha(1-\alpha)\|x - x'\|^2,$$

c es llamado el parámetro de convexidad fuerte.

Proposición 1.1. *(ver [4]) Una función $f(\cdot)$ es fuertemente convexa en C con parámetro de convexidad c si y solo si la función $f(\cdot) - \frac{1}{2}c\|\cdot\|^2$ es convexa en C.*

Teorema 1.1. *Sea $f(\cdot)$ una función diferenciable en un conjunto abierto $\Omega \subset \mathbb{R}^n$ y sea C un conjunto convexo en Ω. Entonces*

(i) $f(\cdot)$ es convexa en C si y solo si

$$f(x) \geq f(x_0) + \bigtriangledown f(x_0)^T(x - x_0)$$

para todo $x, x_0 \in C$.

(ii) $f(\cdot)$ es estrictamente convexa en C si y solo si la desigualdad anterior se sostiene de forma estricta siempre que $x \neq x_0$;

(iii) $f(\cdot)$ es fuertemente convexa en C con parámetro de convexidad fuerte c si y solo si para todo $x_0, x \in C$,

$$f(x) \geq f(x_0) + \bigtriangledown f(x_0)^T(x - x_0) + \frac{1}{2}c\|x - x_0\|^2.$$

Prueba:
(i) y (ii) ver: [4], páginas 183 y 184.

Probemos (iii)
Aplicando la desigualdad dada en (i) a la función diferenciable $f(\cdot) - \frac{1}{2}c\|\cdot\|^2$ para $x, x_0 \in C$ arbitrarios, tenemos,

$$f(x) - \frac{1}{2}c\|x\|^2 \geq f(x_0) - \frac{1}{2}c\|x_0\|^2 + (\bigtriangledown f(x_0) - cx_0)^T(x - x_0).$$

Realizando operaciones y usando la proposición 1.1 obtenemos lo deseado.

■

Teorema 1.2. ***(Taylor de orden*** 0***)*** *(ver [11]) Sea* $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ *una función continuamente diferenciable y* $x \in \mathbb{R}^n$ *entonces para todo* $y \in \mathbb{R}^n$*,*

$$f(y) = f(x) + E_0(y) = f(x) + \int_0^1 \bigtriangledown f(x + \tau(y - x))^T(y - x)d_\tau$$

Teorema 1.3. ***(Valor intermedio)*** *(ver [9]) Si* $g(\cdot)$ *es un a función a valores reales continua en un intervalo* I*, entonces* $g(\cdot)$ *tiene la propiedad del valor intermedio en* I*: siempre que* $a, b \in I$*,* $a < b$ *y* z *este entre* $f(a)$ *y* $f(b)$ *(es decir,* $f(a) < z < f(b)$ *o* $f(b) < z < f(a)$*), existe al menos un* $x \in (a, b)$ *tal que* $f(x) = z$*.*

Definición 1.4. El epígrafo de una función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es el conjunto de puntos situados en o sobre su gráfico, es decir,

$$epif = \{(x, \mu) : x \in \mathbb{R}^n, \mu \in \mathbb{R}, f(x) \leq \mu\}$$

Definición 1.5. Un punto x es combinación lineal convexa de n puntos si se cumple que:

$$x = \sum_{i=1}^{n} \lambda_i x_i,$$

con $\lambda_i \geq 0$ y $\sum_{i=1}^{n} \lambda_i = 1$.

Con el motivo de estudiar la convergencia en el algoritmo de Nesterov se introduce la siguiente definición.

Definición 1.6. Un par de sucesiones $(\phi_k(\cdot))_{k=0}^{\infty}$ y $(\lambda_k)_{k=0}^{\infty}$, $\lambda_k \geq 0$ se llama una sucesión estimada de la función $f(\cdot)$ si

$$\lambda_k \longrightarrow 0$$

y para cada $x \in \mathbb{R}^n$ y todo $k \geq 0$ tenemos,

$$\phi_k(x) \leq (1-\lambda_k)f(x) + \lambda_k\phi_0(x).$$

El siguiente lema muestra el objetivo de esta definición.

Lema 1.1. *(ver [8]) Sea $(\phi_k(\cdot))$ y (λ_k) una sucesión estimada de la función $f(\cdot)$. Si para alguna sucesión (x^k) tenemos que,*

$$f(x^k) \leq \phi_k^* = \min_{x\in\mathbb{R}^n} \phi_k(x)$$

entonces

$$f(x^k) - f^* \leq \lambda_k(\phi_0(x^*) - f^*) \longrightarrow 0.$$

La forma de una sucesión estimada la ofrece el siguiente lema.

Lema 1.2. *(ver [8]) Supongamos que:*

1. *$f(\cdot)$ es una función convexa con parámetro de convexidad $\mu \geq 0$ y continuamente diferenciable con constante de Lipschitz $L > 0$ para el gradiente.*
2. *$\phi_0(\cdot)$ es una función arbitraria de $\mathbb{R}^n$.*
3. *(y^k) es una sucesión arbitraria en $\mathbb{R}^n$*
4. *(α_k): $\alpha_k \in (0,1)$, $\sum_{k=0}^{\infty} \alpha_k = \infty$.*
5. *$\lambda_0 = 1$.*

Entonces el par de sucesiones $(\phi_k(\cdot))$, (λ_k) definidas recursivamente por:

$$\lambda_{k+1} = (1-\alpha_k)\lambda_k,$$

$$\phi_{k+1}(x) = (1-\alpha_k)\phi_k(x) + \alpha_k\left(f(y^k) + \bigtriangledown f(y^k)^T(x-y^k) + \frac{\mu}{2}\|x-y^k\|^2\right),$$

es una sucesión estimada.

Definición 1.7. (Orden de complejidad)

Dada dos sucesiones de escalares (n^k) y (v^k), escribiremos que

$$n^k = O(v^k)$$

si existe una constante $C > 0$ tal que

$$n^k \leq Cv^k$$

para todo k suficientemente grande y

$$n^k = \Omega(v^k)$$

si existe una constante $C_1 > 0$ tal que

$$n_k \geq C_1 v_k$$

para todo k suficientemente grande.

Definición 1.8. Dada una función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$, diremos que:

- x^* es un minimizador local de $f(\cdot)$ si existe una vecindad $\mathcal{N}$ de x^* en $\mathbb{R}^n$ tal que

  $$f(x^*) \leq f(x)$$

  para todo $x \in \mathcal{N}$

- x^* es un minimizador global de $f(\cdot)$ si

  $$f(x^*) \leq f(x)$$

  para todo $x \in \mathbb{R}^n$. En este caso $f(x^*)$ es llamado el valor mínimo de $f(\cdot)$.

Teorema 1.4. *(ver [11])*

Sea $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. Cuando $f(\cdot)$ es convexa, cada minimizador local x^* de $f(\cdot)$ es un minimizador global de $f(\cdot)$. Si adicionalmente $f(\cdot)$ es diferenciable entonces cada punto estacionario x^* es un minimizador global de $f(\cdot)$.

Definición 1.9. Una función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ se dice que es cuadrática si

$$f(x) = \frac{1}{2}x^T Ax + b^T x + c,$$

donde $A \in M_n(\mathbb{R})$ es simétrica, $b \in \mathbb{R}^n$ y $c \in \mathbb{R}$.
Las expresiones del gradiente y del hessiano son:

$$\bigtriangledown f(x) = Ax + b \tag{1.1}$$

$$\bigtriangledown^2 f(x) = A. \tag{1.2}$$

Si $A = aI_n$ con $a \in \mathbb{R}$ e I_n la matriz identidad de orden n, entonces diremos que $f(\cdot)$ es cuadrática simple.

Cabe destacar que la combinación lineal de funciones cuadráticas simples es una función cuadrática simple.

Definición 1.10. Una matriz $A \in M_n(\mathbb{R})$ simétrica es:

- Definida positiva si $x^T Ax > 0$ para todo vector $x \in \mathbb{R}^n$ no nulo.
- Definida negativa si $x^T Ax < 0$ para todo vector $x \in \mathbb{R}^n$ no nulo.
- Semidefinida positiva si $x^T Ax \geq 0$ para todo vector $x \in \mathbb{R}^n$.
- Semidefinida negativa si $x^T Ax \leq 0$ para todo vector $x \in \mathbb{R}^n$.

Proposición 1.2. *(ver [11]) Sean $C \subset \mathbb{R}^n$ convexo, $f : C \longrightarrow \mathbb{R}$ una función dos veces continuamente diferenciable sobre C.*

(a) *Si para todo $x \in C$ $\bigtriangledown^2 f(x)$ es definida positiva, entonces $f(\cdot)$ es estrictamente convexa.*

(b) *Si para todo $x \in C$ $\bigtriangledown^2 f(x)$ es semidefinida positiva, entonces $f(\cdot)$ es convexa.*

(c) *Si para todo $x \in C$ $\bigtriangledown^2 f(x)$ es definida negativa, entonces $f(\cdot)$ es estrictamente cóncava.*

(d) *Si para todo $x \in C$ $\bigtriangledown^2 f(x)$ es semidefinida negativa, entonces $f(\cdot)$ es cóncava.*

Afirmación 1.1. Toda función $f(\cdot)$ cuadrática simple estrictamente convexa se puede expresar de la forma

$$f(x) = f^* + \frac{a}{2}\|x - x^*\|^2,$$

con $a > 0$, x^* el minimizador de $f(\cdot)$ y f^* el valor mínimo de $f(\cdot)$.

En efecto, por la definición 1.9 se tiene que existen $a > 0, b \in \mathbb{R}^n$ y $c \in \mathbb{R}$ tales que

$$f(x) = \frac{1}{2}x^T aIx + b^T x + c,$$

desarrollando esta expresión, se obtiene,

$$\begin{aligned} f(x) &= \frac{1}{2}x^T aIx + b^T + c \\ &= \frac{1}{2}a\left(x^T x + \frac{2}{a}b^T x + \frac{2}{a}c\right) \\ &= \frac{1}{2}a\left(x^T x + \frac{2}{a}b^T x + \frac{1}{a^2}b^T b - \frac{1}{a^2}b^T b + \frac{2}{a}c\right) \\ &= \frac{1}{2}a\left(\left(x + \frac{1}{a}b\right)^T\left(x + \frac{1}{a}b\right) - \frac{1}{a^2}b^T b + \frac{2}{a}c\right) \\ &= \frac{1}{2}a\left\|x - \left(-\frac{1}{a}b\right)\right\|^2 - \frac{1}{2a}\|b\|^2 + c. \end{aligned}$$

Como $f(\cdot)$ es estrictamente convexa su minimizador es único y para calcularlo basta con hallar el punto donde el vector gradiente se anule.
Por (1.1), tenemos que,

$$\bigtriangledown f(x) = aIx + b.$$

Así que $\bigtriangledown f(x) = 0$ si y solo si $ax + b = 0$.
Por lo tanto el minimizador de $f(\cdot)$ es $x^* = -\frac{1}{a}b$ y su valor mínimo viene dado como sigue

$$f^* = f(x^*) = \frac{1}{2a}\|b\|^2 - \frac{1}{a}\|b\|^2 + c = -\frac{1}{2a}\|b\|^2 + c.$$

En consecuencia

$$f(x) = f^* + \frac{1}{2}a\|x - x^*\|^2$$

■

Teorema 1.5. *(Danskin) (ver [2]) Sea $Z \subset \mathbb{R}^m$ un conjunto compacto y sea $\phi : \mathbb{R}^n \times Z \longrightarrow \mathbb{R}$ continua y tal que $\phi(\cdot, z) : \mathbb{R}^n \longrightarrow \mathbb{R}$ es convexa para cada $z \in Z$.*

(a) *La función $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ dada por*

$$f(x) = \max_{z \in Z} \phi(x, z)$$

es convexa y tiene derivada direccional dada por

$$f'(x; y) = \max_{z \in Z(x)} \phi'(x, z; y),$$

donde $\phi'(x, z; y)$ es la derivada direccional de la función $\phi(\cdot, z)$ para cada x en la dirección y, $Z(x)$ es el conjunto dado por

$$Z(x) = \{\bar{z} : \phi(x, \bar{z}) = \max_{z \in Z} \phi(x, z)\}.$$

En particular, si $Z(x)$ consiste de un único punto $\bar{z}$ tal que $\phi(\cdot, \bar{z})$ es diferenciable en x, entonces f es diferenciable en x y $\bigtriangledown f(x) = \bigtriangledown_x \phi(x, \bar{z})$, donde $\bigtriangledown_x \phi(x, \bar{z})$ es el vector con coordenadas

$$\frac{\partial \phi(x, \bar{z})}{\partial x_i}, \quad i = 1, \ldots, n.$$

(b) *Si $\phi(\cdot, \bar{z})$ es diferenciable para todo $z \in Z$ y $\bigtriangledown_x \phi(x, \cdot)$ es continua en Z para cada x, entonces*

$$\partial f(x) = conv\{\bigtriangledown_x \phi(x, z) : z \in Z(x)\}$$

para todo $x \in \mathbb{R}^n$.
En particular, si ϕ es lineal en x para todo $z \in Z$, es decir,

$$\phi(x, z) = a_z' x + b_z,$$

para todo $z \in Z$, entonces

$$\partial f(x) = conv\{a_z : z \in Z(x)\}.$$

Observación 1.1. Sabemos que si una función $f(\cdot)$ es convexa, entonces $-f(\cdot)$ es cóncava en consecuencia el minimizador de $f(\cdot)$ es el maximizador de $-f(\cdot)$, por lo tanto el teorema 1.5 se puede enunciar de manera análoga para funciones cóncavas.

Definición 1.11. Sea $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ una función diferenciable. Se llama dirección de descenso para $f(\cdot)$ en x a todo vector $d \in \mathbb{R}^n$ tal que $d^T \bigtriangledown f(x) < 0$.

Un ejemplo de dirección de descenso para $f(\cdot)$ en x es $- \bigtriangledown f(x)$, que es la llamada dirección de máximo descenso o dirección de Cauchy.

Las estrategias para la resolución del problema (1) se pueden clasificar en dos tipos; Búsqueda lineal y Regiones de confianza. El tipo de algoritmo que se presentará forma parte de la primera, la cual consiste en la construcción iterativa de puntos x^k partiendo de un punto inicial x^0. Los siguientes iterados vienen dados por $x^{k+1} = x^k + s_k d^k$, donde s_k es el tamaño de paso y d^k es la dirección de búsqueda.

El tamaño de paso $s_k > 0$ debe ser tal que halla una disminución del valor de la función $f(\cdot)$ en el punto x^{k+1} respecto al punto x^k.

Existen también dos formas para hallar el tamaño de paso; la primera consiste en resolver el problema unidimensional

$$\min_{s>0} \phi(s) = f(x^k + sd^k). \tag{1.3}$$

En la literatura, existen varios algoritmos que resuelven el problema (1.3); Newton, Sección Dorada, Bisección, entre otros (ver, por ejemplo [1], [5],). Este tipo de búsqueda puede ser muy costoso, y algunas veces ineficiente, es por ello que usualmente se emplean métodos de "búsqueda imprecisa" o inexacta, que consisten en garantizar cierto decrecimiento de la función objetivo $f(\cdot)$, esto se mide con la siguiente desigualdad

$$f(x^k + sd^k) \leq f(x^k) + c_1 s \bigtriangledown f(x^k)^T d^k, \tag{1.4}$$

para alguna constante $c_1 \in (0, 1)$. En otras palabras, la reducción en $f(\cdot)$ deberá ser proporcional a la longitud de paso s_k y a la derivada direccional $\bigtriangledown f(x^k)^T d^k$. La desigualdad (1.4) es llamada la **condición de Armijo**. En [11] recomiendan $c_1 = 10^{-4}$.

La condición de suficiente decrecimiento de la función objetivo $f(\cdot)$ no asegura que el algoritmo tenga un rasonable progreso, para excluir los pasos cortos no aceptables se introduce una segunda condición, llamada la **condición de curvatura**, donde se requiere que s_k satisfaga la siguiente desigualdad:

$$\bigtriangledown f(x^k + sd^k)^T d^k \geq c_2 \bigtriangledown f(x^k)^T d^k,$$

para alguna constante $c_2 \in (c_1, 1)$, donde c_1 es la constante de (1.4).

La condición de Armijo y la condición de curvatura son llamadas en conjunto las condiciones de Wolfe, ver figura 1.1

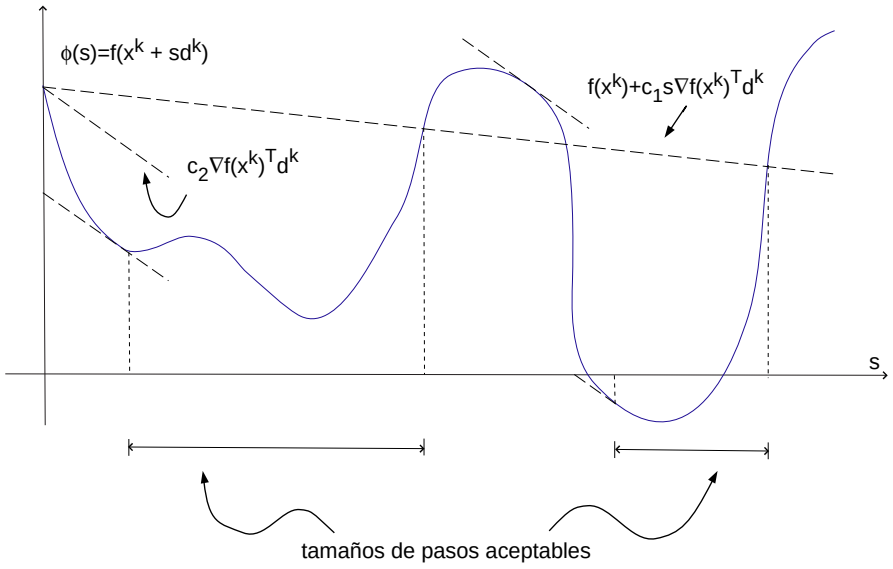

Figura 1.1: Condiciones de Wolfe.

Definición 1.12. Diremos que s_k cumple con la condición de Armijo con parámetro 0 si $f(x^k + s_k d^k) \leq f(x^k)$, y que cumple con la condición de curvatura con parámetro 0 si $\bigtriangledown f(x^k + s_k d^k)^T d^k \geq 0$.

Mas aún diremos que s_k cumple con las condiciones de Wolfe con parámetro 0 si cumple con ambas condiciones expresadas anteriormente, ver figura 1.2.

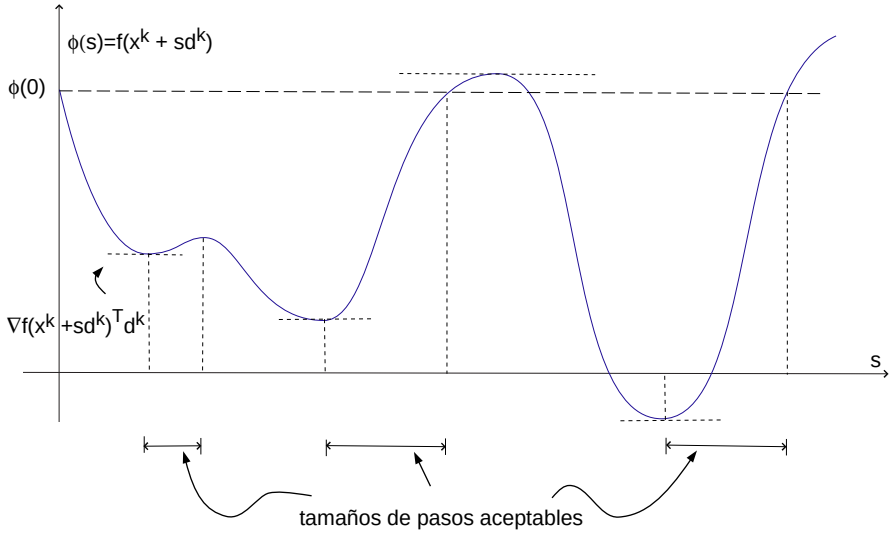

Figura 1.2: Condiciones de Wolfe con parámetro 0

Al igual que las condiciones de Wolfe, las condiciones de Goldstein también aseguran que la longitud de paso s_k logre un suficiente decrecimiento de la función $f(\cdot)$ y al mismo tiempo previene que s_k sea muy pequeña, estas son las siguientes:

$$f(x^k) + (1-c)s \bigtriangledown f(x^k)^T d^k \leq f(x^k + sd^k) \leq f(x^k) + cs \bigtriangledown f(x^k)^T d^k, \qquad (1.5)$$

con $0 < c < \frac{1}{2}$. La segunda desigualdad es la condición de suficiente decrecimiento (1.4), mientras que la primera desigualdad se introduce para evitar que los tamaños de pasos sean muy pequeños, ver figura 1.3.

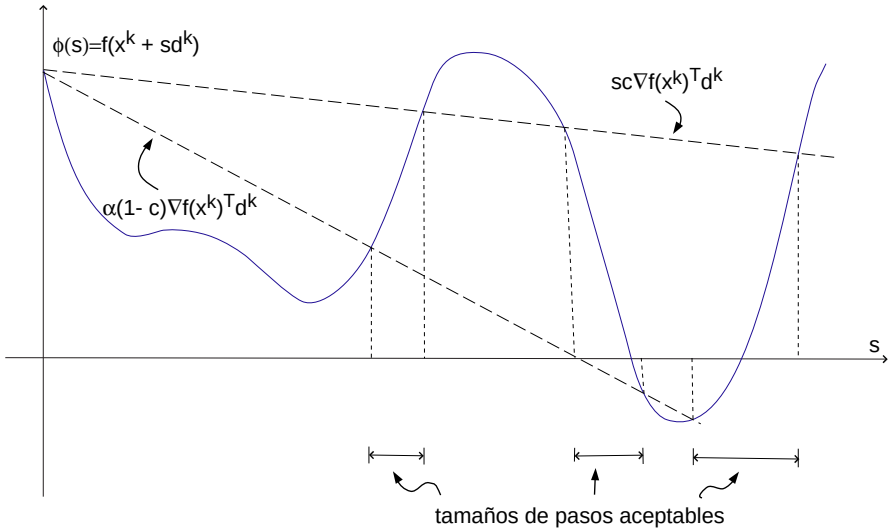

Figura 1.3: Condiciones de Goldstein

Capítulo 2

Algoritmo de Nesterov y variante

En este capitulo se presentaran dos métodos de primer orden para programación convexa, a saber el método de Nesterov y el de Gonzaga y Karas (variante del método de Nesterov) mostrados en [3] los cuales ofrecen una solución optima para el problema de programación no lineal

$$\begin{aligned} &\text{mín} \quad f(x) \\ &s.\,a. \quad x \in \mathbb{R}^n \end{aligned} \tag{2.0.1}$$

donde $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es convexa y continuamente diferenciable, con constante de Lipschitz $L > 0$ para el gradiente de $f(\cdot)$ y con parámetro de convexidad $\mu \geq 0$.

Por la definición 1.1 y el teorema 1.1 tenemos que para todo $x, y \in \mathbb{R}^n$,

$$\|f(x) - f(y)\| \leq L\|x - y\|, \tag{2.0.2}$$

$$f(x) \geq f(y) + \bigtriangledown f(y)^T(x - y) + \frac{1}{2}\mu\|x - y\|^2. \tag{2.0.3}$$

Afirmación 2.0.1. $\mu \leq L$.

En efecto, por (2.0.3) tenemos que para $x, y \in \mathbb{R}^n$ con $x \neq y$,

$$f(x) \geq f(y) + \bigtriangledown f(y)^T(x - y) + \frac{1}{2}\mu\|x - y\|^2,$$

$$f(y) \geq f(x) + \bigtriangledown f(x)^T(y-x) + \frac{1}{2}\mu\|y-x\|^2,$$

Sumando las dos desigualdades anteriores,

$$\mu\|(x-y)\|^2 \leq -\bigtriangledown f(y)^T(x-y) - \bigtriangledown f(x)^T(y-x).$$

Utilizando (2.0.2) y la desigualdad de Cauchy Schwarz, tenemos que,

$$\begin{aligned}
\mu\|(x-y)\|^2 &\leq \bigtriangledown f(x)^T(x-y) - \bigtriangledown f(y)^T(x-y) \\
&= \mid (\bigtriangledown f(x) - \bigtriangledown f(y))^T(x-y) \mid \\
&\leq \|(\bigtriangledown f(x) - \bigtriangledown f(y)\|\|x-y\| \\
&\leq L\|x-y\|^2.
\end{aligned}$$

En consecuencia obtenemos que,

$$\mu \leq L.$$

■

El método de Nesterov usa las propiedades locales de $f(\cdot)$ (las direcciones de máximo descenso) y las propiedades globales de las funciones convexas para generar una sucesión de funcionales acotados superiormente impuestos sobre el epígrafo de $f(\cdot)$. Los funcionales tienen la forma:

$$\phi_k(x) = \phi_k^* + \tfrac{\gamma_k}{2}\|x - v^k\|^2,$$

donde $\phi_k^* \in \mathbb{R}$, $\gamma_k > 0$ y $v^k \in \mathbb{R}^n$. Estas funciones son estrictamente convexas con mínimo ϕ_k^* para v^k. Al inicio $v^0 = x^0$ y $\gamma_0 > 0$ son dados. El método calcula una sucesión de puntos x^k, una sucesión de parámetros positivos $\lambda_k \longrightarrow 0$ y una sucesión de funciones $\phi_k(\cdot)$. La construcción es de tal forma que para cada iteración $\phi_k^* \geq f(x^k)$, para así tener la hipótesis del lema 1.1. En el presente trabajo se requiere que $\phi_k^* = f(x^k)$. Por lo tanto cada solución óptima x^* satisface que $f(x^*) \leq \phi_k^* \leq \phi_k(x^k)$ y así podemos pensar en $\phi_k(\cdot)$ como una cota superior, impuesta para el epígrafo de $f(\cdot)$. Esto se muestra en la Figura 2.1.

Las constantes de segundo orden se definen como $\gamma_k - \mu = \lambda_k(\gamma_0 - \mu)$ y la sucesión (λ_k) se define por recursión como:
$\lambda_0 = 1$,
$\lambda_{k+1} = (1 - \alpha_k)\lambda_k$
con $\alpha_k \in (0, 1)$ y $\sum_{k=0}^{\infty} \alpha_k = \infty$.

Así cuando k crece las funciones $\phi_k(\cdot)$ se convierten en planas como se demuestra en la figura 2.1.

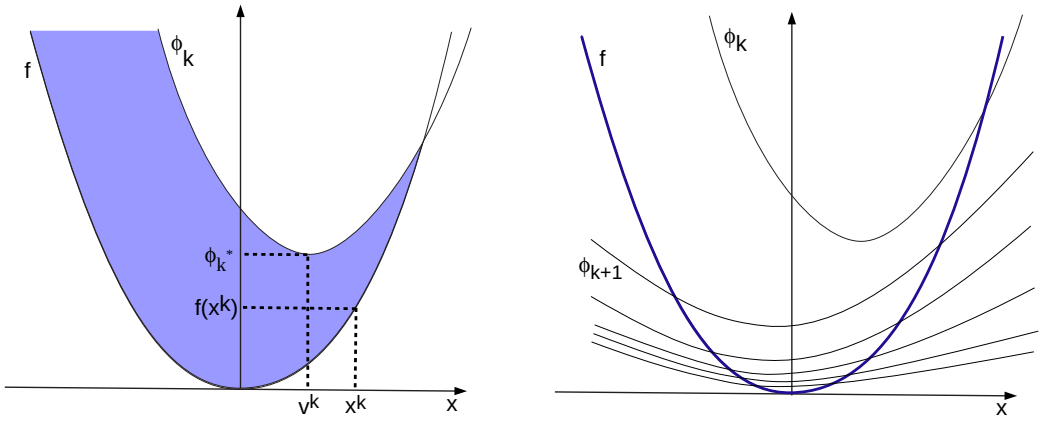

Figura 2.1: Mecánica del Algoritmo de Nesterov

Nesterov muestra como definir estas funciones de modo que

$$f(x^k) \leq \phi_k(\cdot) \leq \lambda_k\phi_0(\cdot) + (1 - \lambda_k)f(\cdot),$$

es decir, que el par de sucesiones $(\phi_k(\cdot))$ y (λ_k) sea una sucesión estimada de $f(\cdot)$. Una examinación inmediata de la solución óptima muestra que

$$f(x^k) \leq \phi_k(x^*) \leq \lambda_k\phi_0(x^*) + (1 - \lambda_k)f(x^*) = f(x^*) + \lambda_k(\phi_0(x^*) - f(x^*)),$$

observando que $f(x^k) \longrightarrow f(x^*)$ cuando $\lambda_k \longrightarrow 0$ (este resultado se muestra en el lema 1.1). El método de Nesterov asegura que $\lambda_k = O(1/k^2)$, es decir un error de $\epsilon > 0$ es guardado en $O(1/\sqrt{\epsilon})$ iteraciones (ver [8]).

La eficiencia práctica de este método depende de la cantidad de veces que en cada iteración se examinen las propiedades locales de $f(\cdot)$ alrededor de cada iterado x^k cuando λ_k disminuya. Por supuesto la complejidad en el peor de los casos no se puede mejorar pero la velocidad se puede mejorar en los problemas prácticos, este es el objetivo de este trabajo.

A continuación se presenta un algoritmo prototipo que incluye la elección de dos parámetros (α_k y θ_k) en cada iteración: diferentes elecciones de estos parámetros dan diferentes algoritmos. Hay varias formas de elegir estos parámetros. Nesterov tiene una regla fija para la elección de estos, basada en la constante de Lipschitz L. En el método propuesto por Gonzaga y Karas [3] sólo la elección de θ_k es necesaria mientras que α_k está determinada por la solución de una ecuación de segundo grado.
Se asume que el parámetro de convexidad μ es dado (posiblemente nulo).

Algoritmo 2.0.1. (Algoritmo prototipo)
Datos: $x^0 \in \mathbb{R}^n$, $v^0 = x^0$, $\gamma_0 > \mu$ ($\gamma_0 = L$ si L is conocida)
$k = 0$
Repetir
$d^k = v^k - x^k$
Elegir $\theta_k \in [0, 1]$.
$y^k = x^k + \theta_k d^k$.
Si $\bigtriangledown f(y^k) = 0$, entonces pare, tomar y^k como una solución optima.
Paso de máximo descenso: $x^{k+1} = y^k - \nu \bigtriangledown f(y^k)$.
Elegir $\alpha_k \in (0, 1]$.
$\gamma_{k+1} = (1 - \alpha_k)\gamma_k + \alpha_k \mu$.
$v^{k+1} = \frac{1}{\gamma_{k+1}}((1 - \alpha_k)\gamma_k v^k + \alpha_k(\mu y^k - \bigtriangledown f(y^k)))$.
$k = k + 1$.
Se puede usar cualquier procedimiento de búsqueda lineal en el paso de máximo descenso.

Un resultado importante para funciones continuamente diferenciables con gradiente Lipschitz se muestra a continuación.

Teorema 2.0.1. *Sea $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ una función continuamente diferenciable con constante de Lipschitz $L > 0$ para el gradiente de $f(\cdot)$, entonces para todo $x, y \in \mathbb{R}^n$ se cumple la siguiente desigualdad*

$$f(y) - f(x) - \bigtriangledown f(x)^T(y - x) \leq \frac{L}{2}\|x - y\|^2$$

Prueba:
como $f(\cdot)$ es continuamente diferenciable, por el teorema de Taylor de orden 0, tenemos que, para todo $x, y \in \mathbb{R}^n$ se cumple lo siguiente:

$$f(y) = f(x) + \int_0^1 (\bigtriangledown f(x + \tau(y - x))^T(y - x)d_\tau. \tag{2.0.4}$$

Sumando y restando $\bigtriangledown f(x)^T(y - x)$ en (2.0.4), tenemos que,

$$f(y) = f(x) + \bigtriangledown f(x)^T(y - x) + \int_0^1 (\bigtriangledown f(x + \tau(y - x)) - \bigtriangledown f(x))^T(y - x)d_\tau.$$

Por la desigualdad de Cauchy Schwarz y (2.0.2) obtenemos que,

$$\begin{aligned}
f(y) - f(x) - \bigtriangledown f(x)^T(y - x) &= \int_0^1 (\bigtriangledown f(x + \tau(y - x) - \bigtriangledown f(x))^T(y - x)d_\tau \\
&\leq \left| \int_0^1 (\bigtriangledown f(x + \tau(y - x) - \bigtriangledown f(x))^T(y - x)d_\tau \right| \\
&\leq \int_0^1 \left| (\bigtriangledown f(x + \tau(y - x) - \bigtriangledown f(x))^T(y - x) \right| d_\tau \\
&\leq \int_0^1 \| \bigtriangledown f(x + \tau(y - x)) - \bigtriangledown f(x)\| \|y - x\| d_\tau \\
&\leq \int_0^1 L\|x + \tau(y - x) - x\| \|y - x\| d_\tau \\
&= \int_0^1 \tau L\|x - y\|^2 d_\tau \\
&= \frac{L}{2}\|y - x\|^2.
\end{aligned}$$

■

Si L se conoce entonces la longitud de paso (o tamaño de paso) $\nu = 1/L$ asegura que,

$$f(y^k) - f(x^{k+1}) \geq \| \bigtriangledown f(y^k)\|^2/2L.$$

En efecto, sustituyendo $y = x^{k+1}$, $x = y^k$ en el resultado anterior tenemos que,

$$f(x^{k+1}) - f(y^k) + \frac{1}{L}\| \bigtriangledown f(y^k)\|^2 \leq \frac{1}{2L}\| \bigtriangledown f(y^k)\|^2.$$

Por lo tanto,

$$f(y^k) - f(x^{k+1}) \geq \frac{1}{2L}\| \bigtriangledown f(y^k)\|^2,$$

completando la prueba.

■

Es necesario que la búsqueda lineal utilizada sea tan buena como ésta, cada vez que se conozca L. Por supuesto una buena búsqueda lineal se haría, pero esto no es factible en la práctica.

Una búsqueda de Goldstein con una buena elección en el parámetro es usualmente una buena elección, e incluso si L no se conoce éste asegura una disminución de, al menos $\| \bigtriangledown f(y^k)\|^2/4L$, la cual por demás aceptable para el análisis de complejidad.

En efecto, por la primera desigualdad de Goldstein dada en (1.5) tenemos que,

$$f(y^k) - f(x^{k+1}) \leq (1-c)\nu\| \bigtriangledown f(y^k)\|^2.$$

Aplicando el teorema 2.0.1 con $x = y^k$ y $y = x^{k+1}$ obtenemos que,

$$\begin{aligned} f(y^k) - f(x^{k+1}) &\geq \bigtriangledown f(y^k)^T(y^k - x^{k+1}) - \frac{L}{2}\|y^k - x^{k+1}\|^2 \\ &\geq \nu\| \bigtriangledown f(y^k)\|^2 - \nu^2\frac{L}{2}\| \bigtriangledown f(y^k)\|^2 \\ &= \nu\left(1 - \nu\frac{L}{2}\right)\| \bigtriangledown f(y^k)\|^2. \end{aligned}$$

Por lo tanto,

$$\nu \geq \frac{2}{L}c.$$

De la segunda desigualdad de Goldstein dada en (1.5) tenemos que,

$$f(y^k) - f(x^{k+1}) \geq c\nu \| \bigtriangledown f(y^k)\|^2.$$

Combinando esta desigualdad con la previa, concluimos que

$$f(y^k) - f(x^{k+1}) \geq \frac{2}{L} c^2 \| \bigtriangledown f(y^k)\|^2.$$

De acá se puede observar que tomando $c \geq \frac{1}{\sqrt{8}}$ se obtiene lo deseado.

■

Por lo tanto supongamos que en todas las iteraciones

$$f(y^k) - f(x^{k+1}) \geq \frac{1}{4L} \| \bigtriangledown f(y^k)\|^2. \tag{2.0.5}$$

2.1. Elección de los parámetros

Para elegir los parámetros (α_k y θ_k) en el método propuesto por Nesterov se requiere del conocimiento de la constante de Lipschitz $L > 0$ para el gradiente de $f(\cdot)$. En vista de que la mayoría de veces se hace difícil calcular tal constante, Gonzaga y Karas (ver [3]) proponen una forma diferente para la elección de éstos sin utilizarla. La elección de los α_k y los θ_k marcan la diferencia entre los dos algoritmos.

2.1.1. Algoritmo de Nesterov:

Elección de θ_k: calcular α_N como la solución positiva de la ecuación

$$2L\alpha^2 - (1-\alpha)\gamma_k - \alpha\mu = 0. \tag{2.1.1}$$

Calcular

$$\theta_k = \frac{\gamma_k}{\gamma_k + \alpha_N \mu} \alpha_N. \tag{2.1.2}$$

Nota 2.1.1. La justificación de esta elección se muestra en el lema 3.0.1.

2.1.2. Algoritmo propuesto por Gonzaga y Karas

Elección de θ_k: Calcular $\theta_k \in [0,1]$ tal que $f(x^k + \theta_k d^k) \leq f(x^k)$ y $\theta_k = 1$ o $f'(x^k + \theta_k d^k, d^k) \geq 0$. (Se trata de una condición de Armijo con parámetro 0 y una condición de curvatura con parámetro 0.)

Elección de α_k: Calcular $\alpha_k \in [0,1]$ como la mayor raíz de la ecuación

$$A\alpha^2 + B\alpha + C = 0 \tag{2.1.3}$$

con

$$\begin{aligned} Q &= \gamma_k \left(\frac{\mu}{2} \|v^k - y^k\|^2 + \bigtriangledown f(y^k)^T (v^k - y^k) \right), \\ A &= Q + \frac{1}{2} \| \bigtriangledown f(y^k)\|^2 + (\mu - \gamma_k)(f(x^k) - f(y^k)), \\ B &= (\mu - \gamma_k)(f(x^{k+1}) - f(x^k)) - \gamma_k (f(y^k) - f(x^k)) - Q, \\ C &= \gamma_k (f(x^{k+1}) - f(x^k)). \end{aligned}$$

Esta ecuación siempre tiene una raíz real y su raíz más grande está en el intervalo $[0,1]$. Esto se muestra en el lema 3.0.2

Elección original de α_k: tomar $\alpha_k = \alpha_N$ o modificar la elección de α_k: calculando α_k como la raíz mas grande de (2.1.3). El método de Nesterov elige $\alpha_k = \alpha_N$ y θ_k como en (2.1.2).

Ventajas del Algoritmo de Nesterov:

- Utiliza sólo un cálculo del gradiente.
- No realiza cálculos de la función por iteración.
- Tiene complejidad óptima.

Desventajas del algoritmo de Nesterov:

- No explora la eficiencia local de los pasos de Cauchy, basándose en el conocimiento de la constante de Lipschitz L para calcular α_k. Como veremos, la velocidad del algoritmo depende de cómo γ_k es reducida, lo que implica que en cada iteración el valor de α_k debe ser tan grande como la teoría lo permita.

- La sucesión de los valores de la función $(f(x^k))$ no es monótona decreciente, aunque converge a $f(x^*)$ con complejidad óptima.

El método propuesto por Gonzaga y Karas [3] no depende del conocimiento de la constante de Lipschitz L y calcula α_k de manera que en "cierto sentido" es el más grande posible.

Ventajas del algoritmo de Gonzaga y Karas:

- No depende del conocimiento de L o incluso de la existencia de L, sino hacer uso de ella siempre que esté disponible. Por supuesto, los resultados de la complejidad sólo tienen sentido cuando existe L (aunque sea desconocida).

- La sucesión $(f(x^k))$ disminuye hacia $f(x^*)$ con complejidad óptima.

- Si L es conocida, la adicional búsqueda lineal puede ser reducida limitando el número de cálculos de la función a un número pre-determinado.

2.2. Combinación lineal de funciones cuadráticas simples

El algoritmo se basa en la construcción de una sucesión de funciones cuadráticas simples $\phi_k(\cdot)$. En esta sección nos concentraremos en el estudio de las combinaciones lineales de dos funciones las cuales definiremos a continuación

$$\phi(x) = \phi^* + \frac{\gamma}{2}\|x - v\|^2,$$

$$l(x) = l_0 + g^T(x - y) + \frac{\mu}{2}\|x - y\|^2,$$

donde $\phi^*, l_0, \mu, \gamma \in \mathbb{R}$, $g, y, v \in \mathbb{R}^n$. En el presente trabajo asumimos que $\gamma > \mu \geq 0$. Ambas funciones son cuadráticas simples, por lo tanto tienen hessiano escalar, es decir $\bigtriangledown^2\phi(x) = \gamma I$ y $\bigtriangledown^2 l(x) = \mu I$, pero $l(\cdot)$ puede ser lineal ya que μ puede ser cero.

De (1.1) tenemos,

$$\bigtriangledown l(x) = \mu I x + g - \mu y \tag{2.2.1}$$

para todo $x \in \mathbb{R}^n$.

Si $\mu > 0$ entonces la función $l(\cdot)$ tiene un único minimizador (ya que $l(\cdot)$ sería estrictamente convexa), éste se obtiene calculando el punto que anula el gradiente, el cual esta dado por (2.2.1) como $\hat{x} = y - \frac{1}{\mu}g$.

Evaluando la expresión que define a $l(\cdot)$ en $\hat{x}$, obtenemos,

$$\begin{aligned} l(\hat{x}) &= l\left(y - \frac{1}{\mu}g\right) \\ &= l_0 - \frac{1}{\mu}\|g\|^2 + \frac{1}{2\mu}\|g\|^2 \\ &= l_0 - \frac{1}{2\mu}\|g\|^2. \end{aligned}$$

En consecuencia,

$$\min_{x\in\mathbb{R}^n} l(x) = l_0 - \frac{1}{2\mu}\|g\|^2.$$

Por la Afirmación 1.1,

$$l(x) = l(\hat{x}) + \frac{\mu}{2}\|x - \hat{x}\|^2, \tag{2.2.2}$$

donde $\hat{x}$ es el minimizador de $l(\cdot)$.

Estudiaremos las combinaciones convexas de estas dos funciones, definiendo para $\alpha \in \mathbb{R}$ y $x \in \mathbb{R}^n$

$$\phi_+(\alpha, x) = (1 - \alpha)\phi(x) + \alpha l(x), \tag{2.2.3}$$

$$\phi_+^*(\alpha) = \inf_{x\in\mathbb{R}^n} \phi_+(\alpha, x) \in \mathbb{R} \cup \{-\infty\}, \tag{2.2.4}$$

$$\gamma_+(\alpha) = (1-\alpha)\gamma + \alpha\mu. \tag{2.2.5}$$

Sabemos que para cada $\alpha \in \mathbb{R}$ la función $\phi_+(\alpha, \cdot)$ es también cuadrática simple, mas aún

$$\begin{aligned} \bigtriangledown^2\phi_+(\alpha, x) &= (1-\alpha)\bigtriangledown^2\phi(x) + \alpha\bigtriangledown^2 l(x) \\ &= (1-\alpha)\gamma I + \alpha\mu I \\ &= \gamma_+(\alpha) I. \end{aligned}$$

Si $\gamma_+(\alpha) > 0$ entonces para cada $x \in \mathbb{R}^n$ la matriz $\bigtriangledown^2\phi_+(\alpha, x)$ es definida positiva, por lo tanto $\phi_+^*(\alpha)$ es finito.

Si $\gamma_+(\alpha) < 0$ entonces para cada $x \in \mathbb{R}^n$ la matriz $\bigtriangledown^2\phi_+(\alpha, x)$ es definida negativa, por lo tanto $\phi_+^*(\alpha) = -\infty$.

La siguiente afirmación muestra para que valores de α se obtienen los casos anteriores.

Afirmación 2.2.1. Si $\gamma > \mu \geq 0$, se tiene que,

$$\{\alpha \in \mathbb{R} : \gamma_+(\alpha) > 0\} = (-\infty, \alpha_{\text{máx}}), \tag{2.2.6}$$

con $\alpha_{\text{máx}} = \frac{\gamma}{\gamma-\mu}$.

En efecto, sea $\alpha \in \mathbb{R}$ tal que $\gamma_+(\alpha) > 0$, de la definición de $\gamma_+(\alpha)$ se obtiene que $(1-\alpha)\gamma + \alpha\mu > 0$, luego $\gamma - \alpha(\gamma-\mu) > 0$, en consecuencia $\alpha < \frac{\gamma}{\gamma-\mu}$. Por lo tanto $\alpha \in (-\infty, \alpha_{\text{máx}})$.

Análogamente se prueba que dado $\alpha \in (-\infty, \alpha_{\text{máx}})$, $\gamma_+(\alpha) > 0$, obteniendo (2.2.6).

■

De la afirmación 2.2.1 podemos observar que,

$$\{\alpha \in \mathbb{R} : \gamma_+(\alpha) < 0\} = (\alpha_{\text{máx}}, +\infty).$$

Ahora estudiemos el caso cuando $\alpha = \alpha_{\text{máx}}$.
Si evaluamos la expresión (2.2.5) en $\alpha = \alpha_{\text{máx}}$, tenemos que,

$$\begin{aligned}\gamma_+(\alpha_{\text{máx}}) &= \left(1 - \left(\frac{\gamma}{\gamma-\mu}\right)\right)\gamma + \left(\frac{\gamma}{\gamma-\mu}\right)\mu \\ &= -\frac{\gamma\mu}{\gamma-\mu} + \frac{\gamma\mu}{\gamma-\mu} = 0. \end{aligned} \tag{2.2.7}$$

Desarrollando la expresión que define a $\phi_+(\alpha, x)$ en (2.2.3), tenemos que,

$$\begin{aligned}\phi_+(\alpha,x) &= (1-\alpha)\phi(x) + \alpha l(x) \\ &= (1-\alpha)\left(\phi^* + \frac{\gamma}{2}\|x-v\|^2\right) + \alpha\left(l_0 + g^T(x-y) + \frac{\mu}{2}\|x-y\|^2\right) \\ &= (1-\alpha)\frac{\gamma}{2}x^Tx - (1-\alpha)\gamma v^Tx + (1-\alpha)\left(\phi^* + \frac{\gamma}{2}\|v\|^2\right) + \\ &\quad +\alpha\frac{\mu}{2}x^Tx + \alpha(g-\mu y)^Tx + \alpha\left(l_0 + \frac{\mu}{2}\|y\|^2 - g^Ty\right) \\ &= \frac{1}{2}\gamma_+(\alpha)x^Tx + (\alpha(g-\mu y) - (1-\alpha)\gamma v)^Tx + \\ &\quad +(1-\alpha)\left(\phi^* + \frac{\gamma}{2}\|v\|^2\right) + \alpha\left(l_0 + \frac{\mu}{2}\|y\|^2 - g^Ty\right). \end{aligned} \tag{2.2.8}$$

En consecuencia,

$$\begin{aligned}\phi_+(\alpha_{\text{máx}},x) &= (\alpha_{\text{máx}}(g-\mu y) - (1-\alpha_{\text{máx}})\gamma v)^Tx + (1-\alpha_{\text{máx}})\left(\phi^* + \frac{\gamma}{2}\|v\|^2\right) + \\ &\quad +\alpha_{\text{máx}}\left(l_0 + \frac{\mu}{2}\|y\|^2 - g^Ty\right) \end{aligned} \tag{2.2.9}$$

es una función lineal.

En general, $\phi_+^*(\alpha_{\text{máx}}) = -\infty$, pero puede suceder que $\phi_+(\alpha_{\text{máx}}, x)$ sea constante y por lo tanto $\phi_+^*(\alpha_{\text{máx}}) > -\infty$.

En el siguiente lema estudiaremos esta situación.

Lema 2.2.1. *Consideremos las funciones cuadráticas simples $\phi(\cdot)$ y $l(\cdot)$ definidas anteriormente con $\gamma > \mu$ y supongamos que $l(\cdot)$ no es constante. Si definimos $\alpha_{\text{máx}} = \frac{\gamma}{\gamma-\mu}$ entonces $\phi_+^*(\alpha_{\text{máx}})$ es finito solo si $\mu > 0$ y los minimizadores de $\phi(\cdot)$ y $l(\cdot)$ coinciden. En este caso $\phi_+^*(\cdot)$ es lineal en $(-\infty, \alpha_{\text{máx}})$.*

Prueba:
De (2.2.9) tenemos que $\phi_+(\alpha_{\text{máx}}, x)$ es lineal.
Supongamos que $\phi_+^*(\alpha_{\text{máx}})$ es finito, entonces sustituyendo $\alpha = \alpha_{\text{máx}}$ en (2.2.3) tenemos que,

$$\phi_+(\alpha_{\text{máx}}, x) = (1 - \alpha_{\text{máx}})\phi(x) + \alpha_{\text{máx}} l(x) \tag{2.2.10}$$

es constante.
Supongamos por reducción al absurdo que $\mu = 0$, entonces

$$\alpha_{\text{máx}} = \frac{\gamma}{\gamma - 0} = 1.$$

Por lo tanto,

$$\phi_+(\alpha_{\text{máx}}, x) = l(x).$$

De acá tenemos que $l(\cdot)$ es constante, contradiciendo la hipótesis, en consecuencia $\mu > 0$.
Por (2.2.2) podemos escribir (2.2.3) como sigue

$$\phi_+(\alpha, x) = (1 - \alpha)\left(\phi^* + \frac{\gamma}{2}\|x - v\|^2\right) + \alpha\left(l(\hat{x}) + \frac{\mu}{2}\|x - \hat{x}\|^2\right)$$

donde v y $\hat{x}$ son los minimizadores de $\phi(\cdot)$ y $l(\cdot)$ respectivamente.
Evaluando la expresión anterior en $\alpha = \alpha_{max}$, tenemos que,

$$\phi_+(\alpha_{\text{máx}}, x) = (1 - \alpha_{\text{máx}})\left(\phi^* + \frac{\gamma}{2}\|x - v\|^2\right) + \alpha_{\text{máx}}\left(l(\hat{x}) + \frac{\mu}{2}\|x - \hat{x}\|^2\right).$$

Diferenciando la expresión anterior respecto a x, tenemos que para cada $x \in \mathbb{R}^n$,

$$(1 - \alpha_{\text{máx}})\gamma(x - v) + \alpha_{\text{máx}}\mu(x - \hat{x}) = 0.$$

Evaluando la expresión anterior para $x = v$, obtenemos que,

$$\alpha_{\text{máx}}\mu(v - \hat{x}) = 0.$$

Como $\alpha_{\text{máx}}, \mu > 0$ concluimos que $v = \hat{x}$.

Ahora bien v es el minimizador de $\phi(\cdot)$ y $l(\cdot)$. Así que para cualquier $\alpha \in (-\infty, \alpha_{\text{máx}})$

$$\phi_+^*(\alpha) = (1 - \alpha)\phi^* + \alpha l(v),$$

es una función lineal, completando la prueba.

■

Lema 2.2.2. *Para cada $\alpha \in (-\infty, \alpha_{\text{máx}})$ la función $x \longrightarrow \phi_+(\alpha, x)$ es una función cuadrática simple dada por*

$$\phi_+(\alpha, x) = \phi_+^*(\alpha) + \frac{\gamma_+(\alpha)}{2}\|x - v_+(\alpha)\|^2, \tag{2.2.11}$$

con

$$\phi_+^*(\alpha) = l_0 - \frac{\alpha^2}{2\gamma_+(\alpha)}\|g\|^2 + +(1-\alpha)\Bigg(\phi^* - l_0 + \frac{\alpha\gamma}{\gamma_+(\alpha)}\Bigg(\frac{1}{2}\mu\|y - v\|^2 + \\ + g^T(v - y)\Bigg)\Bigg), \tag{2.2.12}$$

$$v_+(\alpha) = \frac{1}{\gamma_+(\alpha)}((1-\alpha)\gamma v + \alpha(\mu y - g)). \tag{2.2.13}$$

En particular si $\mu > 0$ entonces

$$\phi_+^*(1) = l_0 - \frac{1}{2\mu}\|g\|^2. \tag{2.2.14}$$

Prueba:
Sabemos que para cada $\alpha \in (-\infty, \alpha_{\text{máx}})$, $\phi_+(\alpha, \cdot)$ es cuadrática simple.
Por (1.1) y (2.2.8), tenemos que,

$$\bigtriangledown\phi_+(\alpha, x) = \gamma_+(\alpha)x + \alpha(g - \mu y) - (1-\alpha)\gamma v.$$

Por lo tanto, $\bigtriangledown\phi_+(\alpha, x) = 0$ si y solo si $\gamma_+(\alpha)x + \alpha(g - \mu y) - (1-\alpha)\gamma v = 0$ o lo que es lo mismo si y solo si $x = \frac{1}{\gamma_+(\alpha)}((1-\alpha)\gamma v + \alpha(\mu y - g))$.

Así que para cada $\alpha \in (-\infty, \alpha_{\text{máx}})$, $v_+(\alpha) = \frac{1}{\gamma_+(\alpha)}\ ((1-\alpha)\gamma v + \alpha(\mu y - g))$ es el único minimizador de $\phi_+(\alpha, \cdot)$ ya que $\phi_+(\alpha, \cdot)$ es estrictamente convexa.

Evaluando (2.2.3) en $v_+(\alpha)$, obtenemos,

$$\begin{aligned}
\phi_+(\alpha, v_+(\alpha)) &= \left((1-\alpha)\left(\phi^* + \frac{\gamma}{2}\left\|\frac{1}{\gamma_+(\alpha)}((1-\alpha)\gamma v + \alpha(\mu y - g)) - v\right\|^2\right) + \right. \\
&\quad +\alpha\left(l_0 + g^T\left(\frac{1}{\gamma_+(\alpha)}((1-\alpha)\gamma v + \alpha(\mu y - g)) - y\right) + \right. \\
&\quad \left. +\frac{\mu}{2}\left\|\frac{1}{\gamma_+(\alpha)}((1-\alpha)\gamma v + \alpha(\mu y - g)) - y\right\|^2\right) \\
&= (1-\alpha)\phi^* + (1-\alpha)\frac{\gamma}{2}\left\|\frac{(1-\alpha)\gamma v + \alpha(\mu y - g) - \gamma_+(\alpha)v}{\gamma_+(\alpha)}\right\|^2 + \\
&\quad +\alpha l_0 + \alpha g^T\left(\frac{(1-\alpha)\gamma v + \alpha(\mu y - g) - \gamma_+(\alpha)y}{\gamma_+(\alpha)}\right) \\
&\quad +\alpha\frac{\mu}{2}\left\|\frac{(1-\alpha)\gamma v + \alpha(\mu y - g) - \gamma_+(\alpha)y}{\gamma_+(\alpha)}\right\|^2 \\
&= (1-\alpha)\phi^* + \frac{(1-\alpha)\alpha^2\gamma}{2\gamma_+(\alpha)^2}\|\mu(y-v) - g\|^2 + \alpha l_0 + \\
&\quad +\frac{\alpha}{\gamma_+(\alpha)}g^T\left((1-\alpha)\gamma(v-y) - \alpha g\right) + \\
&\quad +\frac{\alpha\mu}{2\gamma_+(\alpha)^2}\|(1-\alpha)\gamma(v-y) - \alpha g\|^2. \qquad (2.2.15)
\end{aligned}$$

Tenemos que,

$$\begin{aligned}
\|\mu(y-v) - g\|^2 &= (\mu(y-v) - g)^T(\mu(y-v) - g) \\
&= \mu^2\|y-v\|^2 - 2\mu g^T(v-y) + \|g\|^2. \qquad (2.2.16)
\end{aligned}$$

Por otro lado,

$$\begin{aligned}
\|(1-\alpha)\gamma(v-y) - \alpha g\|^2 &= ((1-\alpha)\gamma(v-y) - \alpha g)^T((1-\alpha)\gamma(v-y) - \alpha g) \\
&= (1-\alpha)^2\gamma^2\|v-y\|^2 - 2(1-\alpha)\alpha\gamma g^T(v-y) + \\
&\quad +\alpha^2\|g\|^2. \qquad (2.2.17)
\end{aligned}$$

Sustituyendo (2.2.16) y (2.2.17) en (2.2.15) y aplicando operaciones, tenemos,

$$\begin{aligned}
& \phi_+(\alpha, v_+(\alpha)) = \\
= \;& (1-\alpha)\phi^* + \frac{(1-\alpha)\alpha^2\gamma}{2\gamma_+(\alpha)^2}(\mu^2\|y-v\|^2 + 2\mu g^T(v-y) + \|g\|^2) + \alpha l_0 + \\
& + \frac{\alpha}{\gamma_+(\alpha)} g^T((1-\alpha)\gamma(v-y) - \alpha g) + \frac{\alpha\mu}{2\gamma_+(\alpha)^2}((1-\alpha)^2\gamma^2\|v-y\|^2 - \\
& -2(1-\alpha)\alpha\gamma g^T(v-y) + \alpha^2\|g\|^2) \\
= \;& (1-\alpha)\phi^* + \frac{(1-\alpha)\alpha^2\gamma\mu^2}{2\gamma_+(\alpha)^2}\|y-v\|^2 + \frac{(1-\alpha)\alpha^2\gamma\mu}{\gamma_+(\alpha)^2} g^T(v-y) + \\
& + \frac{(1-\alpha)\gamma\alpha^2}{2\gamma_+(\alpha)^2}\|g\|^2 + \alpha l_0 + \frac{(1-\alpha)\alpha\gamma}{\gamma_+(\alpha)} g^T(v-y) - \frac{\alpha^2}{\gamma_+(\alpha)}\|g\|^2 + \\
& + \frac{(1-\alpha)^2\alpha\gamma^2\mu}{2\gamma_+(\alpha)^2}\|y-v\|^2 - \frac{(1-\alpha)\alpha^2\gamma\mu}{\gamma_+(\alpha)^2} g^T(v-y) + \frac{\alpha^3\mu}{2\gamma_+(\alpha)^2}\|g\|^2 \\
= \;& (1-\alpha)\phi^* + \frac{(1-\alpha)\alpha\gamma\mu}{2\gamma_+(\alpha)}\|y-v\|^2 + \frac{(1-\alpha)\alpha\gamma}{\gamma_+(\alpha)} g^T(v-y) - \frac{\alpha^2}{2\gamma_+(\alpha)} + \alpha l_0.
\end{aligned}$$

Sumando y restando l_0 en la expresión anterior, tenemos,

$$\begin{aligned}
\phi_+(\alpha, v_+(\alpha)) \;=\; & l_0 - \frac{\alpha^2}{2\gamma_+(\alpha)}\|g\|^2 + (1-\alpha)\Bigg(\phi^* - l_0 + \\
& \frac{\alpha\gamma}{\gamma_+(\alpha)}\left(\frac{\mu}{2}\|y-v\|^2 + g^T(v-y)\right)\Bigg).
\end{aligned}$$

De la afirmación 1.1 se tiene (2.2.11).

Finalmente (2.2.14) se obtiene de (2.2.12) con $\alpha = 1$ y $\gamma_+(\alpha) = \mu > 0$, completando la prueba.

■

Ahora usaremos el teorema 1.5 (Danskin) para probar que $\phi_+^*(\cdot)$ es cóncava.

Lema 2.2.3. *La función $\alpha \in \mathbb{R} \longrightarrow \phi_+^*(\alpha)$ es cóncava en $\mathbb{R}$ y diferenciable en $(-\infty, \alpha_{\text{máx}})$*

Prueba:

Sabemos que $\phi_+^*(\alpha) = -\infty$ para $\alpha > \alpha_{\text{máx}}$, pero $\phi_+^*(\alpha_{\text{máx}})$ puede ser infinito ó

finito, por tal razón se nos presentan los siguientes casos:

Caso 1: $\phi_+^*(\alpha_{\text{máx}}) > -\infty$.
Por el lema 2.2.1 $\phi_+^*(\cdot)$ es lineal en $(-\infty, \alpha_{\text{máx}}]$, por lo tanto diferenciable en $(-\infty, \alpha_{\text{máx}})$ y cóncava en $(-\infty, \alpha_{\text{máx}}]$, por el comentario hecho al inicio, se concluye que $\phi_+^*(\cdot)$ es cóncava en $\mathbb{R}$, completando la prueba.

Caso 2: $\phi_+^*(\alpha_{\text{máx}}) = -\infty$.
Sea $[\alpha_1, \alpha_2] \subset (-\infty, \alpha_{\text{máx}})$ un intervalo arbitrario. Por el lema 2.2.2, para cada $\alpha \in [\alpha_1, \alpha_2]$, $\phi_+(\alpha, \cdot)$ tiene un único minimizador $v_+(\alpha)$. Por la continuidad de $v_+(\cdot)$, $v_+([\alpha_1, \alpha_2])$ es un conjunto compacto, por lo tanto cerrado y acotado, es decir, existe $V = v_+([\alpha_1, \alpha_2])$ tal que

$$\phi_+^*(\alpha) = \min_{x \in V} \phi_+(\alpha, x).$$

Por lo tanto se cumplen las hipótesis del teorema 1.5:

- $\phi_+(\cdot, x)$ es cóncava (mas aun lineal) para cada $x \in V$.
- $\phi_+(\alpha, \cdot)$ tiene un único minimizador en el conjunto compacto V para cada $\alpha \in [\alpha_1, \alpha_2]$.

Concluimos por dicho teorema que $\phi_+^*(\alpha)$ es cóncava y diferenciable en $[\alpha_1, \alpha_2]$. Como $[\alpha_1, \alpha_2]$ es arbitrario, se demuestra la concavidad y diferenciabilidad en $(-\infty, \alpha_{\text{máx}})$. Como $\phi_+^*(\alpha) = -\infty$ para $\alpha \geq \alpha_{\text{máx}}$ entonces $\phi_+^*(\cdot)$ es cóncava en $\mathbb{R}$, completando la prueba.

■

Ahora podemos estudiar la ecuación

$$\phi_+^*(\alpha) = P$$

para un determinado $P \in \mathbb{R}$. Estamos interesados en el caso en que $\phi^* > \inf_{x \in \mathbb{R}^n} l(x)$ y $l(\cdot)$ no sea constante.

Lema 2.2.4. *Supongamos que* $\phi^* \geq \inf_{x\in\mathbb{R}^n} l(x)$. *Dada una constante* $P \geq \inf_{x\in\mathbb{R}^n} l(x)$ *entonces ó* $\phi_+^*(\alpha) < P$ *para todo* $\alpha \in [0,1]$ *o* $\phi_+^*(\alpha) = P$ *tiene una o dos raíces reales y la mayor raíz pertenece al intervalo* $[0,1]$.
Las raíces se calculan resolviendo la ecuación de segundo grado

$$A\alpha^2 + B\alpha + C = 0 \tag{2.2.18}$$

con

$$\begin{aligned} Q &= \gamma\left(\frac{\mu}{2}\|v-y\|^2 + g^T(v-y)\right), \\ A &= Q + \frac{1}{2}\|g\|^2 + (\mu-\gamma)(\phi^* - l_0), \\ B &= (\mu-\gamma)(P-\phi^*) - \gamma(l_0-\phi^*) - Q, \\ C &= \gamma(P-\phi^*). \end{aligned}$$

Prueba:
Si $\phi_+^*(\alpha) < P$ para todo $\alpha \in [0,1]$ entonces no hay soluciones para $\phi_+^*(\alpha) = P$. Ahora supongamos que $\phi_+^*(\alpha) \geq P$ para algún $\alpha \in [0,1]$. Como $\phi_+^*(\cdot)$ es cóncava y continua en [0,1) existe $\phi_+^*(\bar{\alpha}) = \max_{\alpha\in[0,1]} \phi_+^*(\alpha)$.
Si suponemos por reducción al absurdo que

$$\phi_+^*(1) = \phi_+^*(\bar{\alpha})$$

entonces

$$\phi_+^*(1) \geq \phi_+^*(0). \tag{2.2.19}$$

Por hipótesis, tenemos,

$$\phi_+^*(0) = \phi^* \geq \inf_{x\in\mathbb{R}^n} l(x) = \phi_+^*(1). \tag{2.2.20}$$

De (2.2.19) y (2.2.20), tenemos que,

$$\phi_+^*(0) = \phi_+^*(1).$$

Como $\phi_+^*(\cdot)$ es cóncava en $[0,1]$, tenemos que $\phi_+^*(\cdot)$ es constante en $[0,1]$. Contradiciendo (2.2.12). Por lo tanto

$$\phi_+^*(1) < \phi_+^*(\bar{\alpha}),$$

y en consecuencia se tiene que $\bar{\alpha} \in [0, 1)$.
Ya que $\phi_+^*(1) < \phi_+^*(\bar{\alpha})$, $P \geq \inf_{x \in \mathbb{R}^n} l(x) = \phi_+^*(1)$ y $\phi_+^*(\alpha) > P$ para algún $\alpha \in [0, 1]$, por el teorema del valor intermedio, existe al menos un $\alpha' \in [\bar{\alpha}, 1) \subset [0, 1]$ tal que $\phi_+^*(\alpha) = P$ (el cual debe ser único ya que $\phi_+^*(\cdot)$ es cóncava y disminuye de $\bar{\alpha}$ a 1 en $[\bar{\alpha}, +\infty)$).
Sustituyendo $\phi_+^*(\alpha) = P$ en (2.2.12), tenemos que,

$$P = l_0 - \frac{\alpha^2}{2\gamma_+(\alpha)}\|g\|^2 + (1-\alpha)\left(\phi^* - l_0 + \frac{\alpha\gamma}{\gamma_+(\alpha)}\left(\frac{\mu}{2}\|y-v\|^2 + g^T(v-y)\right)\right).$$

Multiplicando en ambos lados de la ecuación anterior por $\gamma_+(\alpha)$ (ver, 2.2.5) y efectuando operaciones, obtenemos,

$$\begin{aligned}
0 &= \gamma_+(\alpha)P + \frac{1}{2}\|g\|^2\alpha^2 - \gamma_+(\alpha)\phi^* + \gamma_+(\alpha)\phi^*\alpha - \gamma_+(\alpha)l_0\alpha - \\
&\quad -\gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha + \gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha^2 \\
0 &= \gamma P(1-\alpha) + \mu P\alpha + \frac{1}{2}\|g\|^2\alpha^2 - \gamma\phi^*(1-\alpha) - \mu\phi^*\alpha + \gamma\phi^*(1-\alpha)\alpha + \\
&\quad +\mu\phi^*\alpha^2 - \gamma l_0(1-\alpha)\alpha - \mu l_0\alpha^2 - \gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha + \\
&\quad +\gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha^2 \\
0 &= \gamma P - \gamma P\alpha + \mu P\alpha + \frac{1}{2}\|g\|^2\alpha^2 - \gamma\phi^* + \gamma\phi^*\alpha - \mu\phi^*\alpha + \gamma\phi^*\alpha - \gamma\phi^*\alpha^2 + \\
&\quad +\mu\phi^*\alpha^2 - \gamma l_0\alpha + \gamma l_0\alpha^2 - \mu l_0\alpha^2 - \gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha + \\
&\quad +\gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\alpha^2 \\
0 &= \left(\gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right) + \frac{1}{2}\|g\|^2 + (\mu-\gamma)(\phi^* - l_0)\right)\alpha^2 + \\
&\quad + \left((\mu-\gamma)(P-\phi^*) - \gamma(l_0-\phi^*) - \gamma\left(\frac{1}{2}\mu\|y-v\|^2 + g^T(v-y)\right)\right)\alpha + \\
&\quad +\gamma(P-\phi^*),
\end{aligned}$$

la cual es la ecuación de segundo grado definida en (2.2.18) cuya mayor raíz debe ser α', completando la prueba.

■

Capítulo 3

Análisis de los Algoritmos

Presentados los dos métodos (Nesterov y Variante) se procederá a realizar un estudio de estos, específicamente en los parámetros α_k y θ_k con el objetivo de probar la buena definición del algoritmo propuesto por Gonzaga y Karas (variante) y analizar el orden de complejidad del mismo (ver, [3]).

Los algoritmos propuestos por Nesterov y Gonzaga y Karas generan una sucesión de puntos $(x_k)_{k=0}^{\infty}$ y una sucesión de funciones $(\phi_k(\cdot))_{k=0}^{\infty}$. Cada iteración construye $\phi_{k+1}(\cdot)$ de tal forma que

$$\phi_{k+1}(\cdot) \leq (1-\alpha_k)\phi_k(\cdot) + \alpha_k f(\cdot).$$

$\phi_k(\cdot)$ es una función cuadrática simple con hessiano $\gamma_k I$ y

$$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k \mu.$$

Por lo tanto γ_k converge a μ con la misma velocidad que $\phi_k(x^*)$ converge a $f(x^*)$, donde x^* es una solución óptima.

Los dos objetivos principales de este trabajo son:

- Demostrar que el algoritmo está bien definido, es decir, que podemos calcular α_k tal que $f(x^{k+1}) = \min\limits_{x\in\mathbb{R}^n} \phi_{k+1}(\alpha_k, x) = \phi_{k+1}^*(\alpha_k)$.

- Demostrar que α_k es grande, es decir, $\alpha_k = \Omega(\sqrt{\gamma_{k+1}})$ lo cual asegura complejidad óptima.

La geometría de una iteración del algoritmo propuesto por Gonzaga y Karas se muestra en la figura 3.1: la figura de la izquierda muestra los puntos presentes en una iteración de un problema de dimensión 2. La figura de la derecha muestra las funciones que participan en una iteración de un problema de dimensión 1.

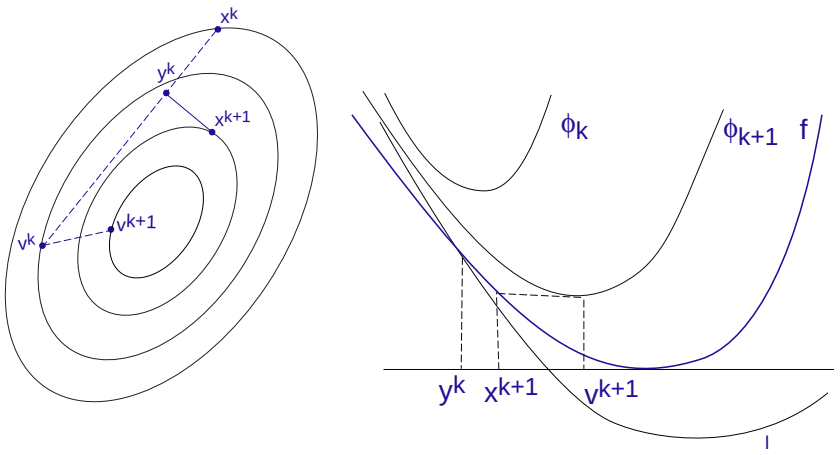

Figura 3.1: Una iteración del algoritmo propuesto por Gonzaga y Karas.

Ahora se describen las variables y funciones asociadas a cada iteración $k = 0, 1, ...$ del algoritmo de Gonzaga y Karas.

- $x^k, v^k \in \mathbb{R}^n$ con x^0 dado y $v^0 = x^0$.
- $\alpha_k \in [0, 1)$.
- $\gamma_k \in \mathbb{R}_{++}$: constantes de segundo orden. $\gamma_0 > \mu$ es dado y debemos tomar $\gamma_0 = L$ cuando L está disponible.
- $y^k \in \mathbb{R}^n$: Un punto en el segmento que une x^k y v^k,

$$y^k = x^k + \theta_k(v^k - x^k).$$

Es apartir de y^k que un paso de máximo descenso se calcula en cada iteración para obtener x^{k+1}. En este trabajo discutiremos con detalle el cálculo de θ_k.

- $\phi_k(\cdot)$: función definida para $x \in \mathbb{R}^n$ por

$$\phi_k(x) = \phi_k^* + \frac{\gamma_k}{2}\|x - v^k\|^2, \tag{3.0.1}$$

con $\phi_0^* = f(x^0)$. Por construcción $\phi_k(\cdot)$ es una función cuadrática simple que asume el mínimo valor ϕ_k^* para v^k. Estas son funciones discutidas al inicio del capítulo, cuya construcción y propiedades se describen a partir de ahora. Nesterov construye estas funciones con el fin de que $\phi_k^* \geq f(x^k)$. En este trabajo se desea que $\phi_k^* = f(x^k)$ para todas las iteraciones.

- $x^{k+1} = y^k - \nu \bigtriangledown f(y^k)$: siguiente iteración, calculada mediante una búsqueda lineal. Como comentamos anteriormente, una buena búsqueda lineal asegura que,

$$f(x^{k+1}) \leq f(y^k) - \frac{1}{4L}\| \bigtriangledown f(y^k)\|^2. \tag{3.0.2}$$

Construcción de las funciones $\phi_k(\cdot)$. Para describir la construcción de las funciones $\phi_k(\cdot)$. Iniciaremos de $x^k, v^k, \phi_k(\cdot), \gamma_k$, los cuales son dados, por lo tanto y^k también es dado (su cálculo será la tarea de este trabajo). Por otro lado supongamos que x^{k+1} se ha calculado y que (3.0.2) se cumple. Una vez que y^k y x^{k+1} son dados, debemos calcular α_k.
La función $\phi_{k+1}(\cdot,\cdot)$ se obtiene mediante la combinación convexa de $\phi_k(\cdot)$ y una aproximación cuadrática inferior de $f(\cdot)$ alrededor de y^k:

$$\phi_{k+1}(\alpha, x) = \phi_{k+1}(x) = (1-\alpha)\phi_k(x) + \alpha l_k(x), \text{ donde} \tag{3.0.3}$$

$$l_k(x) = f(y^k) + \bigtriangledown f(y^k)^T(x - y^k) + \frac{\mu}{2}\|x - y^k\|^2. \tag{3.0.4}$$

Observemos que $\phi_{k+1}(\cdot)$ es una combinación lineal de dos funciones cuadráticas simples cuya construcción proviene del lema 1.2.

Supongamos que en todas las iteraciones $l(\cdot)$ no es constante, de lo contrario, tenemos que,

$$\mu x + \bigtriangledown f(y^k) - \mu y^k = \bigtriangledown l(x) = 0$$

para todo $x \in \mathbb{R}^n$ en particular para $x = y^k$, esto es $\bigtriangledown f(y^k) = 0$, por lo tanto y^k será una solución óptima del problema (2.0.1).
Sea

$$\phi^*_{k+1}(\alpha) = \text{mín}\{\phi_{k+1}(\alpha, x) : x \in \mathbb{R}^n\}$$

para $\alpha \in (-\infty, \alpha_{\text{máx}})$.

La elección de α debe ser tal que $\phi^*_{k+1}(\alpha) = f(x^{k+1})$: tenemos que demostrar en qué condiciones esto es factible.

Obtenemos directamente del lema 2.2.2 con $y = y^k$, $v = v^k$, $\gamma = \gamma_k$, $\gamma_+(\alpha) = \gamma_{k+1}$, $l_0 = f(y^k)$, $g = \bigtriangledown f(y^k)$ lo siguiente:

$$v^{k+1}(\alpha) = \frac{(1-\alpha)\gamma_k v^k + \alpha(\mu y^k - \bigtriangledown f(y^k))}{\gamma_{k+1}}, \tag{3.0.5}$$

$$\begin{aligned}\phi^*_{k+1}(\alpha) &= f(y^k) - \frac{\alpha^2}{2\gamma_{k+1}} \| \bigtriangledown f(y^k)\|^2 + (1-\alpha)\Bigg(\phi^*_k - f(y^k) + \\ &\quad + \frac{\alpha\gamma_k}{\gamma_{k+1}}\left(\frac{\mu}{2}\|y^k - v^k\|^2 + \bigtriangledown f(y^k)^T(v^k - y^k)\right)\Bigg).\end{aligned} \tag{3.0.6}$$

Tomando $y^k = x^k + \theta_k d^k$ con $d^k = v^k - x^k$ y $\theta_k \in [0,1]$ esto puede ser escrito como

$$\phi^*_{k+1}(\alpha) = f(y^k) - \frac{\alpha^2}{2\gamma_{k+1}} \| \bigtriangledown f(y^k)\|^2 + (1-\alpha)\zeta(\alpha, \theta_k) \tag{3.0.7}$$

con

$$\begin{aligned}\zeta(\alpha, \theta_k) &= \phi^*_k - f(x^k + \theta_k d^k) + \frac{\alpha\gamma_k}{\gamma_{k+1}}\Bigg(\frac{\mu}{2}\|x^k + \theta_k d^k - v^k\|^2 + \\ &\quad + \bigtriangledown f(x^k + \theta_k d^k)(v^k - x^k - \theta_k d^k)\Bigg) \\ &= \phi^*_k - f(x^k + \theta_k d^k) + \frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}\Bigg(\frac{\mu}{2}(1-\theta_k)^2\|d^k\|^2 + \\ &\quad + (1-\theta_k) f'(x^k + \theta_k d^k)\Bigg).\end{aligned} \tag{3.0.8}$$

Lema 3.0.1. *Supongamos que* $f(x^k) \leq \phi_k^*$, $y^k = x^k + \theta_k d^k$ *con* $d^k = v^k - x^k$ *y* $\theta_k \in [0,1]$. *Supongamos además que* x^{k+1} *es obtenida mediante un paso de Cauchy de* y^k, *satisfaciendo* (3.0.2), *donde la constante de Lipschitz* L *es posiblemente infinita entonces*

$$\phi_{k+1}^*(\alpha) \geq f(x^{k+1}) + \left(\frac{1}{4L} - \frac{\alpha^2}{2\gamma_{k+1}}\right) \| \bigtriangledown f(y^k)\|^2 + (1-\alpha)\zeta(\alpha, \theta_k). \qquad (3.0.9)$$

Además

$$\zeta(\alpha, \theta_k) \geq \left(-\theta + \frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}(1-\theta_k)\right) f'(x^k + \theta_k d^k, d^k). \qquad (3.0.10)$$

Prueba:
De (3.0.7) y la hipótesis, tenemos que,

$$\begin{aligned}\phi_{k+1}^*(\alpha) &\geq f(x^{k+1}) + \frac{1}{4L}\| \bigtriangledown f(y^k)\|^2 - \frac{\alpha^2}{2\gamma_{k+1}}\| \bigtriangledown f(y^k)\|^2 + (1-\alpha)\zeta(\alpha, \theta_k) \\ &= f(x^{k+1}) + \left(\frac{1}{4L} - \frac{\alpha^2}{2\gamma_{k+1}}\right) \| \bigtriangledown f(y^k)\|^2 + (1-\alpha)\zeta(\alpha, \theta_k).\end{aligned}$$

De (3.0.8), la hipótesis y el hecho de que $\mu \geq 0$, obtenemos,

$$\zeta(\alpha, \theta_k) \geq f(x^k) - f(x^k + \theta_k d^k) + \frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}(1-\theta_k) f'(x^k + \theta_k d^k, d^k).$$

Por la convexidad de $f(\cdot)$ y el hecho que $\mu \geq 0$, se cumple,

$$f(x^k) - f(x^k + \theta_k d^k) \geq -\theta_k f'(x^k + \theta_k d^k, d^k).$$

Por lo tanto,

$$\begin{aligned}\zeta(\alpha, \theta_k) &\geq -\theta_k f'(x^k + \theta_k d^k, d^k) + \frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}(1-\theta_k) f'(x^k + \theta_k d^k, d^k) \\ &= \left(-\theta_k + (1-\theta_k)\frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}\right) f'(x^k + \theta_k d^k, d^k),\end{aligned}$$

completando la prueba.

■

Ahora estamos listos para demostrar los dos objetivos principales anteriormente mencionados.
Comenzaremos por describir la elección de Nesterov.
Elección de Nesterov. Cuando la constante de Lipschitz L se conoce, las elecciones de Nesterov son:

- α_N se calcula de manera que el término central en (3.0.9) sea nulo. Por lo tanto α_N es la solución positiva de la ecuación de segundo grado

$$2L\alpha^2 - (1-\alpha)\gamma_k - \alpha\mu = 0 \tag{3.0.11}$$

- θ_N es tal que el lado derecho de (3.0.10) sea nulo:

$$\theta_N = \frac{\gamma_k}{\gamma_k + \alpha_N \mu}\alpha_N \tag{3.0.12}$$

Nota 3.0.1. : Si L es desconocida, pero existe, α_N y θ_N están bien definidas (pero desconocidas), de lo contrario tomamos $\alpha_N = 0$, $\theta_N = 0$.

El Método de Nesterov toma $\alpha_k = \alpha_N$ y $\theta_k = \theta_N$, obteniendo directamente de (3.0.9) y el hecho de que $\zeta(\alpha_N, \theta_N) \geq 0$ que

$$\phi^*_{k+1} \geq f(x^{k+1}) \qquad y \qquad \alpha_N = \sqrt{\frac{\gamma_{k+1}}{2L}}. \tag{3.0.13}$$

El lema siguiente utiliza la definición de α_N (posiblemente nula si L es desconocida) para demostrar en qué condiciones se puede calcular $\alpha_k \geq \alpha_N$ tal que $\phi^*_{k+1}(\alpha_k) = f(x^{k+1})$. Usaremos directamente el lema 2.2.4.

Por construcción, tenemos que,

$$\phi^*_{k+1}(0) = \phi^*_k \geq f(x^k)$$

$$\phi^*_{k+1}(1) = \inf_{x\in\mathbb{R}^n} l(x) \leq f(x^*) \leq f(x^{k+1}). \tag{3.0.14}$$

Comencemos por eliminar un caso trivial: debido a (3.0.14), si $\phi^*_{k+1}(1) = f(x^{k+1})$ entonces x^{k+1} es una solución óptima y podemos terminar el algoritmo.
Supongamos que $\phi^*_{k+1}(1) < f(x^{k+1})$.

Lema 3.0.2. *Consideremos una iteración k del algoritmo 2.0.1. Supongamos que $\phi^*_{k+1}(1) < f(x^{k+1})$ y que $y^k = x^k + \theta_k d^k$ se selecciona de manera que $\zeta(\alpha_N, \theta_k) \geq 0$, entonces existe un valor $\alpha_k \in [\alpha_N, 1]$ que resuelve la ecuación $\phi^*_{k+1}(\alpha_k) = f(x^{k+1})$ en* (3.0.6)*.*

Prueba:
Como $l_k(\cdot)$ es una aproximación cuadrática inferior de $f(\cdot)$ y $\phi^*_k = f(x^k)$, tenemos que,

$$\phi^*_k = f(x^k) \geq l_k(x^k) \geq \inf_{x \in \mathbb{R}^n} l_k(x).$$

Aplicando el lema 3.0.1 con $\alpha = \alpha_N$ y utilizando la hipótesis, obtenemos que,

$$\phi^*_{k+1}(\alpha_N) \geq f(x^{k+1}).$$

Tomando $P = f(x^{k+1})$ obtenemos por el lema 2.2.4 que la ecuación $\phi^*_{k+1}(\alpha) = f(x^{k+1})$ tiene una o dos raíces reales y la raíz mas grande α' está en el intervalo $[0, 1]$. Notemos que $\alpha' \geq \alpha_N$ ya que α' es la raíz mas grande de la función cóncava $\phi^*_{k+1}(\cdot) - P$.

■

Concluimos de este lema que la igualdad $\phi^*_{k+1}(\alpha_k) = f(x^{k+1})$ se logra siempre que $\zeta(\alpha_N, \theta_k) \geq 0$.

Teorema 3.0.1. *El algoritmo 2.0.1 con ambas elecciones (Nesterov y la de Gonzaga y Karas [3]) de θ_k, está bien definido y genera sucesiones (x^k) y (ϕ^*_k) tales que para todo $k = 0, 1, \ldots$, $\phi^*_k = f(x^k)$. Además de esto, los parámetros α_k y γ_k satisfacen*

$$\gamma_k - \mu = \lambda_k(\gamma_0 - \mu) \qquad y \qquad \alpha_k \geq \sqrt{\frac{\gamma_{k+1}}{2L}}. \tag{3.0.15}$$

Prueba:

(i) Comencemos por demostrar que las dos versiones del algoritmo están bién definidas:
Debemos suponer que $\phi^*_{k+1}(1) < f(x^{k+1})$ ya que de lo contrario el algoritmo terminaría con x^{k+1} como solución óptima del problema (2.0.1) como comentamos anteriormente.

Nesterov: inmediata. Tomando $\theta_k = \theta_N$ obtenemos por la definición de α_N, θ_N y (3.0.10) que $\zeta(\alpha_N, \theta_N) \geq 0$, luego por el lema 3.0.2 tenemos que existe $\alpha_k \in [\alpha_N, 1)$ tal que $\phi^*_{k+1}(\alpha_k) = f(x^{k+1})$.

Método propuesto por Gonzaga y Karas [3]: supongamos que $\phi^*_k = f(x^k)$. Como $\mu \geq 0$, $f(x^k) - f(x^k + \theta_k d^k) \geq 0$ y $f'(x^k + \theta_k d^k, d^k) \geq 0$ tenemos de (3.0.8) que,

$$\begin{aligned} \zeta(\alpha, \theta_k) &\geq f(x^k) - f(x^k + \theta_k d^k) + \frac{\alpha\gamma_k}{(1-\alpha)\gamma_k + \alpha\mu}(1-\theta_k) f'(x^k + \theta_k d^k, d^k) \\ &\geq 0 \end{aligned}$$

para cada $\alpha \in [0, 1)$ en particular para $\alpha = \alpha_N$. Por el lema 3.0.2 existe $\alpha_k \in [\alpha_N, 1)$ tal que $\phi^*_{k+1}(\alpha_k) = f(x^{k+1})$.

(ii) En ambos casos obtenemos que $\zeta(\alpha_k, \theta_k) \geq 0$. Por lo tanto de (3.0.9),

$$f(x^{k+1}) = \phi^*_{k+1}(\alpha_k) \geq f(x^{k+1}) + \left(\frac{1}{4L} - \frac{\alpha^2}{2\gamma_{k+1}}\right) \| \bigtriangledown f(y^k)\|^2.$$

De acá,

$$\left(\frac{1}{4L} - \frac{\alpha^2}{2\gamma_{k+1}}\right) \leq 0$$

y despejando α_k, tenemos que,

$$\alpha_k \geq \sqrt{\frac{\gamma_{k+1}}{2L}}.$$

(iii) Sabemos que $\lambda_0 = 1$ así que

$$\gamma_0 - \mu = \lambda_0(\gamma_0 - \mu),$$

por lo tanto podemos usar inducción.
Por la definición de la sucesión (λ_k) (la cual esta definida por recursión) y la hipótesis inductiva, tenemos que,

$$\begin{aligned}\gamma_{k+1}-\mu &= (1-\alpha_k)\gamma_k+\alpha_k\mu-\mu\\ &= (1-\alpha_k)\gamma_k-(1-\alpha_k)\mu\\ &= (1-\alpha_k)(\gamma_k-\mu)\\ &= (1-\alpha_k)\lambda_k(\gamma_0-\mu)\\ &= \lambda_{k+1}(\gamma_0-\mu),\end{aligned}$$

completando la prueba.

■

Concluimos de este teorema que el algoritmo de Nesterov, el cual usa los valores $\theta_k = \theta_N$ y $\alpha_k = \alpha_N$ puede recalcular α_k usando (3.0.6) y preservar al mismo tiempo $\theta_k = \theta_N$.

Nota 3.0.2. Incluso con un cálculo del mejor valor de α_k, la sucesión $(f(x^k))$ generada por el método de Nesterov no es necesariamente decreciente.

Nota 3.0.3. En el caso de $\mu = 0$, las expresiones son simplificadas, pues α_N es la solución positiva de $2L\alpha_k^2 = (1-\alpha_k)\gamma_k$ y la expresión (3.0.10) se reduce a

$$\begin{aligned}\zeta(\alpha,\theta_k) &\geq \left(-\theta_k+\frac{\alpha}{1-\alpha}(1-\theta_k)\right)f'(x^k+\theta_k d^k,d^k)\\ &= \left(\frac{-\theta_k(1-\alpha)+\alpha(1-\theta_k)}{1-\alpha}\right)f'(x^k+\theta_k d^k,d^k)\\ &= \frac{\alpha-\theta_k}{1-\alpha}f'(x^k+\theta_k d^k,d^k).\end{aligned}$$

Además de la definición de θ_N, tenemos que $\theta_N = \alpha_N$.

3.1. Más información de la elección de θ_k.

La elección de θ_k propuesta por Gonzaga y Karas [3] depende de una búsqueda lineal en la dirección de d^k. El esfuerzo necesario para esta búsqueda depende en

gran medida de las dificultades para calcular los valores comparativos de la función y los gradientes. Si un cálculo del gradiente requiere mucho más tiempo que una evaluación de la función entonces la búsqueda lineal puede ser muy buena. De lo contrario el esfuerzo en la búsqueda lineal debe reducirse. En esta sección se muestran algunos resultados que muestran cómo podemos simplificar esta búsqueda.

El alcance es discutir la cantidad de cálculos necesarios en esta búsqueda lineal y demostrar que cuando L y μ se conocen la búsqueda lineal se puede hacer en un tiempo polinomial.

Concluimos del análisis anterior que lo que debemos demostrar es que $\zeta(\alpha_N, \theta_k) \geq 0$ para una elección dada de $y^k = x^k + \theta_k d^k$ en la iteración k. El siguiente lema resume las condiciones que aseguran esta propiedad las cuales darán lugar a nuevos algoritmos.

Lema 3.1.1. *Sea $y^k = x^k + \theta_k d^k$ la elección dada por el algoritmo 2.0.1 en la iteración k y sea θ_N la elección de Nesterov* (3.0.12)*. Supongamos que una de las siguientes condiciones se cumplen:*

(i) $f(y^k) \leq f(x^k)$ *y* $f'(y^k, d^k) \geq 0$*;*

(ii) $f(y^k) \leq f(x^k)$ *y* $\theta_k \geq \theta_N$*;*

(iii) $f'(y^k, d^k) \geq 0$ *y* $\theta_k \leq \theta_N$*;*

(iv) $f'(x^k, d^k) \geq -\frac{\mu}{2}\|d^k\|^2$ *y* $\theta_k = 0$.

Entonces $\zeta(\alpha_N, \theta_k) \geq 0$ (y $\alpha_k \geq \alpha_N$ puede calcularse utilizando (3.0.6)*).*

Prueba:

(i) Se demostró en el teorema 3.0.1.

Sustituyendo $\theta_k = \theta_N + \Delta\theta$ en (3.0.10) con $\gamma_N = (1-\alpha_N)\gamma_k + \alpha_N\mu$, tenemos que,

$$\zeta(\alpha_N, \theta_k) \geq \left(-\theta_N + \frac{\alpha_N\gamma_k}{\gamma_N}(1-\theta_N) - \Delta\theta - \frac{\alpha_N\gamma_k}{\gamma_N}\Delta\theta\right) f'(x^k + \theta_k d^k, d^k).$$

Por otro lado, obtenemos,

$$-\theta_N + \frac{\alpha_N\gamma_k}{\gamma_N}(1-\theta_N) = \frac{-\theta_N\gamma_N + \alpha_N\gamma_k(1-\theta_N)}{\gamma_N}.$$

Desarrollemos $1-\theta_N$ usando la definición de θ_N (3.0.12),

$$\begin{aligned} 1-\theta_N &= 1 - \frac{\gamma_k}{\gamma_k + \alpha_N\mu}\alpha_N \\ &= \frac{\gamma_k + \alpha_N\mu - \gamma_k\alpha_N}{\gamma_k + \alpha_N\mu} \\ &= \frac{(1-\alpha_N)\gamma_k + \alpha_N\mu}{\gamma_k + \alpha_N\mu} \\ &= \frac{\gamma_N}{\gamma_k + \alpha_N\mu}. \end{aligned}$$

Sustituyendo esta expresión en la anterior, obtenemos que,

$$-\theta_N + \frac{\alpha_N\gamma_k}{\gamma_N}(1-\theta_N) = 0.$$

Por lo tanto

$$\zeta(\alpha_N, \theta_k) \geq -\left(1 + \frac{\alpha_N\gamma_k}{\gamma_N}\right)\Delta\theta f'(x^k + \theta_k d^k, d^k). \tag{3.1.1}$$

(ii) Supongamos que $f(y^k) \leq f(x^k)$ y $\theta_k \geq \theta_N$.

Acá vamos a estudiar dos casos:

Caso 1: $f'(y^k, d^k) \leq 0$

Como $\Delta\theta = \theta_k - \theta_N \geq 0$ se tiene de (3.1.1) que $\zeta(\alpha_N, \theta_k) \geq 0$.

Caso2: $f'(y^k, d^k) \geq 0$

De (i) se obtiene que $\zeta(\alpha_N, \theta_k) \geq 0$.

(iii) Supongamos que $f'(y^k, d^k) \geq 0$ y $\theta_k \leq \theta_N$.
Como $\Delta\theta = \theta_k - \theta_N \leq 0$, obtenemos de (3.1.1) que $\zeta(\alpha_N, \theta_k) \geq 0$.

(iv) Supongamos que $f'(x^k, d^k) \geq -\frac{\mu}{2}\|d^k\|^2$ y $\theta_k = 0$.
Tomando $\alpha = \alpha_N$ en (3.0.8), tenemos que,

$$\begin{aligned}\zeta(\alpha_N, \theta_k) &= \phi_k^* - f(x^k + \theta_k d^k) + \frac{\alpha_N \gamma_k}{(1-\alpha_N)\gamma_k + \alpha_N \mu}\left(\frac{\mu}{2}(1-\theta_k)^2\|d^k\|^2 + \right. \\ &\left. +(1-\theta_k) f'(x^k + \theta_k d^k, d^k)\right).\end{aligned}$$

Por lo supuesto,

$$\begin{aligned}\zeta(\alpha_N, \theta_k) &= \phi_k^* - f(x^k) + \frac{\mu}{2}\frac{\alpha_N \gamma_k}{(1-\alpha_N)\gamma_k + \alpha_N \mu}\|d^k\|^2 + \\ &+ \frac{\alpha_N \gamma_k}{(1-\alpha_N)\gamma_k + \alpha_N \mu} f'(x^k, d^k) \\ &\geq f(x^k) - f(x^k) + \frac{\mu}{2}\frac{\alpha_N \gamma_k}{(1-\alpha_N)\gamma_k + \alpha_N \mu}\|d^k\|^2 - \\ &- \frac{\mu}{2}\frac{\alpha_N \gamma_k}{(1-\alpha_N)\gamma_k + \alpha_N \mu}\|d^k\|^2 \\ &= 0.\end{aligned}$$

■

Así que todo lo que tenemos que hacer es especificar en el algoritmo 2.0.1 la elección de θ_k y α_k como sigue:
Elijamos $\theta_k \in [0, 1]$ satisfaciendo una de las condiciones en el lema 3.1.1.
Calcular α_k tal que $\phi_{k+1}^*(\alpha_k) = f(x^{k+1})$ usando (3.0.6).

Debemos discutir el cálculo de θ_k el cual requiere una búsqueda lineal. Por supuesto, si L es dada podemos utilizar $\theta_k = \theta_N$ y asegurar la complejidad óptima, sin embargo los valores grandes de α_k pueden ser obtenidos.

Se quiere obtener θ_k (y luego α_k) tal que $\zeta(\alpha_k, \theta_k)$ sea tan grande como sea posible. Dado que el cálculo de α_k depende de un paso de máximo descenso de

$y^k = x^k + \theta_k d^k$ no se puede maximizar $\zeta(\cdot, \cdot)$. Pero los valores positivos de $\zeta(\alpha, \theta_k)$ serán obtenidos si θ_k es tal que $f(x^k) - f(x^k + \theta_k d^k)$ y $f'(x^k + \theta_k d^k)$ son los más grandes posibles (por supuesto, estos son objetivos en conflicto).

Ahora se definen (pero no se calculan) dos puntos:

$$\theta' \in argmin\{f(x^k + \theta d^k) : \theta \geq 0\} \tag{3.1.2}$$

$$\theta'' = \text{máx}\{\theta \in [0, 1] : f(x^k + \theta d^k) \leq f(x^k)\}. \tag{3.1.3}$$

Veamos dos casos importantes:

- Si $f'(x^k, d^k) \geq 0$ (lo que significa que d^k no es dirección de descenso) entonces se debe elegir $\theta_k = 0$: un paso de Cauchy apartir de x^k sigue.
- Si $f(x^k + d^k) \leq f(x^k)$ entonces se puede elegir $\theta_k = 1$: un paso de Cauchy apartir de v^k sigue.

Si ninguna de estas dos situaciones especiales ocurre entonces $0 < \theta' < \theta'' < 1$. Cualquier punto $\theta_k \in [\theta', \theta'']$ cumple que $\zeta(\alpha, \theta_k) \geq 0$ para cada $\alpha \in [0, 1]$ y $y^k = x^k + \theta_k d^k$ satisface la condición (i) del lema 3.1.1. Se puede calcular un punto en este intervalo usando el algoritmo *reducción del intervalo* descrito en el apéndice B

No se tiene estimaciones de complejidad para un procedimiento de búsqueda lineal tal como éste si L y μ no están disponibles. Cuando estas constantes son conocidas es posible reducir el número de iteraciones en la búsqueda lineal para mantener la complejidad de Nesterov. Discutiremos brevemente la elección de θ_k en tres posibles situaciones, de acuerdo a nuestro conocimiento de las constantes.

Caso 1: No se tiene conocimiento de L o μ. Se comienza la iteración chequeando los casos especiales anteriores: si $f(x^k + d^k) \leq f(x^k)$ entonces se elige $\theta_k = 1$: se seguirá entonces un paso de Cauchy apartir de v^k. Para el otro caso se debe elegir θ_k siempre que $f'(x^k, d^k) \geq 0$. Esto se puede hacer como sigue:

Se calcula $f(x^k + \tilde{\theta}d^k)$ para un valor pequeño $\tilde{\theta} > 0$.
Si $f(x^k + \tilde{\theta}d^k) \geq f(x^k)$ se calcula $\bigtriangledown f(x^k)$ y luego $f'(x^k, d^k) = \bigtriangledown f(x^k)^T d^k$.
Si no $f'(x^k, d^k) < 0$.
Si $f'(x^k, d^k) \geq 0$ se toma $\theta_k = 0$.
Si no, se hace una búsqueda lineal como en el Apéndice B, iniciando con los puntos 0, $\tilde{\theta}$ y 1.

Por supuesto no se puede especificar el significado de "pequeño $\tilde{\theta}$", pero esto es por lo general fácil en las aplicaciones prácticas. Una suposición "intelectualmente satisfactoria" es la siguiente: sea L una cota superior para la (desconocida) constante de Lipschitz. Calculemos α_N y $\tilde{\theta} = \theta_N$ por (3.0.11) y (3.0.12) respectivamente.

Caso 2: L es dado. En este caso se inicia la búsqueda con θ_N calculada como en (3.0.12). La búsqueda lineal puede ser ajustada para mantener el número de cálculos de la función dentro de una cota dada, resultando en un paso satisfactorio una de las tres primeras condiciones en el lema 3.1.1. Hay dos casos a considerar, dependiendo del signo de $f(x^k) - f(x^k + \theta_N d^k)$.

- $f(x^k + \theta_N d^k) < f(x^k)$: la condición (ii) del lema 3.1.1 se satisface.
  Se puede usar un método de reducción del intervalo (ver Apéndice B) o simplemente tomar pasos para hacer crecer a θ_k. Un método trivial es el siguiente, usando una constante $\beta > 1$:
  $\theta_k = \theta_N$.
  Mientras que $\beta\theta_k \leq 1$ y $f(x^k + \beta\theta_k) \leq f(x^k)$ tomar
  $\theta_k = \beta\theta_k$.

  En cualquier método, el número de cálculos de la función puede ser limitado por una constante dada, adoptando θ_k la longitud de paso más grande calculada la cual cumpla la condición de descenso.

- $f(x^k + \theta_N d^k) > f(x^k)$: se cumple la condición (iii) del lema 3.1.1 para θ_N y se tiene la intención de reducir el valor de θ_k. Para una búsqueda de

reducción del intervalo entre 0 y θ_N se necesita un punto intermedio ν tal que $f(x^k + \nu d^k) \leq f(x^k)$ si este existe. Se Propone el siguiente procedimiento:
Calcular $f(x^k + \nu d^k)$ para $\nu \ll \theta_N$.
Si $f(x^k + \nu d^k) \geq f(x^k)$, tomar $\theta_k = \nu$.
Si no calcular θ_k por un método de reducción del intervalo (ver Apéndice B), comenzando con los puntos 0, ν y θ_N. El algoritmo puede interrumpirse en cualquier momento asignando $\theta_k = B$.

Caso 3: L y μ son dados. Este es un caso especial del caso 2, con una interesante propiedad: un punto que cumple la condición (i) en el lema 3.1.1 se puede calcular en un tiempo polinomial sin la necesidad de calcular $\bigtriangledown f(x^k)$, usando el siguiente lema.

Lema 3.1.2. *Considere $\theta' \in argmin\{f(x^k + \theta d^k) : \theta \geq 0\}$ y $\theta'' > \theta'$ tal que $f(x^k + \theta'' d^k) = f(x^k)$ si este existe. Definir $\bar{\theta} = \frac{\mu}{2L}$.*
Si $f(x^k + \bar{\theta} d^k) \leq f(x^k) - \frac{\mu^2}{8L}\|d^k\|^2$ entonces $\theta' \geq \bar{\theta}$ y $\theta'' - \theta' \geq \bar{\theta}$, de lo contrario $f'(x^k, d^k) \geq -\frac{\mu}{2}\|d^k\|^2$.

Prueba:
Supongamos que,

$$f(x^k + \theta' d^k) \leq f(x^k + \bar{\theta} d^k) \leq f(x^k) - \frac{\mu^2}{8L}\|d^k\|^2. \qquad (*)$$

Como $\theta' \in argmin\{f(x^k + \theta d^k) : \theta \geq 0\}$, entonces

$$f'(x^k + \theta' d^k, d^k) = \bigtriangledown f(x^k + \theta' d^k)^T d^k = 0.$$

Por el teorema 2.0.1, para $\theta \in \mathbb{R}$, tenemos que,

$$\begin{aligned} f(x^k + \theta d^k) &\leq f(x^k + \theta' d^k) + \frac{L}{2}\|(x^k + \theta d^k - x^k - \theta' d^k)\|^2 \\ &= f(x^k + \theta' d^k) + \frac{L}{2}(\theta - \theta')^2\|d^k\|^2. \end{aligned}$$

Usando $(*)$,

$$\begin{aligned} f(x^k + \theta d^k) &\leq f(x^k) + \frac{L}{2}(\theta - \theta')^2\|d^k\|^2 - \frac{\mu^2}{8L}\|d^k\|^2 \\ &= f(x^k) + \left(L(\theta - \theta')^2 - \frac{\mu^2}{4L}\right)\frac{\|d^k\|^2}{2}. \qquad (3.1.4) \end{aligned}$$

Evaluando (3.1.4) en $\theta = 0$, tenemos que,

$$f(x^k) \leq f(x^k) + \left(L\theta'^2 - \frac{\mu^2}{4L}\right)\frac{\|d^k\|^2}{2}.$$

Efectuando operaciones, obtenemos,

$$L\theta'^2 - \frac{\mu^2}{4L} \geq 0$$

o equivalentemente que

$$\theta'^2 \geq \frac{\mu^2}{4L^2}.$$

Por la definición de $\bar{\theta}$ y el hecho de que $\theta', \bar{\theta} \geq 0$ se tiene que, $\theta' \geq \bar{\theta}$, probando la primera desigualdad.

Por otra parte de la hipótesis $f(x^k + \theta''d^k) = f(x^k)$, en consecuencia para $\theta = \theta''$, $L(\theta'' - \theta')^2 - \frac{\mu^2}{4L} \geq 0$.
Por la definición de $\bar{\theta}$ y el hecho que $\theta'' - \theta', \bar{\theta} \geq 0$, tenemos, $\theta'' - \theta' \geq \bar{\theta}$, probando la segunda desigualdad.

Supongamos ahora que $f(x^k + \bar{\theta}d^k) > f(x^k) - \frac{\mu^2}{8L}\|d^k\|^2$.
Por el teorema 2.0.1, tenemos,

$$\begin{aligned} f(x^k + \bar{\theta}d^k) &\leq f(x^k) + \bigtriangledown f(x^k)^T(x^k + \bar{\theta}d^k - x^k) + \frac{L}{2}\|x^k + \bar{\theta}d^k - x^k\|^2 \\ &= f(x^k) + f'(x^k, d^k)\bar{\theta} + \frac{L}{2}\bar{\theta}^2\|d^k\|^2. \end{aligned}$$

Suponiendo por reducción al absurdo que $f'(x^k, d^k) < -\frac{\mu}{2}\|d^k\|^2$ y utilizando la definición de $\bar{\theta}$, obtenemos,

$$\begin{aligned} f(x^k + \bar{\theta}d^k) &< f(x^k) - \frac{\mu}{2}\bar{\theta}\|d^k\|^2 + \frac{L}{2}\bar{\theta}^2\|d^k\|^2 \\ &= f(x^k) - \frac{\mu^2}{4L}\|d^k\|^2 + \frac{\mu^2}{8L}\|d^k\|^2 \\ &= f(x^k) - \frac{\mu^2}{8L}\|d^k\|^2, \end{aligned}$$

contradiciendo la hipótesis y completando la prueba.

■

En este caso, podemos usar una búsqueda dorada de la siguiente manera:
Si $f(x^k + \bar{\theta}d^k) > f(x^k) - \frac{\mu^2}{8L}\|d^k\|^2$ tomar $\theta_k = 0$ y pare (la condición (iv) en el lema 3.1.1 se cumple). Si $f(x^k + d^k) \leq f(x^k)$ tomar $\theta_k = 1$ y pare (la condición (ii) en el lema 3.1.1 se cumple).
(Ahora sabemos que $\theta' \leq \theta'' - \frac{\mu}{2L} < 1$).
Usemos una búsqueda de reducción del intervalo (ver Apéndice B) en el intervalo $[\bar{\theta}, 1]$.

Si usamos una búsqueda de la sección dorada (ver Apéndice A) entonces para cada iteración j antes de encontrarnos con la condición de parada, tenemos $A \leq \theta' \leq \theta'' \leq B$ y la longitud del intervalo satisface que $B - A \leq \beta^j$ con $\beta = (\sqrt{5} - 1)/2 \approx 0{,}62$. Usando el Lema anterior, un punto $B \in [\theta', \theta'']$ será encontrado cuando $\beta^j \leq \frac{\mu}{2L}$.
Por lo tanto el número de iteraciones de la búsqueda estará acotado por

$$j_{max} = \frac{log(\frac{\mu}{2L})}{log\beta}.$$

Esto proporciona una cota en el esfuerzo computacional por iteración de $O(log(\frac{\mu}{2L}))$ evaluaciones de la función.

3.2. Parámetro de convexidad adaptativo μ

En las pruebas computacionales se puede observar que el uso de una constante de convexidad fuerte es muy efectivo en la reducción del numero de iteraciones. Pero muy pocas veces tal constante es asequible. En esta sección fijaremos un algoritmo que utiliza una sucesión decreciente de estimaciones para el valor de la constante μ. Supongamos que un parámetro de convexidad fuerte μ^* es dado y el método generará una sucesión $\mu_k \longrightarrow \mu^*$ comenzando con $\mu_0 \in [\mu^*, \gamma_0)$.
Todavía se necesita una hipótesis adicional: que el conjunto nivel asociado con

x^0 sea acotado. Ya que $f(\cdot)$ es convexa, esto es equivalente a la hipótesis de que el conjunto óptimo sea acotado. Empezaremos por especificar como el algoritmo puede ser aplicado y los comentarios los haremos después.
Las iteraciones del algoritmo son las mismas que en el algoritmo 2.0.1, usando el parámetro $\mu_k \geq \mu^*$. El parámetro es reducido (se hace esto sustituyendo $\mu_k = max\{\mu^*, \frac{\mu_k}{10}\}$) en dos situaciones:

- Cuando $\gamma_k - \mu^* < \beta(\mu_k - \mu^*)$ donde $\beta > 1$ es fijo (se usa $\beta = 1{,}02$) significando que γ_k esta demasiado cerca de μ_k.
- Cuando es imposible satisfacer la condición $\phi_+^*(\alpha) = f(x^{k+1})$ para $\alpha \in [0,1]$. Por el Lema 2.2.4 esto sólo puede ocurrir cuando $P = f(x^{k+1}) < \min\limits_{x \in \mathbb{R}^n} l(x)$. Por (2.2.14) este mínimo viene dado como $\phi_{k+1}^*(1) = f(y^k) - \frac{\|\bigtriangledown f(y^k)\|^2}{2\mu_k}$. Así μ_k no puede ser mayor que

$$\tilde{\mu} = \frac{\| \bigtriangledown f(y^k)\|^2}{2(f(y^k) - f(x^{k+1}))}.$$

Si $\mu_k > \tilde{\mu}$ se reduce μ_k.

Notación: denotemos $\phi_{k+1}(x) = \phi_{k+1}(\alpha_k, x)$ y $\phi_{k+1}^* = \phi_{k+1}^*(\alpha_k)$.

Algoritmo 3.2.1. Algoritmo con parámetro de convexidad adaptivo.
Datos: $x^0 \in \mathbb{R}^n$, $v^0 = x^0$, $\gamma_0 > 0$, $\beta > 1$, $\mu^* = \mu$, $\mu_0 \in [\mu^*, \gamma_0)$.
(Sugerimos $\mu_0 = \text{máx}\{\mu^*, \frac{\gamma_0}{100}\}$, $\beta = 1{,}02$, $\gamma_0 = L$ si L es conocida)
$k = 0$.
Repetir
$d^k = v^k - x^k$.
Elegir $\theta_k \in [0,1]$ como en el Algoritmo 2.0.1.
$y^k = x^k + \theta_k d^k$.
Si $\bigtriangledown f(y^k) = 0$ entonces pare con y^k como una solución óptima.
Paso de máximo descenso: $x^{k+1} = y^k - \nu \bigtriangledown f(y^k)$. Si L es conocida, $\nu \geq \frac{1}{L}$.
Si $\gamma_k - \mu^* < \beta(\mu_k - \mu^*)$ entonces $\mu_k = \text{máx}\{\mu^*, \frac{\mu_k}{10}\}$.

Calcular $\tilde{\mu} = \frac{\|\bigtriangledown f(y^k)\|^2}{2(f(y^k)-f(x^{k+1}))}$.
Si $\mu_k > \tilde{\mu}$ entonces $\mu_k = \text{máx}\{\mu^*, \frac{\tilde{\mu}}{10}\}$.
Calcular α_k como la raíz mas grande de (2.1.3) con $\mu = \mu_k$.
Tomar $\mu_{k+1} = \mu_k$
$\gamma_{k+1} = (1-\alpha_k)\gamma_k + \alpha_k\mu_k$
$v^{k+1} = \frac{1}{\gamma_{k+1}}((1-\alpha_k)\gamma_k v^k + \alpha_k(\mu_k y^k - \bigtriangledown f(y^k)))$
$k = k+1$.

Ahora mostraremos la optimalidad del algoritmo, usando una hipótesis adicional.

Hipótesis. Dado $x^0 \in \mathbb{R}^n$, supongamos que el conjunto nivel asociado con x^0 es acotado, es decir,

$$D = \sup\{\|x-y\| : f(x) \leq f(x^0), f(y) \leq f(x^0)\} < \infty.$$

Definimos $Q = \frac{D^2}{2}$.

Lema 3.2.1. *Considere $\gamma_0 > \mu^*$. Sea x^* una solución óptima del problema* (2.0.1) *entonces para toda iteración del Algoritmo 3.2.1,*

$$\phi_k(x^*) - f(x^*) \leq \frac{\gamma_0 + L}{\gamma_0 - \mu^*} Q(\gamma_k - \mu^*). \tag{3.2.1}$$

Prueba:
Primero probaremos (3.2.1) para $k = 0$. Por definición de $\phi_0(\cdot)$, el lema 2.0.1, la definición de D y Q, tenemos que,

$$\begin{aligned}
\phi_0(x^*) - f(x^*) &= f(x^0) - f(x^*) + \frac{\gamma_0}{2}\|x^* - x^0\|^2 \\
&\leq \frac{L+\gamma_0}{2}\|x^* - x^0\|^2 \\
&\leq \frac{L+\gamma_0}{2} D^2 \\
&= (L+\gamma_0)Q \\
&= \frac{L+\gamma_0}{\gamma_0 - \mu^*} Q(\gamma_0 - \mu^*).
\end{aligned}$$

Por lo tanto podemos usar inducción.
Por (2.0.3), (2.2.3) y el hecho que $\mu_k \geq \mu^*$, tenemos,

$$\begin{aligned}\phi_{k+1}(x) &= (1-\alpha_k)\phi_k(x) + \alpha_k\left(f(y^k) + \bigtriangledown f(y^k)^T(x-y^k) + \frac{\mu_k}{2}\|x-y^k\|^2\right)\\ &\leq (1-\alpha_k)\phi_k(x) + \alpha_k f(x)\\ &\leq (1-\alpha_k)\phi_k(x) + \alpha_k\left(f(x) + \frac{\mu_k-\mu^*}{2}\|x-y^k\|^2\right).\end{aligned}$$

Evaluando la desigualdad anterior en x^* y usando la definición de D y Q,

$$\begin{aligned}\phi_{k+1}(x^*) - f(x^*) &\leq (1-\alpha_k)(\phi_k(x^*) - f(x^*)) + \alpha_k\left(\frac{\mu_k-\mu^*}{2}\|x^*-y^k\|^2\right)\\ &\leq (1-\alpha_k)(\phi_k(x^*) - f(x^*)) + \alpha_k Q(\mu_k-\mu^*)\end{aligned}$$

Usando la hipótesis inductiva, la definición de γ_{k+1} y la hipótesis, tenemos que,

$$\begin{aligned}\phi_{k+1}(x^*) - f(x^*) &\leq (1-\alpha_k)\frac{\gamma_0+L}{\gamma_0-\mu^*}Q(\gamma_k-\mu^*) + \alpha_k Q(\mu_k-\mu^*)\\ &\leq \frac{\gamma_0+L}{\gamma_0-\mu^*}Q((1-\alpha_k)\gamma_k + \alpha_k\mu_k - \mu^*)\\ &= \frac{\gamma_0+L}{\gamma_0-\mu^*}Q(\gamma_{k+1}-\mu^*),\end{aligned}$$

completando la prueba.

■

Ahora probaremos un lema técnico, imitando un resultado muy similar en [8].

Lema 3.2.2. *Consideremos una sucesión positiva (λ_k). Supongamos que existe $M > 0$ tal que $\lambda_{k+1} \leq (1 - M\sqrt{\lambda_{k+1}})\lambda_k$, entonces para todo $k > 0$,*

$$\lambda_k < \frac{4}{M^2}\frac{1}{k^2}.$$

Prueba:

Denotemos $a_k = \frac{1}{\sqrt{\lambda_k}}$. Ya que $\{\lambda_k\}$ es una sucesión decreciente, tenemos,

$$\begin{aligned} a_{k+1} - a_k &= \frac{\sqrt{\lambda_k} - \sqrt{\lambda_{k+1}}}{\sqrt{\lambda_k \lambda_{k+1}}} \\ &= \frac{\lambda_k - \lambda_{k+1}}{\sqrt{\lambda_k \lambda_{k+1}}(\sqrt{\lambda_k} + \sqrt{\lambda_{k+1}})} \\ &= \frac{\lambda_k - \lambda_{k+1}}{\lambda_k \sqrt{\lambda_{k+1}} + \sqrt{\lambda_k \lambda_{k+1}}\sqrt{\lambda_{k+1}}} \\ &\geq \frac{\lambda_k - \lambda_{k+1}}{\lambda_k \sqrt{\lambda_{k+1}} + \sqrt{\lambda_k \lambda_{k+1}}\sqrt{\lambda_k}} \\ &= \frac{\lambda_k - \lambda_{k+1}}{\lambda_k \sqrt{\lambda_{k+1}} + \lambda_k \sqrt{\lambda_{k+1}}} \\ &= \frac{\lambda_k - \lambda_{k+1}}{2\lambda_k \sqrt{\lambda_{k+1}}}. \end{aligned}$$

Usando la hipótesis,

$$\begin{aligned} a_{k+1} - a_k &\geq \frac{\lambda_k - (1 - M\sqrt{\lambda_{k+1}})\lambda_k}{2\lambda_k \sqrt{\lambda_{k+1}}} \\ &= \frac{M\sqrt{\lambda_{k+1}}\lambda_k}{2\lambda_k \sqrt{\lambda_{k+1}}} \\ &= \frac{M}{2}. \end{aligned}$$

Por un proceso recursivo, tenemos,

$$a_k \geq a_0 + \frac{Mk}{2} > \frac{Mk}{2}.$$

Como $a_k = \frac{1}{\sqrt{\lambda_k}}$ concluimos que,

$$\lambda_k < \frac{4}{M^2}\frac{1}{k^2},$$

completando la prueba.

■

Ahora podemos estudiar la velocidad con que $(\gamma_k - \mu^*)$ tiende a cero para calcular

la velocidad de convergencia del algoritmo.

Lema 3.2.3. *Consideremos $\gamma_0 > \mu$, entonces para todo $k > 0$,*

$$\gamma_k - \mu^* < \frac{8\beta^2 L}{(\beta-1)^2}\frac{1}{k^2}.$$

Prueba:
Por la definición de γ_{k+1},

$$\gamma_{k+1} - \mu^* = (1-\alpha_k)(\gamma_k - \mu^*) + \alpha_k(\mu_k - \mu^*).$$

Por el algoritmo 3.2.1, tenemos que $(\gamma_k - \mu^*) \geq \beta(\mu_k - \mu^*)$. Por lo tanto,

$$\begin{aligned} \gamma_{k+1} - \mu^* &\leq (1-\alpha_k)(\gamma_k - \mu^*) + \frac{\alpha_k}{\beta}(\gamma_k - \mu^*) \\ &= \left(1 - \frac{\beta-1}{\beta}\alpha_k\right)(\gamma_k - \mu^*). \end{aligned}$$

Por el teorema 3.0.1,

$$\alpha_k \geq \sqrt{\frac{\gamma_{k+1}}{2L}} \geq \sqrt{\frac{\gamma_{k+1} - \mu^*}{2L}},$$

en consecuencia

$$\gamma_{k+1} - \mu^* \leq \left(1 - \frac{\beta-1}{\beta\sqrt{2L}}\sqrt{\gamma_{k+1} - \mu^*}\right)(\gamma_k - \mu^*).$$

Aplicando el lema 3.2.2 con $\lambda_k = \gamma_k - \mu^*$, para $k = 0, 1, 2\ldots$, $M = \frac{\beta-1}{\beta\sqrt{2L}}$ y efectuando operaciones, tenemos que para todo $k > 0$

$$\gamma_k - \mu^* < \frac{8\beta^2 L}{(\beta-1)^2}\frac{1}{k^2},$$

completando la prueba.

■

El siguiente teorema analiza la cota de complejidad del algoritmo 3.2.1.

Teorema 3.2.1. *Consideremos $\gamma_0 > \mu^* > 0$, entonces el algoritmo 3.2.1 genera una sucesión (x^k) tal que, para todo $k > 0$,*

$$f(x^k) - f(x^*) < \frac{8\beta^2 QL(L+\gamma_0)}{(\beta-1)^2(\gamma_0-\mu^*)k^2}.$$

Prueba:
Por la definición de $\phi_k(\cdot)$ y el hecho que $\phi_k^* = f(x^k)$,

$$f(x^k) \leq \phi_k(x)$$

para todo todo $x \in \mathbb{R}^n$, en particular para x^*. Usando esto, el lema 3.2.1 y por lo tanto el lema 3.2.3, tenemos que para cada $k > 0$,

$$f(x^k) - f(x^*) \leq \frac{L+\gamma_0}{\gamma_0-\mu^*} Q(\gamma_k - \mu^*) < \frac{8\beta^2 QL(L+\gamma_0)}{(\beta-1)^2(\gamma_0-\mu^*)k^2},$$

completando la prueba.

■

Este teorema asegura que un error de $\epsilon > 0$ para el valor final de la función objetivo se obtendrá en $O(\frac{1}{\sqrt{\epsilon}})$ iteraciones.

Apéndices

Apéndice A

Método de la sección dorada

Definición A.0.1. Una función estrictamente unimodal sobre un intervalo $[0, s]$ se define como una función que tiene un único mínimo global α^* en $[0, s]$ y si α_1, α_2 son dos puntos en $[0, s]$ tales que $\alpha_1 < \alpha_2 < \alpha^*$ o $\alpha^* < \alpha_1 < \alpha_2$, entonces $g(\alpha_1) > g(\alpha_2) > g(\alpha^*)$ o $g(\alpha^*) < g(\alpha_1) < g(\alpha_2)$, respectivamente, ver figura A.1.

Un ejemplo de una función estrictamente unimodal en $[0, s]$ es una función estrictamente convexa en $[0, s]$.

Supongamos que $g(\alpha)$ es estrictamente unimodal en el intervalo $[0, s]$. El método de la sección dorada minimiza $g(\cdot)$ sobre $[0, s]$ determinando en la iteración k un intervalo $[\alpha_k, \bar{\alpha_k}]$ que contenga α^* (el cual es el minimizador de $g(\cdot)$). Estos intervalos son obtenidos usando el numero

$$\tau = \frac{3 - \sqrt{5}}{2},$$

el cual satisface que $\tau = (1 - \tau)^2$ y esta vinculado con la sucesión numérica de Fibonacci.
Inicialmente se toma

$$[\alpha_0, \bar{\alpha_0}] = [0, s].$$

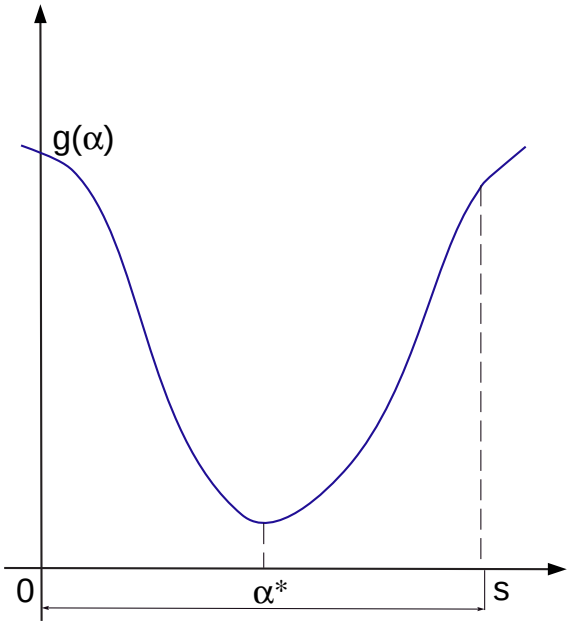

Figura A.1: Función unimodal

Dado $[\alpha_k, \bar{\alpha_k}]$ se determina $[\alpha_{k+1}, \bar{\alpha}_{k+1}]$ tal que $\alpha^* \in [\alpha_{k+1}, \bar{\alpha}_{k+1}]$ como sigue. Se calcula

$$\begin{aligned} b_k &= \alpha_k + \tau(\bar{\alpha_k} - \alpha_k) \\ \bar{b_k} &= \bar{\alpha_k} - \tau(\bar{\alpha_k} - \alpha_k) \end{aligned}$$

y $g(b_k)$, $g(\bar{b_k})$. Entonces:

1. Si $g(b_k) < g(\bar{b_k})$ se toma

$$\begin{aligned} &\alpha_{k+1} = \alpha_k, \quad && \bar{\alpha}_{k+1} = b_k \quad && si \quad && g(\alpha_k) \leq g(b_k) \\ &\alpha_{k+1} = \alpha_k, \quad && \bar{\alpha}_{k+1} = \bar{b_k} \quad && si \quad && g(\alpha_k) > g(b_k). \end{aligned}$$

2. Si $g(b_k) > g(\bar{b_k})$ se toma

$$\begin{aligned} &\alpha_{k+1} = \bar{b}_k, \quad && \bar{\alpha}_{k+1} = \bar{\alpha}_k \quad && si \quad && g(\bar{b}_k) \geq g(\bar{\alpha}_k) \\ &\alpha_{k+1} = b_k, \quad && \bar{\alpha}_{k+1} = \bar{\alpha}_k \quad && si \quad && g(\bar{b}_k) < g(\alpha_k). \end{aligned}$$

3. Si $g(b_k) = g(\bar{b}_k)$ se toma

$$\alpha_{k+1} = b_k, \qquad \bar{\alpha}_{k+1} = \bar{b}_k.$$

Basándose en la definición de función estrictamente unimodal se puede demostrar (ver, figura A.2) que los intervalos $[\alpha_k, \bar{\alpha}_k]$ contiene α^* y que sus longitudes convergen a cero. En la practica, el calculo es terminado cuando $(\bar{\alpha}_k - \alpha_k)$ llega a ser menor que una tolerancia prestablecida. Un importante hecho, que se apoya

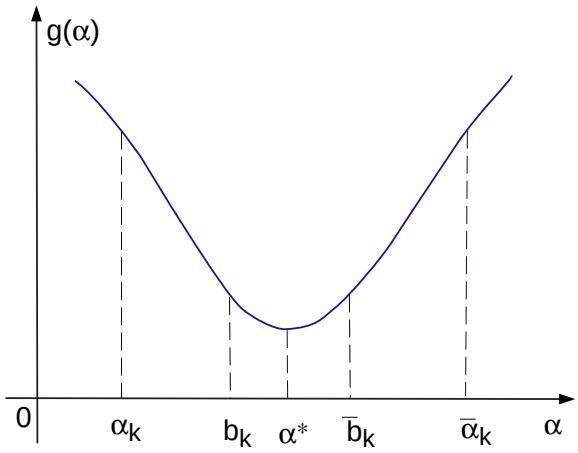

Figura A.2: Método de la sección dorada

en la elección de el numero particular τ es que si $[\alpha_{k+1}, \bar{\alpha}_{k+1}] = [\alpha_k, \bar{b}_k]$, entonces $\bar{b}_{k+1} = b_k$ y si $[\alpha_{k+1}, \bar{\alpha}_{k+1}] = [b_k, \bar{\alpha}_k]$, entonces $b_{k+1} = \bar{b}_k$.

En otras palabras, un punto de prueba b_k o $\bar{b}_k$ que no sea usado como el punto final del siguiente intervalo será un punto de prueba para la siguiente iteración. Esto se puede verificar usando la propiedad

$$\tau = (1 - \tau)^2.$$

Por lo tanto en cualquiera de las dos situaciones anteriores, los valores $\bar{b}_{k+1}$, $g(\bar{b}_{k+1})$ o b_{k+1}, $g(b_{k+1})$ son disponibles y necesariamente no son recalculados para la siguiente iteración, necesitando una única evaluación de la función en lugar de dos.

Apéndice B

Búsqueda de reducción del intervalo

Sea $g : [0,1] \longrightarrow \mathbb{R}$ una función diferenciable convexa. Supongamos que $g(0) = 0$, $g'(0) < 0$ y $g'(1) > 0$. Vamos a estudiar el problema de encontrar $\theta \in [0,1]$ tal que $g(\theta) \leq 0$ y $g'(\theta) \geq 0$.

Esto puede ser visto como un problema de búsqueda lineal con una condición de Armijo con parámetro 0 y una condición de curvatura con parámetro 0, para el cual hay algoritmos (véase, por ejemplo [11]).

En el presente caso, en el que la función es convexa, se puede usar un simple algoritmo de reducción de intervalo dado en[3], el cual esta basado en el siguiente paso:
Suponga que tres puntos $0 \leq A < \nu < B \leq 1$ son dados satisfaciendo que $g(A) \geq g(\nu) \leq g(B)$. Notemos que como consecuencia de la convexidad de $g(\cdot)$, el intervalo [A,B] contiene un minimizador de g y $g'(B) \geq 0$. El problema puede ser resuelto por el siguiente algoritmo de reducción del intervalo:

Algoritmo B.1. Algoritmo de Reducción de intervalo.
Dado $g(B) > 0$.

Elegir $\xi \in [0,1]$, $\xi \neq \nu$.
Tomar $u = \text{mín}\{\nu, \xi\}$, $v = \text{máx}\{\nu, \xi\}$.
Si $g(u) \leq g(v)$ tomar $B = v$, $\nu = u$, si no tomar $A = u$, $\nu = v$.

El valor inicial de ν y los valores de ξ en cada iteración pueden ser tales que u y v definan una sección dorada para el intervalo $[A, B]$ y entonces la longitud del intervalo será reducida por un factor de $\frac{(\sqrt{5}-1)}{2} \approx 0{,}62$ en cada iteración. Se puede además elegir ξ como el minimizador de la función cuadrática a través de $g(A)$, $g(\nu)$, $g(B)$. En este caso debemos evitar que $\xi = \mu$, donde ξ debe ser perturbado.

Conclusiones

- Para la aplicación del método propuesto por Gonzaga y Karas [3] no se requiere del conocimiento de la constante de Lipschitz L para el gradiente de $f(\cdot)$ lo que proporciona una ventaja respecto al método propuesto por Nesterov [8] ya que dicha constante es difícil de calcular la mayoría de veces.

- El método propuesto por Gonzaga y Karas genera más costo computacional por iteración respecto al método de Nesterov puesto que requiere de una búsqueda lineal inexacta en la dirección de d^k para calcular θ_k, lo cual necesita evaluaciones adicionales de la función objetivo.

- La sucesión $(f(x^k))$ es monótona decreciente para los iterados x^k dados por el método propuesto por Gonzaga y Karas debido a la elección de θ_k y los pasos de Cauchy en cambio en el método de Nesterov no se asegura.

- Para ambos métodos se utiliza las propiedades locales de los pasos de Cauchy y se aprovecha al mismo tiempo las propiedades globales de las funciones convexas.

- Con el objetivo de obtener una de las hipótesis del lema 3.1.1 se generan diferentes algoritmos para la elección del parámetro θ_k, los cuales dependen del conocimiento de las constantes L y μ. En consecuencia existen otras formas diferentes de las propuestas por Nesterov y Gonzaga y Karas para la elección de θ_k.

- En el algoritmo dado por Nesterov se puede recalcular α_k usando la propuesta por Gonzaga y Karas y manteniendo al mismo tiempo la elección de

θ_k propuesta por el mismo, lo cual asegura la complejidad óptima.

- La mejora de Gonzaga y Karas mantiene el orden de complejidad $O(\frac{1}{k^2})$ obtenido por Nesterov.

Bibliografía

[1] D. Bertsekas. *Nonlinear Optimization.* Athenas Scientific, Nashua, USA, 1999.

[2] D. Bertsekas, A. Nedić, and A. E. Ozdagar. *Convex Analysis and Optimization.* Athenas Scientific, Belmont, USA, 2003.

[3] C. Gonzaga and E. Karas. Optimal steepest descent algorithms for unconstrained convex problems: fine tuning Nesterov's method. *Optimization online*, 2009.

[4] Urruty. Hiriart and C. Lemarechal. *Convex Analysis and Minimization Algorithm I.* Springer Verlag, New York, 1996.

[5] David G. Luenberger. *Linear and Nonlinear Programming.* Kluwer Academic Publishers, USA, 2004.

[6] A. S. Nemirovski and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization.* John Wiley, New York, 1983.

[7] Y. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o\left(\frac{1}{k^2}\right)$. *Doklady AN SSSR*, 259:543–547, 1983.

[8] Y. Nesterov. *Introductory lectures in convex optimization. A basic course.* Kluwer Academic Publishers, Boston, 2004.

[9] Kenneth Ross. *Elementary Analysis.* Springer- Verlag, New York, 1980.

[10] N. Shor. *Minimization Methods for Non-Differenciable Functions*. Springer Verlag, Berlin, 1985.

[11] S. Wright and J. Nocedal. *Numerical optimization*. Springer series in operations research and financial engineering. Springer, 2^{nd} edition, 2006.