Optimización Continua

Daniel Espinoza, Dpto. Ingeniería Industrial

Universidad de Chile

7 de marzo de 2011

Contenidos

1. Introducción
2. Optimización no constreñida
3. Optimización con Restricciones

Buscando Problemas más generales

- Pensemos en el problema
$$\min f(x) : x \in \mathbb{R}^n,$$
donde $f : \mathbb{R}^n \to \mathbb{R}$.
- Podemos decir algo si f no es continua?
  - Ejemplo: $f(x) = \begin{cases} x^2 & \text{si } x \in \mathbb{Q} \\ x^2 - \pi & \text{si } x \in \mathbb{R} \setminus \mathbb{Q} \end{cases}$.
  - Cómo podemos asegurar si algún punto es óptimo (global o local)?
- Necesitamos algún tipo de regularidad de f para poder resumir su comportamiento.
- Algunas condiciones comunes son $f \in \mathcal{C}^1$ o $f \in \mathcal{C}^2$.
  - Veremos que bajo los supuestos anteriores podremos dar condiciones suficientes y necesarias para optimalidad local.
- Para asegurar optimalidad global necesitamos condiciones globales.
  - Convexidad de f asegurará mínimos locales son mínimos globales.

Anticipando el comportamiento de f

Teorema de Taylor

Sea $f : \mathbb{R}^n \to \mathbb{R}, x, p \in \mathbb{R}^n$.

- Si $f \in \mathcal{C}^1(\mathbb{R}^n)$ entonces $f(x+p) = f(x) + \nabla f(x+tp)^t p$, para algún $t \in (0,1)$.
- Si $f \in \mathcal{C}^2(\mathbb{R}^n)$ entonces

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

para algún $t \in (0,1)$.

- El teorema de Taylor permite anticipar el comportamiento de f en torno a un punto dado $x \in \mathbb{R}^n$.
- Si asumimos optimalidad local de x, entonces podemos deducir condiciones sobre $\nabla f(x)$ y sobre $\nabla^2 f(x)$.

Condiciones Necesarias de Optimalidad

Teorema de condiciones de primer orden

Si $x^* \in \mathbb{R}^n$ es un mínimo local y $f \in \mathcal{C}^1(B(x^*, \varepsilon))$, entonces $\nabla f(x^*) = 0$.

Demostración.

- Si x^* es mínimo local, $\exists \varepsilon \geq 0 : f(x^*) \leq f(x), \forall x \in B(x^*, \varepsilon)$.
- Por Taylor $f(x) = f(x^*) + \nabla f(x^* + t(x - x^*))^t (x - x^*)$.
- Entonces $\forall x \in B(x^*, \varepsilon)$ existe $t \in (0, 1)$ tal que $\nabla f(x^* + t(x - x^*))^t (x - x^*) \geq 0$.
- Tomando limite de $x \to x^*$ y por continuidad de $\nabla f(x)$, tenemos que $\nabla f(x^*) = 0$

$\square$

Definición Punto estacionario

Para $f \in \mathcal{C}^1$, x se dice un punto estacionario de f si $\nabla f(x) = 0$.

Condiciones Suficientes

Teorema condiciones necesarias de segundo orden

Si x^* es un mínimo local de f y $f \in \mathcal{C}^2(B(x^*, \varepsilon))$ entonces $\nabla f(x^*) = 0$ y $\nabla^2 f(x^*) \geq 0$.

Demostración.

- Por las condiciones de primer orden $\nabla f(x^*) = 0$
- Sea $p \in \mathbb{R}^n \setminus \{0\}$, entonces $f(x^* + \varepsilon p) = f(x^*) + \varepsilon^2 \frac{1}{2} p^t \nabla^2 f(x^* + \tilde{\varepsilon} p) p$ para algún $0 < \tilde{\varepsilon} < \varepsilon$.
- Por mínimo local se tiene $0 \leq \frac{1}{2} p^t \nabla^2 f(x^* + \tilde{\varepsilon} p) p$, y tomando límite de $\varepsilon \to 0$ $p^t \nabla^2 f(x^*) p \geq 0$.

$\square$

Teorema condiciones suficientes de segundo orden

Si $f \in \mathcal{C}^2(B(x^*, \varepsilon))$, $\nabla f(x^*) = 0$ y $\nabla^2 f(x^*) > 0$, entonces x^* es un óptimo local.

De óptimo local a óptimo global

Demostración.

Use el argumento de que $\nabla^2 f(x)$ es continua en torno a x^*. □

Teorema óptimos globales para funciones convexas

Si f es una función convexa, cualquier mínimo local es un mínimo global. Si además $f \in \mathcal{C}^1$, entonces cualquier punto estacionario es un óptimo global.

Demostración.

- Para la primera parte, razone por contradicción.
- Para la segunda, use que $\nabla f(x^*)(z - x^*) = \lim\limits_{\lambda \to 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda}$.

□

Algoritmo Básico

- Dado los teoremas anteriores sólo podemos:
  - Encontrar óptimos locales si $\nabla^2 f(x^*) > 0$.
  - Buscar puntos estacionarios.
  - En el caso de f convexa, tenemos óptimos globales.
- La mayoría de los algoritmos buscan condición necesaria de primer orden:
  - $\nabla f(x^*) = 0$.
  - En la práctica buscamos $\|\nabla f(x^*)\| \leq \varepsilon$.
  - ¿Qué podemos asegurar en esas condiciones?

Algoritmos de descenso

Require: $x^o \in \mathbb{R}^n$, $f \in \mathcal{C}^1(\mathbb{R}^n)$, $\varepsilon > 0$.

$k \leftarrow 0$, $\delta^k \leftarrow \|\nabla f(x^k)\|$.

while $\delta^k > \varepsilon$ **do**

    Buscar $x^{k+1} : f(x^k) > f(x^{k+1})$.

    $\delta^{k+1} \leftarrow \|\nabla f(x^{k+1})\|$.

    $k \leftarrow k + 1$.

return $x^* = x^k$.

Buscando x^{k+1}

- Sabemos que $f(x+d) = f(x) + \nabla f(x+\lambda d)^t d$ para algún $\lambda \in (0,1)$.
- Por continuidad de $\nabla f(x)$, para d suficientemente pequeño, sabemos que $\nabla f(x) \approx \nabla f(x+\lambda d)$.
- Podríamos aproximar $f(x+d) \approx f(x) + \nabla f(x)^t d$.
- Bajo este supuesto, si escogemos $x^{k+1} = x^k + d$ tenemos como condición necesaria $\nabla f(x)^t d < 0$.

Definición dirección de descenso

Dada $f \in \mathcal{C}^1, x \in \mathbb{R}^n, \nabla f(x) \neq 0$ se dice que $d \in \mathbb{R}^n$ es una dirección de descenso si $\nabla f(x)^t d < 0$.

Teorema

Para cualquier dirección de descenso d, existe $\alpha > 0$ tal que $x' := x + \alpha d$ satisface $f(x') < f(x)$.

Buscando x^{k+1}

- Dado d dirección de descenso, ¿Cómo buscamos α?

Algoritmo de búsqueda binaria exacta:

Require: $x, d \in \mathbb{R}^n$, $\varepsilon > 0$, f convexa.

$\alpha_{\text{mín}} \leftarrow 0$, $\alpha_{\text{máx}} \leftarrow 1$
while $f(x + \alpha_{\text{máx}} d) < f(x)$ **do**
  $\alpha_{\text{máx}} \leftarrow 2\alpha_{\text{máx}}$.
while $|\alpha_{\text{mín}} - \alpha_{\text{máx}}| > \varepsilon$ **do**
  $x_1 \leftarrow x + \frac{1}{3}(\alpha_{\text{máx}} - \alpha_{\text{mín}})d$, $x_2 \leftarrow x + \frac{2}{3}(\alpha_{\text{máx}} - \alpha_{\text{mín}})d$.
  if $f(x_1) > f(x_2)$ **then**
    $\alpha_{\text{mín}} \leftarrow \frac{1}{3}(\alpha_{\text{máx}} - \alpha_{\text{mín}})$.
  else
    $\alpha_{\text{máx}} \leftarrow \frac{2}{3}(\alpha_{\text{máx}} - \alpha_{\text{mín}})$.
return $\alpha_{\text{mín}}$.

- ¿Por qué el algoritmo es correcto?
- ¿Y si f no es convexa?
- ¿Podríamos terminar antes?

Buscando x^{k+1}

- El algoritmo anterior resuelve $\min f(x + \alpha d)$
  - Existen otros algoritmos para caso no convexo.
- ¿Cómo escogemos d?
  - Podemos hacer un análogo a simplex:
  - Buscar dirección de descenso máximo
  - $\min \nabla f(x)^t d : \|d\| = 1$, i.e. $d = -\nabla f(x)/\|\nabla f(x)\|$.
  - Note que dada $D \in \mathbb{R}^{n\times n} > 0$, podemos escoger $d = -D\nabla f(x)$.
- Si $d = -\nabla f(x)/\|\nabla f(x)\|$, el algoritmo se llama algoritmo de máximo descenso.
  - Este algoritmo es simple, y si f es convexa, existe óptimo, y se usa búsqueda exacta, converge a alguna solución óptima.
  - Convergencia es lenta para problemas muy elipsoidales...
  - Tasa de convergencia lineal $\|x_{k+1} - x^*\|/\|x_k - x^*\| \leq M < 1$.
- ¿Qué alternativa tenemos?
  - Lo anterior se derivó usando aproximación de Taylor de primer orden.
  - ¿Y si consideramos una aproximación de segundo orden?

Método de Newton

- Por Taylor de segundo orden:
  $f(x+d) = f(x) + \nabla f(x)^t d + \frac{1}{2} d^t \nabla^2 f(x+\lambda d) d$ para algún $\lambda \in (0,1)$
- Por continuidad, y para d pequeños, aproximamos $\nabla^2 f(x+\lambda d) = \nabla^2 f(x)$.
- Asumimos $\nabla^2 f(x) > 0$ y minimizamos $f(x+d) = f(x) + \nabla f(x)^t d + \frac{1}{2} d^t \nabla^2 f(x) d$.
  - Óptimo es $d = -\left(\nabla^2 f(x)^{-1}\right) \nabla f(x)$.
  - Note que $d^t \nabla f(x) = -\nabla f(x)^t \left(\nabla^2 f(x)\right)^{-t} \nabla f(x)$.
  - i.e. Si $\nabla^2 f(x) > 0$, d es dirección de descenso.
- Si la aproximación cuadrática de f es buena, deberíamos poder usar $\alpha^k = 1$, a esto se le llama iteración de Newton pura.
- Si además se hace una búsqueda unidireccional (i.e. buscamos α^k), se le llama iteración de Newton.
- ¿Cómo se interpreta d?

Método de Newton

- Ventajas de Newton:
  - Para funciones cuadráticas y estrictamente convexas, Algoritmo de descenso de Newton termina en una iteración.
  - Convergencia cuadrática: $\|x_{k+1} - x^*\| / \|x_k - x^*\|^2 \leq M$.
  - En la práctica muy pocas iteraciones.
- Desventajas de Newton:
  - Computar $\nabla^2 f(x^k)$ en cada iteración.
  - Luego invertir $\nabla^2 f(x^k)$.
  - Muy caro computacionalmente.
- ¿Podemos mejorar?
  - Existen varios métodos para aproximar $\nabla^2 f(x)$:
  - Sólo computar $\partial^2 f(x)/\partial x_i^2$ y usar matriz diagonal asociada.
  - Usar $\nabla f(x^k)$ anteriores para estimar $\nabla^2 f(x^k)$.
  - Computar $\nabla^2 f(x)$ cada m iteraciones....
- Estos métodos se llaman Métodos Quasi-Newton.
  - Si f estrictamente convexa, existen soluciones óptimas, y se usa búsqueda exacta, converge a alguna solución óptima.

Definiciones Básicas

Problema bajo dominio restringido:

Consideramos problemas del tipo:

$$(P): \qquad \min f(x) : x \in \Omega \subseteq \mathbb{R}^n$$

donde $f : \Omega \to \mathbb{R}$, $f \in \mathcal{C}^1(\Omega)$, Ω es un conjunto cerrado. Todo $x \in \Omega$ se dice factible para (P).

Definición Óptimo Global

$x^* \in \Omega$ es un óptimo global para (P) si $f(x^*) \leq f(x)$, $\forall x \in \Omega$.

Definición Óptimo Local

$x^* \in \Omega$ es un óptimo local para (P) si $\exists \varepsilon > 0$ tal que $f(x^*) \leq f(x)$, $\forall x \in \Omega \cap B(x^*, \varepsilon)$.

Resultados Básicos

Definición Problema Convexo

(P) se dice un problema convexo si f es función convexa y si Ω es conjunto convexo.

Teorema de Óptimos Globales para Problemas Convexos

Si (P) es un problema convexo, entonces todos los óptimos locales para (P) son óptimos globales para (P).

- ¿Qué debemos pedir de Ω?
    - Conexidad.... ¿si no?
    - Regularidad (¿Cómo es Ω en torno a x_o?)... ¿si no?
    - En términos de complejidad.... ¿Cómo representamos Ω?
        - Debemos ser capaces de responder $x \in \Omega$ de forma rápida.
    - Una forma clásica es definir

$$\Omega := \{x \in \mathbb{R}^n : g_i(x) \leq 0, \ \forall i \in I\},$$

donde $g_i : \mathbb{R}^n \to \mathbb{R}, g_i \in \mathcal{C}^1(\mathbb{R}^n)$ y donde $|I| = m \in \mathbb{Z}$.

Resultados Básicos

Conjuntos Convexos a partir de Funciones Convexas

Sea $\Omega := \{x \in \mathbb{R}^n : g_i(x) \leq 0, i = 1, \ldots, m\}$, si g_i es convexa en $\mathbb{R}^n$, entonces Ω es un conjunto convexo.

- ¿Cómo caracterizamos optimalidad local sin tener que evaluar todos los vecinos?

Definición de restricciones activas en un punto

Dado $x^o \in \Omega := \{x \in \mathbb{R}^n : g_i(x) \leq 0 \forall i \in I\}$, una restricción $i \in I$ se dice activa en x^o si $g_i(x^o) = 0$. Se define el conjunto de restricciones activas en x^o como $I(x^o) := \{i \in I : g_i(x^o) = 0\}$.

Definición de direcciones factibles en un punto

Dado $x^o \in \Omega := \{x \in \mathbb{R}^n : g_i(x) \leq 0 \forall i \in I\}$, $d \in \mathbb{R}^n$ se dice una dirección factible en x^o, si existe $\varepsilon > 0$ tal que $x^o + \varepsilon' d \in \Omega$ para todo $0 \leq \varepsilon' \leq \varepsilon$. Se define el conjunto de direcciones factibles en x^o como $\mathcal{D}(x^o) := \{d \in \mathbb{R}^n : d \text{ dirección factible para } x^o\}$.

Optimalidad y direcciones factibles

- Para definir el algoritmo de Simplex, redujimos la condición de optimalidad en una vecindad a una condición con las direcciones factibles, ¿podemos hacer lo mismo en nuestro caso?

Condición Necesaria de Direcciones Factibles

Si x^* es un óptimo local, entonces $\nabla f(x^*)^t d \geq 0$ para todo $d \in \mathcal{D}$.

Demostración.

Use definicion de óptimo local y aproximación de Taylor de primer orden. $\square$

- El opuesto no es cierto en general, es decir, si x^* es tal que existe $\varepsilon > 0$ tal que $f(x^*) \leq f(x^* + d)$, $\forall d \in \mathcal{D}(x^*) \cap B(0, \varepsilon)$, x^* no es necesariamente óptimo local.
  - Ejemplo: Considere $\Omega = \{x \in \mathbb{R}^2 : -x_2 + x_1^5 \leq 0\}$ y $f(x_1, x_2) = x_2$.
  - Note que $f \in \mathcal{C}^2(\mathbb{R}^2)$ y $g_1 \in \mathcal{C}^1(\mathbb{R}^2)$.
  - El punto $(0, 0)$ satisface condición anterior pero no es óptimo local.
- Note que si Ω y f son convexas, esto es imposible!

Caracterizando Direcciones Factibles y Óptimos Locales

- ¿Cómo caracterizamos $\mathcal{D}(x)$ en general?
- Buscamos una forma sucinta para esto.

Sobre-estimando $\mathcal{D}(x)$

Consideremos $\widetilde{\mathcal{D}}(x) := \{d \in \mathbb{R}^n : \nabla g_i(x)^t d \leq 0 \forall i \in I(x)\}$.

- Demuestre que $\mathcal{D}(x) \subseteq \widetilde{\mathcal{D}}(x)$.
- Note que si Ω es convexo y $\nabla g_i(x) \neq 0 \forall i \in I(x)$, entonces $\overline{\mathcal{D}(x)} = \widetilde{\mathcal{D}}(x)$.

¿Qué nos gustaría?

Nos gustaría decir que si x^* es óptimo local entonces $\{d \in \mathbb{R} : \nabla f(x^*)^t d < 0\} \cap \widetilde{\mathcal{D}}(x) = \emptyset$

- En general no es cierto.
- Si Ω, f convexas y $\nabla g_i(x) \neq 0 \forall i \in I(x)$, entonces sí es cierto.

Calificación de Restricciones

Definición calificación de restricciones

Se dice que $x \in \Omega$ satisface las condiciones de calificación de restricciones (CCR) si $\overline{\mathcal{D}(x)} = \widetilde{\mathcal{D}}(x)$.

Teorema Condiciones necesarias de Primer Orden

Sea $x^* \in \Omega$ óptimo local satisfaciendo CCR, entonces existe $\lambda \in \mathbb{R}^I_+$ tal que :

1. $\nabla f(x^*) + \sum(\lambda_i \nabla g_i(x^*) : i \in I) = 0.$
2. $\lambda_i g_i(x^*) = 0 \forall i \in I.$

- El teorema anterior se debe a Karush-Kuhn-Tucker, 1939, y se denominan condiciones de KKT.

Demostración.

Use CCR, el teorema anterior y Lema de Farkas. □