

Optimización

Escuela de Ingeniería Informática de Oviedo

Introducción

Problema general de optimización (minimización)

Dado $f : \Omega \subset \mathbb{R}^n \rightarrow \mathbb{R}$ find $x^* \in \Omega$ tal que $f(x^*) \leq f(x)$ para todo $x \in \Omega$.
 f se llama **función objetivo**, y Ω el **conjunto de soluciones posibles**.

Escribimos

$$\min_{x \in \Omega} f(x).$$

Casos principales

- Optimización sin constricciones: $\Omega = \mathbb{R}^n$
- Optimización con constricciones: $\Omega \subsetneq \mathbb{R}^n$, habitualmente determinada por un conjunto de constricciones dadas por igualdades o desigualdades,

$$h(x) = 0, \quad g(x) \leq 0, \quad \text{etc.}$$

Hecho:

No hay técnicas generales para resolver el problema de optimización global.

Por lo tanto, habitualmente se resuelve de forma débil.

Optimización local

Encontrar $x^* \in \Omega$ tal que $f(x^*) \leq f(x)$ para todo x tal que $\|x - x^*\| \leq R$,

Excepción

Si f es una función estrictamente convexa y Ω es un conjunto estrictamente convexo, entonces f tiene un mínimo local y único (y global) en Ω .

Repaso de la teoría de optimización local

- **En una variable:** Resolver el problema de optimización para $f : \mathbb{R} \rightarrow \mathbb{R}$:
 - Encontrar el conjunto de puntos críticos x_c ($f'(x_c) = 0$).
 - Si $f''(x_c) > 0$ entonces x_c es un mínimo local.

- **n variables:** Resolver el problema de optimización para $f : \mathbb{R}^n \rightarrow \mathbb{R}$:
 - Encontrar los puntos críticos x_c , que satisfacen $\nabla f(x_c) = 0$, es decir,

$$\partial_{x_1} f(x_c) = 0, \quad \partial_{x_2} f(x_c) = 0, \dots, \partial_{x_n} f(x_c) = 0$$

- Calcular el Hessiano en x_c

$$H(f)(x_c) = \left(\partial_{x_i x_j} f(x_c) \right)_{i,j=1}^n.$$

Si esta matriz es definida positiva, x_c es un mínimo local.

Métodos de descenso

Señalar que:

Encontrar un mínimo local es, en general, más fácil que el problema de resolver ecuaciones no lineales

$$g(x^*) = \nabla f(x^*) = 0$$

porque

- Podemos evaluar f , además de ∇f ,
- La matriz Hessiana es definida positiva cerca de la solución.

Métodos de descenso

Si tenemos un punto x^k , y conocemos una **dirección de descenso** d^k , es decir, una dirección en la cual

$$f(x^k + \alpha d^k) < f(x^k) \quad \text{for all } 0 < \alpha \leq \alpha_{\max},$$

entonces podemos descender y avanzar un punto hacia el mínimo:

$$x^{k+1} = x^k + \alpha_k d^k,$$

donde α_k es una **longitud de paso**.

Método del gradiente

Si la función es diferenciable podemos usar la fórmula de Taylor

$$f(x^k + \alpha_k d^k) \approx f(x^k) + \alpha_k (\nabla f)^T d^k = f(x^k) + \alpha_k \sum_{i=1}^n \partial_{x_i} f(x^k) d_i^k.$$

Esto significa que el **decremento local más rápido** para la función objetivo se consigue cuando nos movemos en dirección opuesta al gradiente en **la dirección de máxima pendiente**

$$d^k = -\nabla f(x^k) = -g_k.$$

Una forma para escoger la longitud de paso es realizar una minimización sobre una línea unidimensional:

$$\alpha_k = \arg \min_{\alpha} f(x^k + \alpha d^k),$$

de la que solo necesitamos **una solución aproximada**.

Método del gradiente

Si suponemos que tenemos la línea exacta $\alpha_k = \arg \min_{\alpha} \phi(\alpha)$, donde $\phi(\alpha) = f(x^k + \alpha d^k)$. Entonces tenemos que resolver

$$\phi'(\alpha) = \nabla f(x^k + \alpha d^k)^T d^k = 0,$$

lo que implica

$$\nabla f(x^k + \alpha d^k)^T \perp d^k = -\nabla f(x^k).$$

Esto supone que el método sigue una **trayectoria en zigzag** hacia el mínimo.

Error

El método del gradiente tiene convergencia lineal con coeficiente de convergencia

$$C \propto \frac{1-r}{1+r}, \quad \text{donde } r = \frac{\lambda_{\min}(H)}{\lambda_{\max}(H)} = \frac{1}{\kappa_2(H)},$$

es inversamente proporcional al número de condición de la matriz Hessiana.

Por lo tanto, el método del gradiente puede tener una convergencia muy lenta si la matriz Hessiana está mal condicionada.

Método de Newton

Sea la fórmula de Taylor de orden dos de f

$$f(x^k + \Delta x) = f(x^k) + \nabla f(x^k)^T \Delta x + \frac{1}{2}(\Delta x)^T H(x^k) \Delta x,$$

con $\Delta x = x - x^k$. Tenemos un extremo cuando el diferencial con respecto a Δx es cero, es decir, cuando

$$\nabla f(x^k)^T + H(x^k) \Delta x = 0.$$

Entonces

$$\Delta x = -H(x^k)^{-1} \nabla f(x^k)^T \Rightarrow x^{k+1} = x^k - H(x^k)^{-1} \nabla f(x^k)^T.$$

Señalar que:

- **El método es exacto para funciones objetivo cuadráticas.** En este caso $H(x)$ es constante.
- Equivale a usar el método de Newton-Raphson para resolver el sistema no lineal $\nabla f(x^*) = 0$.

Problemas del método de Newton

- Para funciones objetivo no lineales, el método de Newton necesita resolver un sistema lineal en cada paso **costoso**.
- Puede no converger si el punto inicial no es adecuado, o puede converger a un punto de silla o a un máximo: **no fiable**.
- Para tratar de salvar estas dificultades se usan variantes o **métodos de cuasi-Newton**:

$$x^{k+1} = x^k - \alpha_k H_k^{-1} \nabla f(x^k)^T,$$

donde $0 < \alpha_k < 1$ y H_k es una aproximación de la matriz Hessiana.

Formulación general

Problema general de optimización con restricciones

Dado $f : \mathbb{R}^n \rightarrow \mathbb{R}$, encontrar $x \in \mathbb{R}^n$ que satisfaga

$$\min_{x \in \mathbb{R}^n} f(x),$$

$$h(x) = 0 \quad (\text{restricciones igualdad}),$$

$$g(x) \leq 0 \quad (\text{restricciones desigualdad}).$$

Asumimos que las funciones f , g y h son suaves.

Multiplicadores de Lagrange: una única igualdad

Una restricción igualdad $h(x) = 0$ corresponde a una superficie de $n - 1$ dimensiones cuyo vector normal es ∇h .

Para una única restricción igualdad suave, el gradiente de la función objetivo debe ser paralelo al vector normal a la superficie de restricción:

$$\nabla f \parallel \nabla h \Rightarrow \text{existe un } \lambda \text{ tal que } \nabla f + \lambda \nabla h = 0,$$

donde λ es el **multiplicador de Lagrange** correspondiente a la restricción $h(x) = 0$.

Para calcular el mínimo hemos de tener en cuenta las dos ecuaciones:

$$\begin{aligned}\nabla f + \lambda \nabla h &= 0 \\ h(x) &= 0\end{aligned}$$

Multiplicadores de Lagrange: m igualdades

Cuando tenemos m igualdades,

$$h_1(x) = h_2(x) = \dots = h_m(x),$$

generalizamos haciendo que la dirección de descenso $-\nabla f$ debe estar en el subespacio generado por los vectores normales a las restricciones:

$$\nabla f + \sum_{i=1}^m \lambda_i \nabla h_i = \nabla f + (\nabla h)^T \lambda = 0,$$

El Jacobiano tiene como vectores fila los vectores normales

$$\nabla h = \left(\frac{\partial h_i}{\partial x_j} \right)_{ij}.$$

Esta es una condición necesaria de óptimo de primer orden.

Multiplicadores de Lagrange: única desigualdad

En la solución x^* , dada una restricción desigualdad $g_i(x) \leq 0$ esta puede ser

$$\begin{array}{ll} \text{activa,} & \text{si } g_i(x^*) = 0, \\ \text{inactiva,} & \text{si } g_i(x^*) < 0. \end{array}$$

Para desigualdades, existe un signo o sentido para los vectores normales a la restricción:

para una restricción activa, te puedes mover en el sentido de $-\nabla g$ pero no en el sentido de ∇g .

Esto significa, que en este caso

$$\nabla f = -\mu \nabla g, \quad \text{donde } \mu > 0.$$

Multiplicadores de Lagrange: r desigualdades

La generalización es la misma que para las igualdades

$$\nabla f + \sum_{i=1}^r \mu_i \nabla g_i = \nabla f + (\nabla g)^T \mu = 0,$$

pero con la condición

$$\begin{aligned} \mu_i &= 0 && \text{para restricciones inactivas,} \\ \mu_i &> 0 && \text{para restricciones activas.} \end{aligned}$$

Poniendo igualdades y desigualdades juntas tenemos la **condición necesaria de primer orden de Karush-Kuhn-Tucker**: Existen los multiplicadores de Lagrange $\lambda \in \mathbb{R}^m$ y $\mu \in \mathbb{R}^r$ tal que:

$$\nabla f + (\nabla h)^T \lambda + (\nabla g)^T \mu = 0, \quad \mu \geq 0 \quad \text{y} \quad \mu^T g(x) = 0.$$

Función Lagrangiana

Podemos reescribir la ecuación

$$\nabla f + (\nabla h)^T \lambda + (\nabla g)^T \mu = 0$$

como

$$\nabla_x L = 0,$$

donde L es la **función Lagrangiana**

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

o, en notación vectorial,

$$L(x, \lambda, \mu) = f(x) + \lambda^T h(x) + \mu^T g(x).$$

Constricciones igualdad

Las condiciones necesarias de primer orden para problemas de restricción igualdad vienen, por lo tanto, dadas por las condiciones estacionarias:

$$\begin{aligned}\nabla_x L(x^*, \lambda^*) &= \nabla f(x^*) + \nabla h(x^*)^T \lambda^* = 0, \\ \nabla_\lambda L(x^*, \lambda^*) &= h(x^*) = 0.\end{aligned}$$

Tener en cuenta que también hay **condiciones necesarias y suficientes de segundo orden** similares a las de la optimización sin restricciones.

La solución no es un máximo ni un mínimo de la función Lagrangiana (de hecho, para problemas convexos es un punto de silla, mínimo para x y máximo para λ).

El enfoque por penalización

La idea es convertir un problema con constricciones:

$$\begin{aligned} \min_{x \in \mathbb{R}^n} f(x), \\ h(x) = 0, \end{aligned}$$

en un problema sin constricciones, minimizando la función con un término de penalización

$$L_\alpha(x) = f(x) + \alpha \|h(x)\|_2^2 = f(x) + \alpha h(x)^T h(x),$$

donde α es un **parámetro de penalización**.

Se pueden usar **funciones de penalización** que no sean una suma de cuadrados.

Si la restricción se cumple de forma exacta, entonces $L_\alpha(x) = f(x)$. Si $\alpha \rightarrow \infty$ el no cumplir la restricción se penaliza cada vez más, y la igualdad debe ser satisfecha con mayor precisión.

Método de penalización

Teniendo en cuenta lo anterior: para una sucesión monótona divergente $\alpha_1 < \alpha_2 < \dots$, si resolvemos una sucesión de problemas sin restricción

$$x^k = x(\alpha_k) = \arg \min_x \left\{ L_k(x) = f(x) + \alpha_k h(x)^T h(x) \right\},$$

y la solución debería converger al óptimo x^* ,

$$x_k \rightarrow x^* = x(\lim_{k \rightarrow \infty} \alpha_k).$$

Podríamos usar x^k como punto inicial, por ejemplo, del método de Newton. Tener en cuenta que el problema se vuelve peor condicionado cuando α crece.

Un enfoque mejor usa los multiplicadores de Lagrange en conjunción con el método de penalización (Lagrangiana aumentada).