

Análisis Teórico y Métodos de Optimización Aplicados a

$$f(x, y) = y^2 + \log(1 + x^2)$$

Alina María de la Noval Armenteros
Grupo: C-311

1. Modelo a analizar

Se considera el problema irrestricto de optimización en \mathbb{R}^2 :

$$\min_{(x,y) \in \mathbb{R}^2} f(x, y), \quad f(x, y) = y^2 + \log(1 + x^2).$$

El dominio natural de la función es \mathbb{R}^2 . Observamos que la función se puede descomponer como suma de funciones univariadas $f(x, y) = g(x) + h(y)$ con $g(x) = \log(1 + x^2)$ y $h(y) = y^2$, lo que simplifica tanto el análisis teórico como el tratamiento numérico.

2. Análisis de los modelos

2.1. Regularidad y diferenciabilidad

Dado que $1 + x^2 > 0$ para todo $x \in \mathbb{R}$, la función $g(x) = \log(1 + x^2)$ es de clase C^∞ en \mathbb{R} . El término $h(y) = y^2$ es un polinomio y, por tanto, también pertenece a C^∞ . En consecuencia, $f \in C^\infty(\mathbb{R}^2)$: existen derivadas de cualquier orden y son continuas en todo \mathbb{R}^2 . Esta regularidad justifica el uso de métodos de optimización que requieren gradiente y Hessiano.

2.2. Gradiente y Hessiano

Las derivadas parciales primeras son

$$\frac{\partial f}{\partial x}(x, y) = \frac{2x}{1 + x^2}, \quad \frac{\partial f}{\partial y}(x, y) = 2y.$$

Por tanto, el gradiente se escribe

$$\nabla f(x, y) = \begin{pmatrix} \frac{2x}{1 + x^2} \\ 2y \end{pmatrix}.$$

Las segundas derivadas son

$$\frac{\partial^2 f}{\partial x^2}(x, y) = \frac{2(1-x^2)}{(1+x^2)^2}, \quad \frac{\partial^2 f}{\partial y^2}(x, y) = 2, \quad \frac{\partial^2 f}{\partial x \partial y}(x, y) = 0.$$

Por tanto, el Hessiano es la matriz diagonal

$$\nabla^2 f(x, y) = \begin{pmatrix} \frac{2(1-x^2)}{(1+x^2)^2} & 0 \\ 0 & 2 \end{pmatrix}.$$

Estas expresiones se utilizarán para clasificar puntos estacionarios y para determinar regiones de convexidad, así como para justificar la aplicabilidad de métodos de primer y segundo orden.

2.3. Convexidad local y global

Una función C^2 es convexa en un dominio si su Hessiano es semidefinido positivo en dicho dominio. En nuestro caso el Hessiano es diagonal; la componente correspondiente a y es constante y positiva (igual a 2), mientras que la componente asociada a x es

$$a(x) = \frac{2(1-x^2)}{(1+x^2)^2}.$$

El signo de $a(x)$ depende de x : para $|x| < 1$ se tiene $a(x) > 0$; para $|x| = 1$ se tiene $a(x) = 0$; y para $|x| > 1$ se tiene $a(x) < 0$. Por tanto, el Hessiano deja de ser semidefinido positivo cuando $|x| > 1$. En consecuencia, f no es convexa globalmente, aunque sí es convexa en la banda $|x| \leq 1$ (estrictamente convexa en $|x| < 1$).

Esta estructura implica que muchos resultados teóricos de convergencia global que requieren convexidad total no son aplicables sin precauciones; sin embargo, la existencia y unicidad del mínimo global compensan en la práctica la falta de convexidad: los algoritmos locales bien regulados tienden a converger hacia el mínimo desde una amplia región de inicialización.

2.4. Existencia y unicidad del mínimo

Los puntos estacionarios se obtienen resolviendo $\nabla f(x, y) = 0$. Del sistema

$$\frac{2x}{1+x^2} = 0, \quad 2y = 0,$$

se deduce $x = 0$ e $y = 0$. Por tanto, el único punto estacionario es $(0, 0)$.

Evalutando el Hessiano en $(0, 0)$ se obtiene

$$\nabla^2 f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

que es definida positiva. Por la condición suficiente de segundo orden, $(0, 0)$ es un mínimo local estricto.

Para estudiar la existencia del mínimo global consideramos el comportamiento en el infinito. Cuando $|y| \rightarrow \infty$ se tiene $y^2 \rightarrow \infty$, y cuando $|x| \rightarrow \infty$ se cumple $\log(1+x^2) \rightarrow \infty$ (asintóticamente $\log(1+x^2) \sim 2 \log |x|$). Por tanto, si $\|(x, y)\| \rightarrow \infty$ entonces $f(x, y) \rightarrow +\infty$; es decir, f es coerciva.

Una función continua coerciva definida en \mathbb{R}^n alcanza al menos un mínimo global. Dado que $f(x, y) \geq 0$ para todo (x, y) y $f(0, 0) = 0$, y puesto que $(0, 0)$ es el único punto estacionario, concluimos que $(0, 0)$ es el único mínimo global de f .

3. Descripción de los Algoritmos Utilizados

3.1. Fundamento general del descenso por direcciones

Para resolver el problema de minimización de

$$f(x, y) = y^2 + \log(1 + x^2),$$

considero métodos iterativos de optimización sin restricciones que construyen una sucesión de puntos

$$x_{k+1} = x_k + \alpha_k d_k,$$

donde d_k representa la dirección de búsqueda y $\alpha_k > 0$ el tamaño de paso. El objetivo es que los valores de la función disminuyan progresivamente, acercándose a un punto estacionario donde $\nabla f(x, y) = 0$. En este contexto, el concepto esencial es la dirección de descenso. La derivada direccional permite cuantificar cómo cambia la función si me desplazo desde x_k en la dirección d_k :

$$f'(x_k; d_k) = \nabla f(x_k)^\top d_k.$$

Una dirección d_k es de descenso si $f'(x_k; d_k) < 0$; esto significa que el valor de f disminuye cuando avanzo ligeramente en esa dirección. Por tanto, la elección adecuada de d_k y α_k determina la eficiencia y estabilidad del método.

3.2. Método del gradiente descendente y condición de Armijo

El punto de partida de los métodos de primer orden es el método de máximo descenso, donde la dirección de búsqueda se elige como el negativo del gradiente:

$$d_k = -\nabla f(x_k).$$

Esto se debe a que el gradiente indica la dirección de crecimiento más rápido, por lo que avanzar en la dirección contraria garantiza decrecimiento. De hecho, se cumple que:

$$\nabla f(x_k)^\top d_k = -\|\nabla f(x_k)\|^2 < 0,$$

confirmando que siempre se trata de una dirección de descenso.

Una vez que se ha elegido una dirección de descenso d_k el siguiente paso consiste en determinar cuánto avanzar sobre esa dirección. Este tamaño de paso, denotado por $\alpha_k > 0$, cumple un papel crucial: si es demasiado pequeño, el método avanza con lentitud; si es demasiado grande, puede sobrepasar el mínimo o incluso hacer que la función aumente. Para manejar este equilibrio, se introduce la búsqueda lineal, que consiste en estudiar cómo varía la función únicamente en función de α , manteniendo fija la dirección d_k . Así, el problema original, que es multivariable, se transforma en la minimización unidimensional de:

$$\Phi(\alpha) = f(x_k + \alpha d_k).$$

De esta manera, el análisis del comportamiento de f se reduce a observar su evolución a lo largo de la recta generada por d_k . Sin embargo, hallar el valor exacto de α_k que minimiza $\Phi(\alpha)$ puede ser costoso o incluso innecesario. Por ello, en la práctica se adopta una búsqueda lineal inexacta, cuyo objetivo no es encontrar el mejor α_k , sino uno que garantice un descenso suficiente de la función. Este criterio es precisamente el fundamento de la Regla de Armijo. La idea detrás de Armijo parte de la expansión de Taylor de primer orden:

$$f(x_k + \alpha d_k) \approx f(x_k) + \alpha \nabla f(x_k)^\top d_k.$$

Esta expresión muestra que, para valores pequeños de α , el comportamiento de la función es aproximadamente lineal en la dirección d_k . La derivada direccional inicial, $\nabla f(x_k)^\top d_k$, nos indica la pendiente con la que la función comienza a decrecer. La regla de Armijo introduce una línea de referencia suavizada que actúa como límite inferior aceptable para el descenso. En lugar de exigir que la función real siga estrictamente la pendiente de la aproximación lineal (lo que sería demasiado rígido), se permite un descenso más moderado, controlado por un factor $c \in (0, 1)$, usualmente pequeño (por ejemplo, $c = 10^{-4}$). Esta línea de referencia se define como:

$$L_k(\alpha) = f(x_k) + c\alpha \nabla f(x_k)^\top d_k.$$

La condición de Armijo establece entonces que el valor real de la función en el nuevo punto debe situarse por debajo de esta línea suavizada:

$$f(x_k + \alpha_k d_k) \leq f(x_k) + c\alpha_k \nabla f(x_k)^\top d_k.$$

Este planteamiento tiene una justificación intuitiva: dado que $\nabla f(x_k)^\top d_k < 0$, en torno a $\alpha = 0$ la función realmente decrece más rápido que la línea de referencia $L_k(\alpha)$. Por tanto, para pasos suficientemente pequeños, la desigualdad de Armijo siempre se cumple. A partir de esa observación, el método implementa un procedimiento adaptativo conocido como backtracking, que comienza con un valor inicial $\bar{\alpha}$ (a menudo $\bar{\alpha} = 1$) y lo reduce progresivamente hasta que la condición de descenso suficiente se verifique. Este mecanismo logra un equilibrio entre eficiencia y estabilidad: asegura que cada iteración produzca una reducción real de f sin que el algoritmo dé pasos excesivos o inestables. Además, bajo supuestos razonables —como la diferenciabilidad y el acotamiento inferior de f en la dirección de búsqueda— el proceso de Armijo siempre encuentra un α_k adecuado en un número finito de pasos. En resumen, la condición de Armijo no solo garantiza la estabilidad del método de Newton o del gradiente, sino que también traduce, de manera controlada, la información del gradiente en una magnitud de avance que respeta tanto la forma local de la función como su comportamiento direccional.

3.3. Método de Newton

El método de Newton se fundamenta en la idea de que, en torno al punto actual x_k , la función $f(x)$ puede aproximarse mediante su expansión de Taylor de segundo orden. Si d representa el vector de desplazamiento desde x_k , esta aproximación se expresa como:

$$f(x_k + d) \approx f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2}d^\top H(x_k)d,$$

donde $\nabla f(x_k)$ es el gradiente en el punto actual y $H(x_k) = \nabla^2 f(x_k)$ es la matriz Hessiana, que contiene la información sobre la curvatura local de la función. Esta formulación permite construir el modelo cuadrático auxiliar

$$F(d) = f(x_k) + \nabla f(x_k)^\top d + \frac{1}{2}d^\top H(x_k)d,$$

el cual describe cómo se comporta la función en el entorno de x_k . El principio del método de Newton establece que el desplazamiento más eficiente es aquel que minimiza este modelo cuadrático local. La idea es intuitiva: si $F(d)$ representa bien la forma local de $f(x)$, moverse hacia el mínimo de $F(d)$ debería acercarnos directamente al mínimo real de la función. Para encontrar dicho desplazamiento, aplicamos la condición necesaria de optimalidad de primer orden, que establece que en un mínimo el gradiente debe anularse. Como estamos minimizando con respecto a d , derivamos $F(d)$ respecto de d y lo igualamos a cero:

$$\nabla_d F(d) = \nabla f(x_k) + H(x_k)d = 0.$$

De esta expresión se obtiene la dirección de Newton:

$$d_k = -[H(x_k)]^{-1}\nabla f(x_k).$$

Esta dirección combina la información del gradiente (que indica hacia dónde decrece más rápidamente la función) y de la Hessiana (que ajusta la magnitud y orientación del paso según la curvatura local). Así, Newton no solo sigue la pendiente, sino que la corrige teniendo en cuenta cómo varía el gradiente en distintas direcciones, lo que acelera la convergencia hacia el mínimo. La actualización del método se expresa como:

$$x_{k+1} = x_k + \alpha_k d_k,$$

donde $\alpha_k > 0$ es el tamaño de paso. En muchos casos puede tomarse $\alpha_k = 1$, aunque en la práctica suele emplearse una búsqueda de línea (por ejemplo, con los criterios de Armijo o de Wolfe) para garantizar que cada iteración reduzca efectivamente el valor de $f(x)$. Ahora bien, para que la dirección d_k conduzca realmente hacia un mínimo, no basta con que el gradiente se anule; también es necesario que el punto en cuestión sea de curvatura positiva. Si en un punto x_0 el gradiente se anula ($\nabla f(x_0) = 0$), la expansión de Taylor de segundo orden se simplifica a:

$$f(x_0 + d) \approx f(x_0) + \frac{1}{2}d^\top H(x_0)d,$$

Para que x_0 sea un mínimo, el valor de la función en cualquier dirección d debe ser mayor o igual al valor en x_0 , es decir: $f(x_0 + d) \geq f(x_0)$ para todo d . Esto solo se cumple si la forma cuadrática $d^\top H(x_0)d$ es no negativa para todo vector d , lo que significa que la matriz Hessiana $H(x_0)$ es semidefinida positiva. Si, además,

$$d^\top H(x_0)d > 0, \quad \forall d \neq 0,$$

entonces $H(x_0)$ es definida positiva (PD), lo cual confirma que el mínimo es estricto. En nuestra función, esta condición es crucial para interpretar el comportamiento del método. Por ejemplo, en regiones donde la curvatura de $f(x, y)$ respecto a x cambia de signo —como ocurre cuando $|x| > 1$ — la Hessiana deja de ser definida positiva, lo que puede provocar que la dirección de Newton deje de ser de descenso. En estos casos, es necesario ajustar el paso mediante técnicas de búsqueda de línea para asegurar que cada iteración mantenga la propiedad de descenso. En resumen, el método de Newton combina de forma precisa la información del gradiente y la Hessiana para avanzar hacia el mínimo de una función, pero su eficacia depende directamente de que la matriz Hessiana sea semidefinida positiva, garantizando así que la dirección obtenida efectivamente apunte hacia una región de menor valor de $f(x)$.

3.4. Método Cuasi-Newton (BFGS)

Aunque el método de Newton proporciona convergencia rápida, su costo computacional es alto, ya que requiere calcular e invertir la matriz Hessiana en cada iteración. Los métodos Cuasi-Newton, en particular el BFGS (Broyden–Fletcher–Goldfarb–Shanno), buscan un equilibrio entre precisión y eficiencia, aproximando la curvatura de forma progresiva sin calcularla directamente. El BFGS parte de una matriz inicial $H_0 = I$ que aproxima la inversa del Hessiano. En cada iteración, la dirección de búsqueda se define como:

$$d_k = -H_k \nabla f(x_k),$$

y tras avanzar al nuevo punto, se calculan los vectores de actualización:

$$s_k = x_{k+1} - x_k, \quad y_k = \nabla f(x_{k+1}) - \nabla f(x_k).$$

La ecuación secante $H_{k+1}y_k = s_k$ establece la coherencia entre el cambio en el gradiente y el desplazamiento realizado. La actualización estándar del método BFGS es:

$$\rho_k = \frac{1}{y_k^\top s_k},$$

$$H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top.$$

Esta fórmula garantiza que H_{k+1} permanezca simétrica y definida positiva, siempre que $y_k^\top s_k > 0$. Las condiciones de Wolfe, utilizadas en la búsqueda de línea, aseguran precisamente esta propiedad, manteniendo la dirección d_k como una dirección de descenso válida.

En la función analizada, esta característica resulta esencial: dado que el término logarítmico introduce regiones no convexas, la actualización de BFGS actúa como un mecanismo de regularización, asegurando la estabilidad y evitando direcciones que conduzcan a incrementos de la función. El método combina la robustez de los métodos de primer orden con la rapidez de los de segundo orden, logrando una convergencia superlineal.