Optimización

Primer Cuatrimestre 2023

Práctica N° 2: Métodos de descenso.

Notación: $\mathbf{g}_k = \nabla f(\mathbf{x}_k)$, $\mathbf{H}_k = Hf(\mathbf{x}_k)$.

Métodos básicos de descenso.

Ejercicio 1 Dada $f : \mathbb{R}^n \to \mathbb{R}$ y un punto $\mathbf{x} \in \mathbb{R}^n$, escribir una función que estime $\nabla f(\mathbf{x})$ mediante diferencias centradas. La función debe admitir el ingreso del paso h, para el cual debe haber un valor por default.

Ejercicio 2 Utilizando la función del ejercicio anterior, escribir una función que calcule $Hf(\mathbf{x})$, aplicándole diferencias centradas a las coordenadas del gradiente. La función debe devolver una matriz simétrica y no hacer cuentas innecesarias.

Ejercicio 3 Sea $f : \mathbb{R}^n \to \mathbb{R}$, dada por $f(x) = \frac{1}{2}\mathbf{x}^t\mathbf{A}\mathbf{x} - \mathbf{b}^t\mathbf{x} + c$, con A simétrica y definida positiva. Dados un punto $\mathbf{x}_k$ y una dirección $\mathbf{d}_k$, definimos $\phi(t) = f(\mathbf{x}_k + t\mathbf{d}_k)$. Probar que el mínimo de ϕ se realiza en

$$t^* = \frac{\mathbf{d}_k^t\mathbf{r}_k}{\mathbf{d}_k^t\mathbf{A}\mathbf{d}_k},$$

donde $\mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k = -\mathbf{g}_k$ es el residuo correspondiente a $\mathbf{g}_k$.

Ejercicio 4 Implementar el algoritmo de Newton puro. El usuario debe tener la posibilidad de ingresar a mano el gradiente y la matriz hessiana. Testear el algoritmo con las funciones del Ejercicio 10.

Ejercicio 5 Una función $f : \mathbb{R} \to \mathbb{R}$ se dice *unimodal* en el intervalo $[a, b]$ si existe $x^* \in (a, b)$ tal que f es estrictamente decreciente en (a, x^*) y estrictamente creciente en (x^*, b). Probar que si f es unimodal, entonces, dados α, β tales que $a < \alpha < \beta < b$ vale que:

(a) Si $f(\alpha) < f(\beta)$, entonces f es unimodal en $[a, \beta]$.

(b) Si $f(\alpha) > f(\beta)$, entonces f es unimodal en $[\alpha, b]$.

(c) ¿Qué ocurre si $f(\alpha) = f(\beta)$?

Ejercicio 6 Dada una función f unimodal en $[a, b]$, se propone el siguiente algoritmo para buscar su mínimo x^*:

1. Se fijan $a_0 = a$, $b_0 = b$.
2. Para $k = 0, 1, \ldots$ se eligen α_k, β_k tales que $a_k < \alpha_k < \beta_k < b_k$. $\varepsilon < 1$.
   - 2.1. Si $f(\alpha_k) < f(\beta_k)$, se toman: $a_{k+1} = a_k$, $b_{k+1} = \beta_k$.
   - 2.2. Si $f(\alpha_k) > f(\beta_k)$, se toman: $a_{k+1} = \alpha_k$, $b_{k+1} = b_k$

Mostrar que si la elección de α_k y β_k garantiza que $b_{k+1} - a_{k+1} \leq \varepsilon(b_k - a_k)$ para algún $\varepsilon < 1$, entonces $x^* = \lim a_k = \lim b_k$. ¿Cuántas veces debe evaluarse f para calcular $[a_{n+1}, b_{n+1}]$?

Ejercicio 7 (Búsqueda por la razón dorada) Se desea fijar un criterio para la elección de α_k, β_k en el algoritmo anterior, de manera tal que se cumplan:

- que en cada paso el intervalo se vea reducido en un factor fijo η:

$$b_{k+1} - a_{k+1} = \eta(b_k - a_k),$$

- que en cada paso sea necesario evaluar f una sola vez. Es decir, que alguno de los nuevos puntos: α_{k+1} ó β_{k+1} coincida con alguno de los anteriores α_k ó β_k.

(a) Escribir las fórmulas para α_{k+1} y β_{k+1} en función de a_k, b_k y η para que se satisfaga la primera condición.

(b) Calcular el valor de η para que se cumpla la segunda.

(c) Detallar el algoritmo que surge de esta estrategia y mostrar que para una f unimodal en $[a, b]$, las sucesiones a_k y b_k convergen al mínimo de f.

Ejercicio 8 Implementar un algoritmo que reciba como entrada una función ϕ (que se asume unimodal), un intervalo $[a, b]$, y una tolerancia δ y calcule el mínimo de ϕ en $[a, b]$ con error menor o igual que δ, mediante el algortimo de búsqueda por la razón dorada.

Ejercicio 9 Dada $\phi : \mathbb{R} \to \mathbb{R}$, implementar funciones que realicen una búsqueda inexacta del valor de t que minimiza $\phi(t)$ para $t > 0$, según:

(a) La regla de Armijo.

(b) La regla de Goldstein.

(c) El criterio de Wolfe.

Los parámetros deben tener un valor predefinido, pero el usuario debe tener la opción de elegirlos. En el caso del criterio Wolfe, implementar dos métodos para la función: uno en el que el usuario sólo provee ϕ y la derivada se estima por diferencias finitas, y uno en el que usuario ingresa también ϕ'.

Ejercicio 10 Implementar el algoritmo de máximo descenso. El criterio utilizado para realizar la búsqueda lineal en cada iteración debe ser opcional. A su vez, el usuario debe tener la posibilidad de ingresar a mano el gradiente. En caso de que esto no ocurra, el programa debe calcular el gradiente mediante el programa del Ejercicio 1.

Testear el algoritmo con las funciones:

(a) ***Rosenbrock:*** $f(x, y) = 100(y - x^2)^2 + (x - 1)^2$, cuyo mínimo es $(x, y) = (1, 1)$.

(b) ***Wood:*** $f(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 + 90(x_4 - x_3^2)^2 + (1 - x_3)^2 + 10(x_2 + x_4 - 2)^2 + 0{,}1(x_2 - x_4)^2$, cuyo mínimo se encuentra en $x = (1, 1, 1, 1)$.

(c) ***Freudenstein y Roth:*** $f(\mathbf{x}) = (-13 + x_1 + ((5 - x_2)x_2 - 2)x_2)^2 + (-29 + x_1 + ((x_2 + 1)x_2 - 14)x_2)^2$, cuyo mínimo está en $(5, 4)$, pero que también tiene un mínimo local en $(11{,}41\ldots, -0{,}89\ldots)$.

Teoría de convergencia.

Ejercicio 11 Dada una constante $b \in \mathbb{R}$, decidir si la siguiente función punto a conjunto es cerrada.

$$f(\mathbf{x}) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{y}^t\mathbf{x} \leq b\}.$$

Ejercicio 12 Sean $f : X \to Y$ y $g : Y \to Z$ dos funciones punto a conjunto. Probar que si f es cerrada en $\mathbf{x}$, g es cerrada en $f(\mathbf{x})$ e Y es compacto, entonces $g \circ f$ es cerrada en $\mathbf{x}$.

Ejercicio 13 Sean $f : X \to Y$ punto a punto y $g : Y \to Z$ punto a conjunto. Probar que si f es continua en $\mathbf{x}$ y g es cerrada en $f(\mathbf{x})$, entonces $g \circ f$ es cerrada en $\mathbf{x}$.

Ejercicio 14 Dada $\delta > 0$, sea la función punto a conjunto $\mathbf{S}^\delta$ definida como

$$\mathbf{S}^\delta(\mathbf{x}, \mathbf{d}) = \Big\{\mathbf{y} : \mathbf{y} = \mathbf{x} + \alpha\mathbf{d}, \quad 0 \leq \alpha \leq \delta; \quad f(\mathbf{y}) = \min_{0 \leq \beta \leq \delta} f(\mathbf{x} + \beta\mathbf{d})\Big\}.$$

Explicar lo que hace $\mathbf{S}^\delta$ y probar que si f es continua, entonces $\mathbf{S}^\delta(\mathbf{x}, \mathbf{d})$ es cerrada en $(\mathbf{x}, \mathbf{d})$. ¿Por qué es importante este resultado?

Ejercicio 15 Dada $\varepsilon > 0$, sea la función punto a conjunto $\mathbf{S}^\varepsilon$ definida como

$$\mathbf{S}^\varepsilon(\mathbf{x}, \mathbf{d}) = \Big\{\mathbf{y} : \mathbf{y} = \mathbf{x} + \alpha\mathbf{d}, \quad \alpha \geq 0; \quad f(\mathbf{y}) \leq \min_{\beta \geq 0} f(\mathbf{x} + \beta\mathbf{d}) + \varepsilon\Big\}.$$

Explicar lo que hace $\mathbf{S}^\varepsilon$ y probar que si f es continua y $\mathbf{d} \neq \mathbf{0}$, entonces $\mathbf{S}^\varepsilon(\mathbf{x}, \mathbf{d})$ es cerrada en $(\mathbf{x}, \mathbf{d})$. ¿Por qué es importante este resultado?

Ejercicio 16 Mostrar que si A es una aplicación punto a punto, en el Teorema de Convergencia Global puede eliminarse la hipótesis de que los puntos $\mathbf{x}_k$ caigan sobre un compacto.

Ejercicio 17 (Búsqueda compacta) Sea $f : \mathbb{R}^n \to \mathbb{R}$ de clase C^2 estrictamente convexa con mínimo $\mathbf{x}^*$.

(a) Sea $D = \{\mathbf{e}_1, -\mathbf{e}_1, \mathbf{e}_2, -\mathbf{e}_2, \ldots, \mathbf{e}_n, -\mathbf{e}_n\}$. Probar que para todo $\mathbf{x} \neq \mathbf{x}^*$ existe una dirección de descenso $\mathbf{d} \in D$ para f en $\mathbf{x}$.

(b) Sea $\mathbf{x} \in \mathbb{R}^n$ y $\alpha > 0$, consideramos la función punto a conjunto:

$$A(\mathbf{x}, \alpha) = \{\mathbf{y} : \mathbf{y} = \mathbf{x} + \alpha\mathbf{d}, \quad \mathbf{d} \in D\}.$$

Probar que A es cerrada en $(\mathbf{x}, \alpha)$ si $\alpha \neq 0$.

(c) Consideramos el algoritmo de Búsqueda Compacta dado por:

1. Tomar $\mathbf{x}_0 \in \mathbb{R}^n$, α_0, $k = 0$.
2. Se busca $\mathbf{y} \in A(\mathbf{x}_k, \alpha_k)$ tal que $f(\mathbf{y}) < f(\mathbf{x})$.
   2.1. Si existe $\mathbf{y}$, se toman: $\mathbf{x}_{k+1} = \mathbf{y}$, $\alpha_{k+1} = \alpha_k$, $k = k + 1$.
   2.2. Si no, se toma $\alpha_k = \frac{\alpha_k}{2}$.
3. Ir a 2.

Probar que $\mathbf{x}_k \to \mathbf{x}^*$ cuando $k \to \infty$.

Ejercicio 18 Probar que la condición de Goldstein determina un algoritmo de búsqueda cerrado.

Ejercicio 19 Probar que bajo hipótesis suficientes de suavidad, la condición de Wolfe determina un algoritmo de búsqueda cerrado.

Ejercicio 20 Un defecto del método del gradiente es que su convergencia puede resultar muy lenta, incluso para funciones cuadráticas. Esto se debe a que el gradiente se ve afectado por problemas de *escala*, que podría resolverse mediante un re-escale de las variables. Una heurística para lograr esto es tomar como dirección de descenso $-\mathbf{D}_k\mathbf{g}_k$, donde $\mathbf{D}_k$ es una matriz diagonal que aproxima la diagonal de $\mathbf{H}_k^{-1}$. Implementar este algoritmo.

Para las funciones del Ejercicio 10, comparar el número de iteraciones y el tiempo de ejecución (mediante el macro `@time`), de los métodos del gradiente, de Newton y el de gradiente re-escalado.

Ejercicio 21 El método de Levenberg-Marquardt consiste en tomar como dirección de descenso $-(\mathbf{H}_k + \mu_k\mathbf{I})^{-1}g_k$, donde μ_k se elige en cada paso de manera tal que si $\mathbf{H}_k$ es definida positiva, $\mu_k = 0$, y en caso contrario, μ_k es tal que $\mathbf{H}_k + \mu_k\mathbf{I}$ resulte definida positiva. En la práctica, esto equivale a un método intermedio entre Newton ($\mu_k = 0$) y el gradiente (μ_k grande). La elección de μ_k puede resultar compleja, dado que uno querría que $\mu_k \sim -\lambda$, donde λ es el autovalor negativo de $\mathbf{H}_k$ de máximo módulo, y el cálculo de autovalores es un problema complicado.

Sin embargo, en este hilo se discute la razonabilidad de ese enfoque y se argumenta en favor de un método que consiste esencialmente en lo siguiente. Si $\mathbf{H}_k = \mathbf{U}^t\mathbf{D}\mathbf{U}$, con $\mathbf{D}$ diagonal, se calcula $\mathbf{H}_k^\ddagger = \mathbf{U}^t\mathbf{D}^\ddagger\mathbf{U}$, de modo que $\mathbf{H}^\ddagger$ opere como una versión aceptable de $\mathbf{H}_k^{-1}$. Para ello, se toma $\mathbf{D}^\ddagger$ diagonal tal que:

- Si d_{ii} es grande (por ejemplo, $|d_{ii}| > \delta \,\text{máx}\,|\mathbf{D}|$ para algún δ), entonces $d_{ii}^\ddagger = |d_{ii}|^{-1}$.
- Si d_{ii} es pequeño (no cumple la condición anterior), entonces $d_{ii}^\ddagger = \delta \,\text{máx}\,|\mathbf{D}|$.

Discutir las ideas detrás de esta propuesta. Suponiendo que la matriz $\mathbf{H}_k$ tiene autovalores negativos de módulo grande, ¿Qué ocurre en cada método con ellos? ¿Qué ocurre con los autovalores de módulo chico? ¿Cómo afecta esto la dirección de búsqueda?

El paquete `PositiveFactorizations` implementa estas ideas. Por ejemplo: mediante el comando `cholesky(Positive,H)` se recibe una matriz simétrica $\mathbf{H}$ (no necesariamente definida positiva) y se construye una descomposición de Cholesky de $\mathbf{H}^\ddagger$ (en el proceso de realizar la factorización, se determina cómo deben modificarse la matriz).

(a) Implementar el algoritmo de Levenberg-Marquardt tomando μ_k como el mínimo autovalor de $\mathbf{H}_k$ + un valor ε_0.

(b) Implementar un algoritmo tipo Newton pero que en lugar de invertir $\mathbf{H}_k$, calcule $\mathbf{H}^\ddagger$ usando `PositiveFactorizations.cholesky`.

(c) Aplicar ambos métodos a la función de Beale:

$$f(x,y) = (1{,}5 - x + xy)^2 + (2{,}25 - x + xy^2)^2 + (2{,}625 - x + xy^3)^2,$$

cuyo mínimo se encuentra en $(3, 0{,}5)$ y a la función de Ackley:

$$f(x,y) = -20e^{-0{,}2\sqrt{0{,}5(x^2+y^2)}} - e^{0{,}5(\cos 2\pi x + \cos 2\pi y)} + e + 20,$$

cuyo mínimo absoluto está en el origen. En cada caso, graficar la sucesión generada por cada método.

Aplicaciones.

Ejercicio 22 (Braquistócrona) Considerar el problema de la *braquistócrona* consistente en hallar la curva que minimiza el tiempo de caída de una particula por efecto de la gravedad. Concretamente, buscamos una función $y : [0, \frac{\pi}{2}] \to [0, 1]$ tal que $y(0) = 1$, $y(\frac{\pi}{2}) = 0$ que minimice el funcional:

$$T(y) = \int_0^{\frac{\pi}{2}} \sqrt{\frac{1 + y'(x)^2}{2g(1 - y(x))}} dx,$$

donde $g = 9{,}81 m/s^2$ es la aceleración gravitatoria. Para resolver este problema se realiza una discretización en la variable x: $0 = x_0 < x_1 < \ldots < x_n < x_{n+1} = \frac{\pi}{2}$, con $x_i = ih$, para $h = \frac{\pi}{2(n+1)}$. A su vez, notamos y_i a nuestra aproximación de $y(x_i)$. Por comodidad, asumiremos que y es lineal en cada intervalo $[x_i, x_{i+1}]$.

(a) Probar que con estas hipótesis el funcional T puede pensarse como $T : \mathbb{R}^n \to \mathbb{R}$ dado por:

$$T(\mathbf{y}) = \sqrt{\frac{2}{g}} \sum_{i=0}^{n} \sqrt{1 + \left(\frac{h}{y_{i+1} - yi}\right)^2} \left(\sqrt{1 - y_{i+1}} - \sqrt{1 - y_i}\right).$$

Observar que y_0 e y_{n+1} son datos.

(b) Calcular analíticamente el gradiente y el hessiano de la expresión anterior de T.

(c) Implementar funciones que reciban como input el número n de incógnitas de la discretización y devuelvan el funcional T, el gradiente y el hessiano de T.

(d) Implementar funciones que reciban como input el número n de incógnitas de la discretización y devuelvan el funcional T, el gradiente y el hessiano de T.

(e) Resolver el problema de minimización anterior con los métodos vistos.

Ejercicio 23 (Ajuste algebraico de circunferencias) Dados n puntos en el plano $(x_1, y_1), \ldots, (x_n, y_n)$ provenientes de mediciones que se sabe deberían corresponder a una circunferencia, se desea hallar aquella:

$$C : x^2 + y^2 - 2\alpha x - 2\beta y - \gamma = 0,$$

que mejor ajusta los datos. El enfoque algebraico propone realizar el ajuste sobre la *ecuación* del circunferencia. Es decir, plantea buscar los valores de los parámetros α, β, γ de manera que se minimice el funcional:

$$F(\alpha, \beta, \gamma) = \sum_{i=1}^{n} \left(x_i^2 + y_i^2 - 2\alpha x_i - 2\beta y_i - \gamma\right)^2 = \sum_{i=1}^{n} \left(2\alpha x_i + 2\beta y_i + \gamma - r_i^2\right)^2,$$

donde en la última expresión tomamos $r_i^2 = x_i^2 + y_i^2$.

(a) Mostrar que la expresión para C (elegida por conveniencia) es general y puede describir cualquier circunferencia.

(b) Dar las ecuaciones del centro y del radio de C en función α, β y γ.

(c) Resolver el problema de minimización anterior con los métodos vistos usando solamente aproximaciones del gradiente y del hessiano.

Nota: Los puntos se pueden conseguir sorteando n valores en $[0, 2\pi]$ y reemplazando en la ecuación paramétrica de una circunferencia $c(t) = (A\cos(t) + h, A\,\mathrm{sen}(t) + k)$. Revisar la función `rand` y en caso de querer una distribución más específica, el paquete `Distributions`.

(d) Observar que es un problema de minimización cuadrática. Plantear las ecuaciones que resuelven el problema de forma exacta y resolver analíticamente.

(e) Evaluar el desempeño graficando para cada método simultáneamente el error en norma 2 de las estimaciones de los parámetros en función del número de iteraciones.

Ejercicio 24 (Regresión logística) Dadas n observaciones $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\} \subset \mathbb{R}^d$ y sus respectivas respuestas $\{y_1, \ldots, y_n\} \subset \{0, 1\}$, se desea modelar la probabilidad de que una variable aleatoria $Y \in \{0, 1\}$ clasifique como 1 dado cierto vector aleatorio $\mathbf{X} \in \mathbb{R}^d$ de la siguiente forma:

$$\mathbb{P}(Y = 1|\mathbf{X}) = \text{expit}(\mathbf{a}^t\mathbf{X}),$$

donde $\mathbf{a} \in \mathbb{R}^d$ y $\text{expit}(t) = \frac{1}{1+e^{-t}}$. Notar que es equivalente a decir que $Y|\mathbf{X}$ se distribuye $\text{Ber}(\text{expit}(\mathbf{a}^t\mathbf{X}))$. Se desea hallar el parámetro $\mathbf{a}$ que sea óptimo en el sentido que $\mathbb{P}(Y = 1|\mathbf{X_i} = \mathbf{x_i})$ sea parecido a y_i para cada $1 \leq i \leq n$. Una propuesta es hallándolo mediante la maximización de la función de verosimilitud para la muestra aleatoria $\{(\mathbf{X_1}, Y_1), \ldots, (\mathbf{X_n}, Y_n)\}$ para las observaciones ya vistas:

$$\begin{aligned} L(\mathbf{a}) &= \mathbb{P}_\alpha(Y_1 = y_1, \ldots, Y_n = y_n|\mathbf{X_1} = \mathbf{x_1}, \ldots, \mathbf{X_n} = \mathbf{x_n}) \\ &= \prod_{i=1}^{n} [\text{expit}(\mathbf{a}^t\mathbf{x_i})]^{y_i}[1 - \text{expit}(\mathbf{a}^t\mathbf{x_i})]^{1-y_i}. \end{aligned}$$

(a) Probar que es equivalente a hallar el parámetro $\mathbf{a}$ que minimiza la expresión:

$$-\sum_{i=1}^{n} \left[y_i \ln\left(\text{expit}(\mathbf{a}^t\mathbf{x_i})\right) + (1 - y_i) \ln\left(1 - \text{expit}(\mathbf{a}^t\mathbf{x_i})\right)\right].$$

(b) Implementar una función que, dados los datos $\{\mathbf{x_1}, \ldots, \mathbf{x_n}\}$ (como matriz), sus respectivas respuestas $\{y_1, \ldots, y_n\}$ y el método, devuelva el parámetro $\mathbf{a}$ óptimo para el problema del ítem anterior usando solamente aproximaciones del hessiano.

(c) Implementar una función que, dado un parámetro $\mathbf{a}$, una observación $\mathbf{x}$ y un parámetro opcional de umbral (que por default debe valer 0,5), clasifique a la observación de la siguiente forma:

$$C(\mathbf{x}) = \begin{cases} 1 \text{ si } \text{expit}(\mathbf{a}^t\mathbf{x}) \geq \text{umbral} \\ 0 \text{ si no} \end{cases}.$$

Usando el $\mathbf{a}$ óptimo y las observaciones $\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$ usadas para estimarla, controlar si se respetan las clasificaciones originales.

Nota: La función expit es una biyección entre $\mathbb{R}$ y el intervalo $(0, 1)$.