Capítulo 11

Métodos Sin Derivadas.

Todos los métodos de gradiente examinados hasta ahora requieren cálculos de al menos el gradiente $\nabla f(x^k)$ y posiblemente la matriz Hessiana $\nabla^2 f(x^k)$ en cada punto generado x^k. En muchos problemas, cualquiera de estos elementos pueden no estar disponibles en forma explícita o están dados por expresiones muy complicadas. En tales casos, puede ser preferible usar los mismos algoritmos que anteriormente con todas las derivadas no disponibles aproximadas por diferencias finitas.

Las primeras derivadas deben ser aproximadas por la fórmula de la diferencia adelantada:

$$\frac{\partial f(x^k)}{\partial x_i} \approx \frac{1}{h}(f(x^k + he_i) - f(x^k)) \tag{11.1}$$

o por la fórmula de la diferencia central:

$$\frac{\partial f(x^k)}{\partial x_i} \approx \frac{1}{2h}(f(x^k + he_i) - f(x^k - he_i)). \tag{11.2}$$

En estas relaciones, h es un escalar positivo pequeño y e_i es el i-ésimo vector (i-ésima columna de la matriz identidad). En algunos casos el mismo valor de h puede ser utilizado para todas las derivadas parciales, pero en otros casos, particularmente cuando el problema está mal escalado, es esencial utilizar un valor diferente de h para cada derivada parcial. Éste es un proceso difícil que a menudo requiere tanteo; ver la siguiente discusión.

La fórmula de la diferencia central requiere dos veces más operaciones que la fórmula de la diferencia adelantada. De cualquier manera, es mucho más exacta. Esto puede ser visto formando la correspondiente expansión de la serie de Taylor, y verificando que (en aritmética exacta) el valor absoluto del error entre la aproximación y la derivada actual es $O(h)$ para la fórmula de la diferencia adelantada, mientras que es $O(h^2)$ para la fórmula de la diferencia central. Nótese además que si la usada es la fórmula de la diferencia central, y ésta no supone mucho más esfuerzo computacional, esencialmente puede ser utilizada para aproximar cada elemento de la diagonal del Hessiano; utilizando la fórmula:

$$\frac{\partial^2 f(x^k)}{(\partial x_i)^2} \approx \frac{1}{h^2}(f(x^k + he_i) + f(x^k - he_i) - 2f(x^k)).$$

Estas aproximaciones pueden ser utilizadas en esquemas basados en el escalado de la diagonal.

Para reducir el error de aproximación, deberíamos elegir la diferencia finita del intervalo h tan pequeño como sea posible. Desafortunadamente, hay un límite en como h puede ser reducido debido al error de redondeo que ocurre cuando se trabaja con cantidades muy pequeñas. En particular, un error ∂ debido a una precisión aritmética finita evaluando el numerador en la ecuación 1 (o en la 2), resulta en un error de ∂/h (o $\partial/2h$, respectivamente) en la evaluación de la primera derivada. El error de redondeo es particularmente evidente en las fórmulas aproximadas (1) y (2) cerca de un punto estacionario donde ∇f es cercano a cero, y el tamaño del error relativo en la aproximación del gradiente llega a ser muy grande.

La experiencia práctica sugiere que una buena política es mantener el escalar h para cada derivada como un valor fijo, que aproximadamente equilibra el error de aproximación con el error de redondeo. Los cálculos anteriores, nos llevan a los principios generales:

$$\frac{\partial}{h} = O(h) \quad \acute{o} \quad h = O(\partial^{1/2}),$$

para la fórmula de la diferencia adelantada (1),

$$\frac{\partial}{2h} = O(h^2) \quad \acute{o} \quad h = O(\partial^{1/3}),$$

para la fórmula de la diferencia central (2), donde ∂ es el error debido a la precisión aritmética finita en la evaluación del numerador en la ecuación (1) (o en la 2). Así, un valor muy grande de h puede ser utilizado en la fórmula de la diferencia central. Una buena regla práctica es utilizada en la fórmula adelantada (1) hasta que el valor absoluto de la correspondiente derivada aproximada llegue a ser menor que una cierta tolerancia, es decir:

$$\left| \frac{f(x^k + he_i) - f(x^k)}{h} \right| \leq \varepsilon,$$

donde $\varepsilon > 0$ es algún escalar no especificado. En este punto, debería de hacerse un cambio en la fórmula de la diferencia central.

Las segundas derivadas deben ser aproximadas por la fórmula de la diferencia adelantada:

$$\frac{\partial^2 f(x^k)}{\partial x_i \partial x_j} \approx \frac{1}{h} \left(\frac{\partial f(x^k + he_j)}{\partial x_i} - \frac{\partial f(x^k)}{\partial x_i} \right) \tag{11.3}$$

o la fórmula de la diferencia central:

$$\frac{\partial^2 f(x^k)}{\partial x_i \partial x_j} \approx \frac{1}{2h} \left(\frac{\partial f(x^k + he_j)}{\partial x_i} - \frac{\partial f(x^k - he_j)}{\partial x_i} \right) \tag{11.4}$$

La experiencia práctica sugiere que en una forma discreta del método de Newton, no es importante una exactitud excesiva en la aproximación de las derivadas segundas en término del porcentaje de la convergencia. Por esta razón, el uso exclusivo de la fórmula de la diferencia adelantada (3) es adecuado en muchos casos. De cualquier manera,

se debería controlar el valor de la aproximación del Hessiano discreto e introducir modificaciones si es necesario, como discutimos anteriormente.

11.1. Coordenadas de descenso

Hay varios métodos sin derivadas para minimizar funciones diferenciables. Particularmente un algoritmo importante es el *método de la coordenada de descenso*. Aquí el coste es minimizado a lo largo de una dirección de una coordenada en cada iteración. El orden en el cual las coordenadas son escogidas debe cambiar en el curso del algoritmo. En el caso donde el orden es cíclico, dado x^k, la i-ésima coordenada de x^{k+1} es determinada por:

$$x_i^{k+1} = arg \min_{\xi \in \mathbb{R}} f(x_1^{k+1}, ..., x_{i-1}^{k+1}, \xi, x_{i+1}^k, ..., x_n^k); \tag{11.5}$$

ver la Figura 1.8.1. El método puede también utilizarse para la minimización de f sujeta a los límites superior e inferior sobre las variables x^i (la minimización sobre $\xi \in \mathbb{R}$ en la ecuación anterior es reemplazada por la minimización sobre el intervalo apropiado). Analizaremos el método dentro de este contexto más general en el capítulo siguiente.

Una ventaja importante del método de la coordenada de descenso es que es conveniente para la *computación paralela*. En particular, suponemos que hay un subconjunto de coordenadas $x_{i_1}, x_{i_2}, ..., x_{i_m}$, las cuales no verifican la función de costo, esto es, $f(x)$ puede ser escrita como $\sum_{r=1}^m f_r(x_{i_r})$, donde para cada r, $f_r(x_{i_r})$ no depende de las coordenadas x_{i_s} para todo $s \neq r$. Entonces uno puede interpretar las m iteraciones de la coordenada de descenso:

$$x_{i_r}^{k+1} = arg \min_{\xi} f_{i_r}(x^k + \xi e_{i_r}), \quad r = 1, ..., m,$$

independientemente y en paralelo. De este modo, en problemas con especial estructura donde el conjunto de coordenadas pueden ser divididas en p subconjuntos con la propiedad de independencia justamente descrita, uno puede interpretar un ciclo lleno de iteraciones de la coordenada de descenso en p pasos paralelos (en oposición a n), asumiendo desde luego que un número suficiente de procesos paralelos está disponible.

El método de la coordenadas de descenso generalmente tiene propiedades de convergencia similares al descenso máximo. Para funciones objetivo continuas y derivables, se puede demostrar que los puntos límites de las secuencias generadas son estacionarios, aunque las pruebas de ésto algunas veces son complicadas y requieren hipótesis adicionales (ver se suele requerir una convergencia estricta de la función objetivo a lo largo de cada coordenada). Existe un notable acuerdo entre los investigadores de que el uso de la coordenada de descenso es particularmente favorable en la resolución de problemas duales (ver Sección 6.2). El porcentaje de convergencia de la coordenada de descenso suele ser lineal o sublineal, hacia mínimos locales (regulares

o no regulares respectivamente). A menudo, los cambios entre las coordenadas de descenso y descenso máximo son dictados por la estructura de la función de costo. Ambos métodos pueden ser muy lentos, pero para muchos contextos prácticos, pueden ser bastantes efectivos.

11.2. Métodos de búsqueda directa

En el método de la coordenada de descenso nosotros buscábamos a lo largo de un conjunto fijo de direcciones en una coordenada y garantizábamos un valor objetivo mejorado porque estas direcciones son linealmente independientes. Esta idea puede ser generalizada utilizando diferentes conjuntos de direcciones y de vez en cuando cambiando este conjunto de direcciones con el propósito de acelerar la convergencia. Hay un número de métodos de este tipo: el método de Rosenbrock [Ros60a], el modelo de investigación algorítmica de Hooke y Jeeves [HoJ61], y los algoritmos del símplex de Spendley, Hext, y Himsworth [SHH62], y Nelder y Mead [NeM65]. Desafortunadamente la racionalidad de estos métodos es meramente heurística, y sus propiedades teóricas de convergencia son a menudo insatisfactorias. Sin embargo, estos métodos son con frecuencia bastante simples de implementar y no requieren los cálculos del gradiente. A continuación describimos el método símplex de Nelder y Mead (no confundirlo con el método símplex de programación lineal), que ha obtenido una considerable popularidad.

En una iteración estandar de este método, comenzamos con un *símplex*, que es, la envolvente convexa de $n+1$ puntos, $x^0, x^1, ..., x^n$, y terminamos con otro símplex. Por x_{min} y x_{max} denotamos el *mejor* y el *peor* de los vértices del símplex, satisfaciendo éstos:

$$f(x_{min}) = \min_{i=0,1,..,n} f(x^i) \tag{11.6}$$

$$f(x_{max}) = \max_{i=0,1,..,n} f(x^i). \tag{11.7}$$

También $\hat{x}$ denota el centroide de la cara del símplex formado por los vértices una vez eliminad x_{max}:

$$\hat{x} = \frac{1}{n}\left(-x_{max} + \sum_{i=0}^{n} x^i\right). \tag{11.8}$$

La iteración sustituye el peor vértice x_{max} por uno "mejor". En particular, se calcula el *punto reflejo* $x_{ref} = 2\hat{x} - x_{max}$ que se encuentra en la recta que pasa por x_{max} y $\hat{x}$, y es simétrico a x_{max} con respecto a $\hat{x}$. Dependiendo del valor objetivo en x_{ref} relativo a los otros puntos del símplex y x_{max}, un nuevo vértice x_{new} es calculado, y un nuevo símplex es formado a partir del antiguo reemplazando el vértice x_{max} por x_{new}, mientras mantenemos los restantes n vértices.

PASO 1:(PASO REFLEJO)

Calcula:

$$x_{ref} = 2\hat{x} - x_{max}. \tag{11.9}$$

Entonces calcula x_{new} de acuerdo a los siguientes tres casos:

1. (x_{ref} tiene mínimo valor objetivo) Si $f(x_{min}) > f(x_{ref})$, ir al Paso 2.
2. (x_{ref} tiene valor objetivo intermedio) Si el máx$\{f(x^i)|x^i \neq x_{max}\} > f(x_{ref}) \geq f(x_{min})$, ir al Paso 3.
3. (x_{ref} tiene máximo valor objetivo) Si $f(x_{ref}) \geq$ máx$\{f(x^i)|x^i \neq x_{max}\}$, ir al Paso 4.

PASO 2: (INTENTO DE EXPANSIÓN)

Calcula:

$$x_{exp} = 2x_{ref} - \hat{x}. \tag{11.10}$$

Define:

$$x_{new} = \begin{cases} x_{exp} & \text{si } f(x_{exp}) < f(x_{ref}), \\ x_{ref} & \text{en otro caso ,} \end{cases}$$

y se forma el nuevo símplex reemplazando el vértice x_{max} con x_{new}.

PASO 3: (USANDO EL REFLEJO)

Define $x_{new} = x_{ref}$, y forma el nuevo símplex reemplazando el vértice x_{max} con x_{new}.

PASO 4: (CONTRACCIÓN REALIZADA)

Define:

$$x_{new} = \begin{cases} \frac{1}{2}(x_{max} + \hat{x}) & \text{si } f(x_{max}) \leq f(x_{ref}), \\ \frac{1}{2}(x_{ref} + \hat{x}) & \text{en otro caso,} \end{cases} \tag{11.11}$$

y se forma el nuevo símplex reemplazando el vértice x_{max} con x_{new}.

De cualquier forma no se conocen resultados de convergencia para este método. Además, cuando la función objetivo no es convexa, es posible que el vértice del nuevo símplex x_{new} tenga un valor objetivo mayor que el antiguo vértice x_{max}. En este caso se ha sugerido una modificación que consiste en comprimir el antiguo símplex hacia el mejor vértice x_{min}, esto es, formar un nuevo símplex reemplazando todos los vértices x^i, $i = 0, 1, .., n$, por:

$$x^i = \frac{1}{2}(x^i + x_{min}), \quad i = 0, 1, .., n.$$

Este método parece funcionar razonablemente bien en la práctica, particularmente para problemas de dimensión relativamente pequeña (por debajo de 10). De cualquier manera, no hay garantía de tener propiedades de convergencia y en realidad un contraejemplo de su convergencia es dado en [McK94]. La referencia [Tse95a] proporciona una modificación relativamente simple con propiedades de convergencia satisfactorias. Hay también un número de métodos relacionados, algunos de los cuales tienen propiedades de convergencia demostrables; ver [DeT91] y [Tor91]. Nótese que las constantes usadas en las ecuaciones (9), (10) y (11) son arbitrarias y están sugeridas por la interpretación geométrica del método. Las forma más general de estas ecuaciones es:

$$x_{ref} = \hat{x} + \beta(\hat{x} - x_{max}),$$

$$x_{exp} = x_{ref} + \gamma(x_{ref} - \hat{x}),$$

$$x_{con} = \begin{cases} \theta x_{max} + (1-\theta)\hat{x} & \text{si } f(x_{max}) \leq f(x_{ref}), \\ \theta x_{ref} + (1-\theta)\hat{x} & \text{en otro caso,} \end{cases}$$

donde $\beta > 0$, $\gamma > 0$ y $\theta \in (0,1)$ son escalares conocidos como el *coeficiente reflejo*, el *coeficiente de expansión* y el *coeficiente de contracción*, respectivamente. Las fórmulas de las ecuaciones (9), (10) y (11) corresponden a $\beta = 1$, $\gamma = 1$ y $\theta = 1/2$, respectivamente.