Elementos de programación diferenciable

Alejandro Sánchez Yalí

1. ¿Qué es la programación diferenciable?

Un **programa de computadora** consiste en una serie de instrucciones básicas que efectúan una tarea específica. En el ámbito de las ciencias computacionales, estos programas son creados por **programadores**. Sin embargo, existen tareas, especialmente aquellas que involucran patrones intricados y decisiones complejas, como el reconocimiento de imágenes o la generación de texto, donde escribir un programa tradicional resulta extremadamente difícil, sino imposible.

En contraste, las redes neuronales modernas ofrecen diferentes enfoques. Estas se construyen mediante la combinación de funciones parametrizadas y se entrenan directamente de los datos usando optimización basada en el **gradiente**. Durante este proceso de entrenamiento, las redes neuronales aprenden simultáneamente la extracción de características y la ejecución de tareas, lo que les permite desarrollar tareas complejas que antes se consideraban inalcanzables para los programas tradicionales. Este nuevo paradigma de programación ha sido denominado como «programación diferenciable» o «software 2.0», términos que han sido popularizados por LeCun (2018) y Karpathy (2017).

Definición 1.1 (Programación diferenciable)**. Programación diferenciable** es un paradigma de programación donde los programas (incluyendo flujos de control y estructuras de datos) pueden ser derivados automáticamente, permitiendo la optimización de parámetros basada en gradiente.

1.1. Redes neuronales modernas como programas parametrizados

En programación diferenciable, un programa de computadora clásico se puede definir como la composición de operaciones elementales, formando un **grafo computacional**. La diferencia clave es que los programas (como las redes neuronales) contienen parámetros que pueden ser ajustados a partir de los datos y pueden ser derivados usando la derivación automática (*autodiff*). Típicamente, se asume que los programas definen funciones matemáticamente válidas: la función debe retornar valores idénticos para argumentos idénticos y no debe tener efectos colaterales. Además, la función debe tener derivadas bien definidas, asegurando que se pueda usar la optimización basada en el algoritmo del gradiente. Por lo tanto, la programación diferenciable no es solo el arte de derivar a través de programas, sino también de diseñarlos cuidadosamente.

1.2. ¿Por qué las derivadas son importantes?

El aprendizaje computacional típicamente se reduce a la optimización de una función objetivo determinada, que es la composición de una función de pérdida y una función del modelo (red neuronal). La optimización sin derivadas se denomina **optimización de orden cero**. En este caso, solo se asume que podemos evaluar la función objetivo que deseamos optimizar. Lamentablemente, este método sufre de la maldición de la dimensionalidad, es decir, solo es viable para problemas de baja dimensión, con menos de 10 dimensiones. La optimización basada en derivadas, por otro lado, es mucho más eficiente y puede escalar a millones o miles de millones de parámetros. Los algoritmos que utilizan primeras y segundas derivadas se conocen, respectivamente, como algoritmos de **primer orden** y **segundo orden**.

1.3. ¿Por qué la diferenciación automática es importante?

Antes de la revolución de la diferenciación automática, investigadores y practicantes necesitaban implementar manualmente el gradiente de las funciones que deseaban optimizar. Calcular gradientes manualmente podía convertirse en algo tedioso para funciones complicadas. Además, cada vez que la función era modificada (por ejemplo, para probar una nueva idea), el gradiente necesitaba ser recalculado. La diferenciación automática representa un cambio radical porque permite a los usuarios enfocarse en la experimentación creativa con funciones para sus tareas específicas.

1.4. Programación diferenciable no es solo aprendizaje computacional

Aunque existe un solapamiento entre el aprendizaje computacional y la programación diferenciable, sus enfoques son diferentes. El aprendizaje computacional estudia las redes neuronales compuestas de múltiples capas, que les permiten aprender **representaciones intermedias** de los datos. Por ejemplo, las redes neuronales convolucionales están diseñadas para el procesamiento de imágenes, mientras que las redes recurrentes están diseñadas para secuencias. Por otro lado, la programación diferenciable estudia las técnicas para diseñar programas complejos y diferenciables. Su uso va más allá del aprendizaje computacional: por ejemplo, en el aprendizaje por refuerzo, la programación probabilística y la computación científica en general.

1.5. Programación diferenciable no es solo diferenciación automatizada

Si bien la diferenciación automatizada es un ingrediente clave de la programación diferenciable, no es el único. La programación diferenciable también está comprometida con el diseño de operaciones que sean diferenciables en principio. De hecho, gran parte de la investigación sobre programación diferenciable se ha dedicado a hacer que las operaciones de la programación computacional clásica sean compatibles

con la programación diferenciable. Cabe destacar que muchas relajaciones diferenciables pueden interpretarse en un marco probabilístico. El tema central de este libro es la interacción entre la optimización, la probabilidad y la diferenciación. La diferenciación es útil para la optimización y, recíprocamente, la optimización puede ser útil para diseñar operadores diferenciables.

2. Fundamentos

En este capítulo revisaremos los conceptos clave de la diferenciación. En particular, enfatizaremos el papel fundamental que juegan las transformaciones lineales.

2.1. Funciones univariadas

2.1.1. Derivadas

Para estudiar funciones, así como sus derivadas, necesitamos capturar sus variaciones infinitesimales alrededor de los puntos tal como se ha definido para la noción de límite.

Definición 2.1 (Límite)**.** Sea $f : \mathbb{R} \mapsto \mathbb{R}$ una función y sean $x_0, c \in \mathbb{R}$. Diremos que c es el límite de f cuando x tiende a x_0 si para todo $\varepsilon > 0$, existe $\delta > 0$ tal que para todo $x \in \mathbb{R}$ que satisface $0 < |x - x_0| < \delta$, se cumple que $|f(x) - c| < \varepsilon$. En este caso escribimos:

$$\lim_{x \to x_0} f(x) = c.$$

Los límites se preservan bajo operaciones algebraicas básicas. En efecto, si tenemos $f, g : \mathbb{R} \to \mathbb{R}$ y suponemos que existen los límites

$$\lim_{x \to x_0} f(x) = c \quad \text{y} \quad \lim_{x \to x_0} g(x) = d.$$

Entonces, para cualesquiera $a, b \in \mathbb{R}$:

1. **Linealidad:** Si definimos $(af + bg)(x) := af(x) + bg(x)$, entonces

$$\lim_{x \to x_0} (af + bg)(x) = ac + bd$$

2. **Multiplicación:** Si definimos $(fg)(x) := f(x)g(x)$, entonces

$$\lim_{x \to x_0} (fg)(x) = cd.$$

Con la noción de límite, podemos definir la clase de funciones que presentan un comportamiento regular, es decir, aquellas funciones donde el límite en cualquier punto coincide con el valor de la función evaluada en ese punto. Esta propiedad fundamental se conoce como **continuidad**.

Definición 2.2 (Funciones continuas)**.** Una función $f : \mathbb{R} \to \mathbb{R}$ es continua en un punto $x_0 \in \mathbb{R}$ si

$$\lim_{x \to x_0} f(x) = f(x_0).$$

Además, diremos que f es continua (globalmente) si es continua en todo punto $x_0 \in \mathbb{R}$.

Aunque la noción de continuidad parece una suposición benigna, varias funciones sencillas, como la función escalón de Heavyside, no son continuas y requiren un tratamiento especial.

Observación (Notación de Landau). A lo largo de este texto usaremos la notación o pequeña de Landau. Para dos funciones $f, g : \mathbb{R} \to \mathbb{R}$, escribiremos $g(x) = o(f(x))$ cuando $x \to x_0$ si

$$\lim_{x \to x_0} \frac{|g(x)|}{|f(x)|} = 0.$$

Intuitivamente, esto significa que g es asintóticamente dominada por f cuando $x \to x_0$. Como caso particular, podemos caracterizar la continuidad de una función f en un punto x_0 mediante esta notación: f es continua en x_0 si y solo si

$$f(x_0 + h) = f(x_0) + o(1) \quad \text{cuando } h \to 0.$$

Consideremos ahora una función $f : \mathbb{R} \to \mathbb{R}$. Su valor en un intervalo $[x_0, x_0 + h]$ puede ser aproximado por la secante entre los puntos $(x_0, f(x_0))$ y $(x_0 + h, f(x_0 + h))$ como una función lineal con pendiente $(f(x_0 + h) - f(x_0))/h$. En el límite cuando la variación h tiende a cero alrededor de x_0, la secante converge a la **recta tangente** de f en x_0, y su pendiente se define como la derivada de f en x_0. La siguiente definición formaliza esta intuición.

Definición 2.3 (Derivada)**.** La derivada de una función $f : \mathbb{R} \to \mathbb{R}$ en un punto $x_0 \in \mathbb{R}$ se define como

$$f'(x_0) := \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h},$$

siempre que este límite exista. En tal caso, diremos que f es **diferenciable** en x_0.

Si f es diferenciable en cualquier $x \in \mathbb{R}$, diremos que es **diferenciable en todas partes**. Si f es diferenciable en un x_0 dado, entonces es necesariamente continua en x_0 como veremos en la siguiente proposición. Sin embargo continuidad no implica diferenciable como se ilustra con función de Kink.

Teorema 2.1 (Diferenciabilidad implica continuidad)**.** *Si $f : \mathbb{R} \to \mathbb{R}$ es diferenciable en $x_0 \in \mathbb{R}$, entonces es continua en $x_0 \in \mathbb{R}$.*

Demostración. Como f es diferenciable en x_0, existe el límite

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) - f(x_0)}{h}.$$

Por lo tanto, podemos escribir la diferencia como

$$f(x_0 + h) - f(x_0) = f'(x_0)h + r(h),$$

donde $r(h)$ es un término residual que satisface $r(h) = o(h)$ cuando $h \to 0$. En particular,

$$\lim_{h \to 0}[f(x_0 + h) - f(x_0)] = \lim_{h \to 0}[f'(x_0)h + r(h)] = 0,$$

ya que $\lim_{h \to 0} h = 0$ y $\lim_{h \to 0} r(h) = 0$. Por lo tanto,

$$\lim_{h \to 0} f(x_0 + h) = f(x_0),$$

es decir, f es continua en x_0. □

La derivada nos permitir construir una aproximación lineal de una función f en una vecindad de x_0, ya que representa la pendiente de la recta tangente de f en x_0. Por otro lado, también nos da información sobre la **monotonicidad** de f alrededor de x_0.

Si $f'(x_0)$ es positiva, la función es creciente alrededor de x_0. Si $f'(x_0)$ es negativa, la función es decreciente. Esta propiedad puede aprovecharse para desarrollar algoritmos iterativos minimicen f, generando una sucesión de valores de la forma $x_{t+1} = x_t - \gamma f'(x_t)$ para $\gamma > 0$, donde t representa el número de iteración. Estos valores se desplazan siguiendo las direcciones de f alrededor de x_t.

Para muchas funciones elementales tales como x^n, e^x, $\ln x$, $\cos x$, $\sin x$, sus derivadas se pueden calcular directamente aplicando la definición 2.3 como ilustraremos en el siguiente ejemplo.

Ejemplo 2.1 (Derivada de la función potencia). Considere la función $f(x) = x^n$ para $x \in \mathbb{R}$, $n \in$

$\mathbb{N}\setminus\{0\}$. Para cualquier $h\in\mathbb{R}$, tenemos:

$$\begin{aligned}\frac{f(x+h)-f(x)}{h} &= \frac{(x+h)^n-x^n}{h}\\ &= \frac{\sum_{k=0}^{n}\binom{n}{k}h^k x^{n-k}-x^n}{h}\\ &= \sum_{k=1}^{n}\binom{n}{k}h^{k-1}x^{n-k}\\ &= \binom{n}{1}x^{n-1}+\sum_{k=2}^{n}\binom{n}{k}h^{k-1}x^{n-k},\end{aligned}$$

en donde, en la segunda linea, usamos el teorema del binomio. Dado que:

$$\binom{n}{1}=n \text{ y } \lim_{h\to 0}\sum_{k=2}^{n}\binom{n}{k}h^{k-1}x^{n-k}=0,$$

obtenemos que $f'(x)=nx^{n-1}$.

Observación (Funciones definidas en un subconjunto $U\subset\mathbb{R}$). Si bien la definición de derivada se presentó inicialmente para funciones definidas en todo $\mathbb{R}$, esta puede extenderse naturalmente a funciones $f:U\subset\mathbb{R}\to\mathbb{R}$ definidas en un subconjunto U de los números reales (como es el caso de $f(x)=\sqrt{x}$ definida sobre $\mathbb{R}^+$). Para $x\in U$, la derivada de f en x viene dada por el límite de la definición 2.1, siempre que f esté bien definida en una vecindad de x. Es decir, debe existir $r>0$ tal que $x+\varepsilon\in U$ para todo $|\varepsilon|\leq r$. Decimos que f es **diferenciable en todas partes** si es diferenciable en cada punto x del **interior** de U, donde el interior es el conjunto de puntos $x\in U$ tales que existe $r>0$ con $x+\varepsilon:|\varepsilon|\leq r\subseteq U$. Para los puntos ubicados en la frontera de U (como a y b cuando $U=[a,b)$), se pueden definir las derivadas laterales: la derivada por derecha en a y la derivada por izquierda en b, tomando los límites correspondientes al acercarnos a estos puntos desde el interior del intervalo.

Para $x_0\in\mathbb{R}$ y funciones $f,g:\mathbb{R}\to\mathbb{R}$ diferenciables en x_0, entonces:

- **Linealidad:** $(af+bg)'(x_0)=af'(x_0)+bg'(x_0)$, $a,b\in\mathbb{R}$
- **Regla del producto:** $(fg)'(x_0)=f'(x_0)g(x_0)+f(x_0)g'(x_0)$

La linealidad se deriva directamente de la definición de derivada y las propiedades de límites. Para la regla del producto, usando notación $o(h)$:

$$\begin{aligned}(fg)(x_0+h) &= f(x_0+h)g(x_0+h)\\ &= \big(f(x_0)+f'(x_0)h+o(h)\big)\big(g(x_0)+g'(x_0)h+o(h)\big)\\ &= f(x_0)g(x_0)+f'(x_0)g(x_0)h+f(x_0)g'(x_0)h+o(h),\end{aligned}$$

de donde se deduce que:

$$\frac{(fg)(x_0+h)-f(x_0)g(x_0)}{h} = f'(x_0)g(x_0)+f(x_0)g'(x_0)+\frac{o(h)}{h}.$$

Si g es diferenciable en x_0 y f es diferenciable en $g(x_0)$, entonces:

- **Regla de la cadena:** $(f\circ g)'(x_0) = f'\big(g(x_0)\big)g'(x_0)$.

Como hemos visto, la linealidad y la regla del producto son subproductos de la regla de la cadena, siendo esta última fundamental para la diferenciabilidad.

Consideremos una función expresada mediante sumas, productos o composición de funciones elementales, como $f(x) = e^x \ln x + \cos x^2$. Su derivada puede calcularse descomponiendo la función en operaciones elementales y aplicando las reglas de linealidad, producto y composición, como ilustraremos a continuación.

Ejemplo 2.2. Aplicando la reglas de diferenciabilidad Consideremos la función $f(x) = e^x \ln x + \cos x^2$. La derivada de f sobre $x > 0$ puede calcularse paso a paso como sigue, denotando $\mathrm{sq}(x) := x^2$,

$$\begin{aligned}
&f'(x) = (\exp\cdot\ln)'(x) + (\cos\circ \mathrm{sq})'(x) && \text{(Linealidad)}\\
&(\exp\cdot\ln)'(x) = \exp'(x)\cdot\ln(x) + \exp(x)\cdot\ln'(x) && \text{(Regla del producto)}\\
&(\cos\circ \mathrm{sq})'(x) = \cos'(\mathrm{sq}(w))\,\mathrm{sq}'(x) && \text{(Regla de la cadena)}\\
&\exp'(x) = \exp(x), \qquad \ln'(x) = 1/x, && \text{(Función elemental)}\\
&\mathrm{sq}'(x) = 2x, \qquad \cos'(x) = -\sin(x). && \text{(Función elemental)}
\end{aligned}$$

Para finalmente obtener que $f'(x) = e^x \ln x + e^x/x - 2x\sin x^2$.

El proceso de derivación es puramente mecánico y conduce en si mismo a un proceso automatizado, el cual constituye la idea principal de la diferenciación automática.

2.1.2. Notación de Leibniz

La idea de derivada fue introducida independiente por Newton y Leibniz en el siglo XVIII [Ball, 1960]. Posteriormente, las derivadas fueron consideradas como el cociente de variaciones infinitesimales. Específicamente, denotando $u = f(x)$ como una variable que depende de x a través de f, Leibniz consideró la derivada de f como el cociente

$$f'(x) = \left.\frac{du}{dx}\right|_x$$

donde du y dx denota las variaciones infinitesimales de u y x respectivamente y el símbolo $|_x$ denota la evaluación de la derivada en un punto x. Esta notación simplifica la enunciación de la regla de la cadena. En efecto, si tenemos $v = g(x)$ y $u = f(v)$, entonces

$$\frac{du}{dx} = \frac{du}{dv}\frac{dv}{dx}.$$

Estos nos ayuda a ver que las derivadas son multiplicadas cuando consideremos composiciones. En la evaluación, la regla de cadena en la notación de Leibniz expresa de la siguiente forma:

$$\left.\frac{du}{dx}\right|_x = \left.\frac{du}{dv}\right|_{g(x)} \left.\frac{dv}{dx}\right|_x = f'\big(g(x)\big)g'(x) = (f \circ g)'(x).$$

La capacidad de la notación de Leibniz para capturar la regla de la cadena como un simple producto de cocientes la hizo popular a los largo de los siglos, especialmente en mecánica [Ball, 1960]. La lógica detrás de la notación de Leibniz, es decir, el concepto de «variaciones infinitesimales», fue cuestionada porsteriormente por matemáticos debido a sus posibles problemas lógicos [Ball, 1960]. La notación $f'(x)$, introducida por primera vez por Euler y posteriormente popularizada por Lagrange [Cajori, 2007], ha predominado en numerosos libros de texto matemáticos. El concepto de variaciones infinitesimales ha sido definido rigurosamente al considerar el conjunto de los números hiperreales. Esto amplían el conjunto de números reales al considerar cada número como la suma de una parte no infinitesimal y una parte infinitesimal [Hewitt, 1948]. El formalismo de las variaciones infinitesimales también sustenta el desarrollo de algoritmos de diferenciación automática mediante el concepto de números duales.

2.2. Funciones multivariadas

2.2.1. Derivada direccional

Consideremos ahora una función $f : \mathbb{R}^m \to \mathbb{R}$ con $\boldsymbol{x} = x_0, \ldots, x_m \in \mathbb{R}^m$. La mayoría de la ejemplos importantes en aprendizaje computacional es una función en la cual $\boldsymbol{x} \in \mathbb{R}^n$ son los parámetros de una red

neuronal, asociados a una función de perdida con valores en $\mathbb{R}$. la variaciones de f necesitan ser definidas a través de direcciones específicas, como por ejemplo la variación $f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{x})$ de f alrededor de $\boldsymbol{x}\in\mathbb{R}^m$ en la dirección de $\boldsymbol{v}\in\mathbb{R}^m$ por una cantidad $h>0$. Esta consideración nos conduce a la definición de derivada direccional.

Definición 2.4 (Derivada direccional)**.** La **derivada direccional** de f en $\boldsymbol{x}$ en la dirección $\boldsymbol{v}$ esta dada por:

$$\partial f(\boldsymbol{x})[\boldsymbol{v}] := \lim_{h\to 0}\frac{f(\boldsymbol{x}+h\boldsymbol{v})-f(\boldsymbol{v})}{h},$$

siempre que la derivada exista.

Un ejemplo de derivada direccional consiste en calcular la derivada de la función f en $\boldsymbol{x}$ en cualquiera de las direcciones canónicas

$$\boldsymbol{e}_i := (0,\ldots,\underbrace{1}_{i},0,\ldots,0).$$

Esto nos permite definir la noción de **derivadas parciales**, denotada para $i\in[0,m]_{\mathbb{Z}}$

$$\partial_{x_i} f(\boldsymbol{x}) := \partial f(\boldsymbol{x})[\boldsymbol{e}_i] = \lim_{h\to 0}\frac{f(\boldsymbol{x}+h\boldsymbol{e}_i)-f(\boldsymbol{x})}{h}.$$

Al moverse solamente a los largo de la coordenada i - ésima de la función, la derivada parcial es similar a usar la función $\psi(x_i)=f(x_1,\ldots,x_i,\ldots,x_m)$ alrededor de x_1, manteniendo todas las demás coordenadas fijas en sus valores x_i.

2.2.2. Gradientes

Ahora introduciremos el vector gradiente, el cual reune a todas las derivadas parciales. Recordemos primero las definiciones de transformación lineal y forma lineal.

Definición 2.5 (Transformación lineal, formal lineal)**.** Una función $L:\mathbb{R}^m\to\mathbb{R}^n$ es una **transformación lineal** si para cualquier $a,b\in\mathbb{R}$ y $\boldsymbol{u},\boldsymbol{v}\in\mathbb{R}^m$,

$$L(a\boldsymbol{u}+b\boldsymbol{v}) = aL(\boldsymbol{u})+bL(\boldsymbol{v}).$$

Una transformación lineal con valores en $\mathbb{R}$, $L:\mathbb{R}^m\to\mathbb{R}$, se le conoce como **forma lineal**.

La linealidad juega un rol crucial en la diferenciabilidad de una función.

Definición 2.6 (Derivada de una función $f:\mathbb{R}^m \to \mathbb{R}$)**.** . Una función $f:\mathbb{R}^m \to \mathbb{R}$ es diferenciable en $\boldsymbol{x} \in \mathbb{R}^m$ si las derivadas direccionales esta definidas en cualquier dirección, son lineales en cualquier dirección y se cumple que

$$\lim_{||\boldsymbol{v}||_2 \to 0} \frac{|f(\boldsymbol{x}+\boldsymbol{v}) - f(\boldsymbol{v}) - \partial f(\boldsymbol{x})[\boldsymbol{v}]|}{||\boldsymbol{v}||_2} = 0.$$

A continuación introducimos la idea de gradiente.

Definición 2.7 (Gradiente)**.** El **gradiente** de una función diferenciable $f:\mathbb{R}^m \to \mathbb{R}$ en un punto $\boldsymbol{x} \in \mathbb{R}^m$ es definido como el vector de derivadas parciales

$$\nabla f(\boldsymbol{x}) := \begin{bmatrix} \partial_{x_1} f(\boldsymbol{x}) \\ \vdots \\ \partial_{x_m} f(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} \partial f(\boldsymbol{x})[\boldsymbol{e}_1] \\ \vdots \\ \partial f(\boldsymbol{x})[\boldsymbol{e}_m] \end{bmatrix}$$

En virtud de la linealidad, la derivada direccional de f en $\boldsymbol{x}$ en la dirección de $\boldsymbol{v} = \sum_{i=1}^{m} v_i \boldsymbol{e}_i$ está dada por

$$\partial f(\boldsymbol{x})[\boldsymbol{v}] = \sum_{i=1}^{m} v_i \partial f(\boldsymbol{x})[\boldsymbol{e}_i] = \langle \boldsymbol{v}, \nabla f(\boldsymbol{x}) \rangle.$$

Aquí, $\langle \cdot, \cdot \rangle$ denota el producto interno.

En la definición anterior, el hecho de poder usar el gradiente para calcular la derivada direccional, es consecuencia de la linealidad. Sin embargo, para casos más abstractos, el gradiente se define a través de esta propiedad.

Un ejemplo, cualquier transformación lineal de la forma $f(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle = \sum_{i=1}^{m} a_i x_i$ es diferenciable, efecto tenemos que

$$\lim_{||\boldsymbol{v}||_2 \to 0} \frac{\langle \boldsymbol{a}, \boldsymbol{x}+\boldsymbol{v} \rangle - \langle \boldsymbol{a}, \boldsymbol{x} \rangle - \langle \boldsymbol{a}, \boldsymbol{v} \rangle}{||v||_2} = 0.$$

De donde se concluye que $\nabla f(\boldsymbol{x}) = \boldsymbol{a}$.

Generalmente, para mostrar que una función es diferenciable y calcular su gradiente, un enfoque es aproximarse a $f(\boldsymbol{x}+\boldsymbol{v})$ alrededor de $\boldsymbol{v} = 0$. Si podemos encontrar un vector $\boldsymbol{g}$ tal que

$$f(\boldsymbol{x}+\boldsymbol{v}) = f(\boldsymbol{x}) + \langle \boldsymbol{g}, \boldsymbol{v} \rangle + o(||\boldsymbol{v}||_2),$$

entonces f es diferenciable en $\boldsymbol{x}$ dado que $\langle \boldsymbol{g}, \cdot \rangle$ es lineal. Por lo tanto, $\boldsymbol{g}$ sería el gradiente de f en $\boldsymbol{x}$.

Observación (Funciones diferenciables de Gateux y Fréchet). Existen diferentes definiciones de diferenciabilidad. Una de ellas es la definición 2.6 de **diferenciabilidad de Fréchet**. Alternativamente, si $f : \mathbb{R}^m \to \mathbb{R}$ tiene derivadas direccionales bien definidas a través de todas las direcciones entonces se dice que la función es **diferenciable de Gateaux**. Observe que el la existencia de las derivadas direccionales en cualquiera de sus direcciones no es condición suficiente para que la función sea diferenciable. En otras palabras, cualquier función diferenciable en el sentido de Fréchet en diferenciable en el sentido de Gateaux, en el sentido inverso no es verdad. Como contrajemplo, una puede verificar que la función $f(x_1, x_2) = \frac{x_1^3}{x_1^2 + x_2^3}$ es Gateaux diferenciable pero no lo es Fréchet diferenciable en 0 (porque la derivada direcional no es lineal en 0).

Algunos autores también exigen que las funciones diferenciables de Gateaux tengan derivadas lineales direccionales a lo largo de cualquier dirección. Estas funciones aún no son funciones diferenciables de Fréchet. De hecho, el límite de la definición 2.6 es sobre cualquier vector que tienda a 0 (potencialmente de forma patológica), mientras las derivadas direccionales consideran tales límites exclusivamente en términos de una única dirección.

En el resto del capítulo, todas las definiciones de diferenciabilidad se basan en la diferenciabilidad en el sentido de Fréchet. En el siguiente ejemplo se ilustra cómo calcular el gradiente de la perdida logística y validad sus diferenciabilidad.

Ejemplo 2.3 (Gradiente de la pérdida logística). Consideremos la función de pérdida logística

$$l(\boldsymbol{\theta}, \boldsymbol{y}) := -\langle \boldsymbol{y}, \boldsymbol{\theta} \rangle + \log \sum_{i=1}^{M} \exp(\theta_i),$$

que mide el error en las predicciones de los logits $\boldsymbol{\theta} \in \mathbb{R}$ para la etiqueta correcta $\boldsymbol{y} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_M\}$. Lo que nos permite computar el gradiente de esta función de pérdida con respecto a $\boldsymbol{\theta}$ para un valor fijo de $\boldsymbol{y}$, es decir, nosotros queremos calcular el gradiente de $f(\boldsymbol{\theta}) := l(\boldsymbol{\theta}, \boldsymbol{y})$. Esto nos permite descomponer a f como $f = l + \text{logsumexp}$ con $l(\boldsymbol{\theta}) := \langle -\boldsymbol{y}, \boldsymbol{\theta} \rangle$ y

$$\text{logsumexp}(\boldsymbol{\theta}) := \log \sum_{i=1}^{M} \exp(\theta_i),$$

la función log-sum-exp. La función l es lineal y diferenciable con gradiente $\nabla l(\theta) = -\boldsymbol{y}$. Nosotros por lo tanto nos focalizamos sobre logsumexp. Denotando $\exp(\boldsymbol{\theta}) = \big(\exp(\theta_1), \ldots, \exp(\theta_M)\big)$, usando $\exp(x) = 1 + x + o(x)$, $\log(1+x) = x + o(x)$, y denotando por $\odot$ el producto elemento a elemento, así nosotros tenemos:

$$\begin{aligned}
\operatorname{logsumexp}(\boldsymbol{\theta} + \boldsymbol{v}) &= \log(\langle \exp(\boldsymbol{\theta} + \boldsymbol{v}), \mathbf{1} \rangle) \\
&= \log\big(\langle \exp(\boldsymbol{\theta}) \odot \exp(\boldsymbol{v}), \mathbf{1} \rangle\big) \\
&= \log\big(\langle \exp(\boldsymbol{\theta}) \odot (1 + \boldsymbol{v} + o(\|\boldsymbol{v}\|_2)), \mathbf{1} \rangle\big) \\
&= \log\big(\langle \exp(\boldsymbol{\theta}), \mathbf{1} \rangle + \langle \exp(\boldsymbol{\theta}), \boldsymbol{v} \rangle + o(\|\boldsymbol{v}\|_2)\big) \\
&= \log\big(\langle \exp(\boldsymbol{\theta}), \mathbf{1} \rangle\big) + \left\langle \frac{\exp(\boldsymbol{\theta})}{\langle \exp(\boldsymbol{\theta}), \mathbf{1} \rangle}, \boldsymbol{v} \right\rangle + o(\|\boldsymbol{v}\|_2),
\end{aligned}$$

La anterior descomposición de $\operatorname{logsumexp}(\boldsymbol{\theta} + \boldsymbol{v})$ muestra que es diferenciable, y que $\nabla \operatorname{logsumexp}(\boldsymbol{\theta}) = \operatorname{softargmax}(\boldsymbol{\theta})$ donde,

$$\operatorname{softargmax}(\boldsymbol{\theta}) := \left(\frac{\exp(\theta_1)}{\sum_{j=1}^{M} \exp(\theta_j)}, \ldots, \frac{\exp(\theta_M)}{\sum_{j=1}^{M} \exp(\theta_j)} \right).$$

En total, tenemos que $\nabla f(\boldsymbol{\theta}) = -\boldsymbol{y} + \operatorname{softargmax}(\boldsymbol{\theta})$.

Referencias

W.W.R. Ball. *A Short Account of the History of Mathematics*. Dover Books on Mathematics Series. Dover Publications, 1960. ISBN 9780486206301.

F. Cajori. *A History of Mathematical Notations*. Number v. 1 in A History of Mathematical Notations. Cosimo, Incorporated, 2007. ISBN 9781602066854.

M.P. Deisenroth, A.A. Faisal, and C.S. Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020. ISBN 9781108470049.

E. Hewitt. Rings of real-valued continuous functions. i. *Transactions of the American Mathematical Society*, 64(1):45–99, 1948.