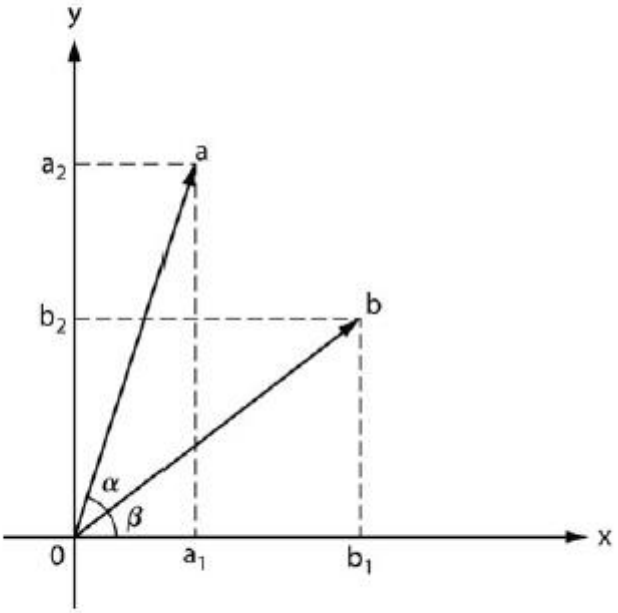
Lecturas complementarias

Norma vectorial

La norma de un vector puede ser pensada como su longitud. De hecho, la denominada *distancia euclídea* es una forma particular de norma. Con el siguiente gráfico, se puede apreciar cómo la norma de un vector **a**, denotada como ||**a**|| puede obtenerse siguiendo este razonamiento a través del teorema de Pitágoras:

$$\|\mathbf{a}\| = \sqrt{a_1^2 + a_2^2}$$



En general, si denominamos a_j a las distintas componentes de un vector y ‹**a**,**a**› al producto vectorial de a consigo mismo, se puede escribir para vectores de más dimensiones:

$$\|\mathbf{a}\| = \left(\sum_{j=1}^{n} a_j^2\right)^{\frac{1}{2}} = \langle \mathbf{a}, \mathbf{a} \rangle^{\frac{1}{2}}$$

Autovalores y autovectores

Algunos conceptos algebraicos son difíciles de asociar a situaciones y conceptos relacionados con problemas reales. Este puede ser el caso de los autovalores y autovectores. Ambos conceptos tienen contrapartes físicas. Por ejemplo, las frecuencias de una cuerda vibratoria son autovalores en realidad. Pero no ejemplos intuitivamente atractivos o fáciles en otros campos. Sin embargo, sí que tienen una contraparte en las cadenas de Markov (usados en el Deep Learning), con una explicación adecuada. Por lo tanto, el siguiente material, al menos al principio, se verá demasiado abstracto. Pero tened paciencia porque entender estos conceptos les darán una ventaja comparativa importante.

Los autovalores y autovectores en sí mismos son de importancia e inmensa ayuda en álgebra matricial, y lo que es más importante, permiten una factorización de matrices en tres, y en el caso de ciertas matrices en dos (algo parecido a tomar la raíz cuadrada de una matriz). Este dispositivo facilita enormemente la comprensión de muchas técnicas de estimación en y la solución de sistemas de ecuaciones diferenciales. De esta forma, si **A** es una matriz cuadrada de orden n, es decir, con n filas y n columnas, si podemos encontrar un escalar λ (número real) y un vector **x** tal que:

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$$

lo que es equivalente a

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

entonces, λ y **x** son denominados autovalor y autovector de la matriz **A.** Obviamente, la segunda ecuación tiene una solución trivial en **x=0,** solución que en un principio no queremos. Para evitarlo, forzamos a (**A-** λ**I**) a que tenga un determinante igual a cero:

$$|\mathbf{A} - \lambda\mathbf{I}| = 0$$

De esta forma, el polinomio formado por

$$P(\lambda) = |\mathbf{A} - \lambda\mathbf{I}|$$

Se denomina "polinomio característico" de la matriz **A**. Y su solución al igualarlo a cero nos proporciona los autovalores λ de la matriz **A**. Los autovectores **x** se obtienen a partir del sistema de ecuaciones obtenido de

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

Muchas veces, se encuentran autovalores repetidos, es decir, con multiplicidad mayor que uno. En este caso, todo permanece invariable, pero el cálculo posterior de autovectores puede verse afectado y habría que recurrir al concepto de autovector generalizado, que escapa en cierta manera al ámbito de este curso, pues no afecta a los algoritmos que se presentarán más adelante.

Derivadas parciales y Vectores gradientes

Derivadas parciales

El concepto de vector gradiente surge de la generalización del concepto de derivada a funciones de más de una variable. La mayor parte de las relaciones que se intentan predecir en Deep Learning son multidimensionales, y muchas veces, puede interesarnos medir el efecto que provoca en el valor de la función, un cambio en solo una de las variables. Ese es el concepto de derivada parcial.

La idea de derivada de una función de una variable que vimos en el documento de la primera videoconferencia puede ser extendido a funciones de varias variables. Consideremos por ejemplo una función de dos variables

$$y = f(x_1,\ x_2)$$

La derivada de *y* respecto a x_1 identifica el cambio provocado en *y* ante una variación (tendente a cero o infinitesimal) de x_1 manteniendo x_2 constante. Lo mismo se aplica a la derivada parcial de y respecto a x_2: identifica el cambio provocado en *y* ante una variación (tendente a cero o infinitesimal) de x_2 manteniendo x_1 constante. La notación para las derivadas parciales es:

$$\frac{\partial y}{\partial x_1} = f_1(x_1, x_2)$$

$$\frac{\partial y}{\partial x_2} = f_2(x_1, x_2)$$

Obviamente, esta definición es extensible a funciones de más de dos variables.

Vector Gradiente

Consideremos la siguiente función de n variables:

$$y = f(\mathbf{x})$$
$$\mathbf{x} = [x_1, x_2, \ldots, x_n]'$$

Entonces, el vector formado por las derivadas parciales de la función *f* respecto a todas sus variables es el vector gradiente de dicha función, y se denota:

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \partial f/\partial x_1 \\ \vdots \\ \partial f/\partial x_n \end{bmatrix}$$

El vector gradiente es un concepto importante, pues indica la dirección de cambio de la función cuando todas las variables que la componen cambian simultáneamente. Se usa mucho en métodos numéricos de optimización, pues un gradiente positivo indica que el valor de la función está creciendo al variar todas sus variables. También es la “pendiente” del hiperplano

tangente a la función, por lo que, en un punto máximo o mínimo, el gradiente de una función es igual a cero.

Derivadas de segundo orden y matriz hessiana

Derivadas de segundo orden y orden superior

Las derivadas de una función son, a su vez, funciones, y como tales, si cumplen las características de continuidad y diferenciabilidad, pueden ser derivadas de nuevo, obteniendo de esta forma las derivadas de segundo orden de la función original. Incluso, si la función resultante de la segunda derivada es derivable, el proceso podría volver a repetirse.

Para una función de una variable $y=f(x)$, la notación usada será:

Para la derivada de primer orden (la normal)

$$y' = \frac{dy}{dx}$$

para la segunda derivada o derivada de segundo orden

$$y'' = \frac{d}{dx}\frac{dy}{dx} = \frac{d^2y}{dx^2}$$

Para la derivada de tercer orden

$$y''' = \frac{d^3y}{dx^3}$$

Y en general, para la derivada de orden n

$$y^{(n)} = f^{(n)}(x) = \frac{d^n y}{dx^n}$$

Matriz Hessiana

Volvamos a considerar la función de n variables anteriormente usada para el vector gradiente:

$$y = f(\mathbf{x})$$
$$\mathbf{x} = [x_1, x_2, \ldots, x_n]'$$

La matriz que contiene las derivadas parciales de segundo orden de la función respecto a cada una de las variables se denomina matriz hessiana, y se denota por:

$$\nabla^2 f(\mathbf{x}) = \left[\frac{\partial^2 f}{\partial x_i \partial x_j}\right] \qquad i,j = 1, \ldots n$$

Es decir, el elemento a_{ij} de la matriz hessiana es el elemento que contiene la derivada respecto a la variable x_j de la derivada de $f(\mathbf{x})$ respecto a la variable x_i, es decir, incluye las derivadas de

segundo orden sobre cada una de las derivadas de primer orden que componen el vector gradiente. Por decirlo de alguna manera, cada fila i del Hessiano es el vector gradiente del componente i del vector gradiente original. En los ejercicios resueltos se van a ver muy claramente.

Optimización sin restricciones

Como ya se introdujo, se trata de encontrar máximos o mínimos de funciones, donde normalmente, las variables están sujetas a algún tipo de restricción:

Como regla general, puede verse como en los puntos máximos y mínimos, la pendiente de la función es cero, es decir, en un máximo o mínimo x*:

$$f'(x^*) = 0$$

Si es un máximo, tras ese punto, la función decrecerá, por lo que la segunda derivada será negativa:

$$f''(x^*) < 0$$

Esta segunda derivada será positiva si el punto es un mínimo. Todo esto nos ofrece las condiciones que debe cumplir un máximo y un mínimo.

En funciones de varias variables, en lugar de la derivada usamos el vector gradiente, y en lugar de la segunda derivada, usamos la matriz hessiana, de forma que, la primera condición quedaría

$$\nabla f(\mathbf{x}^*) = 0$$

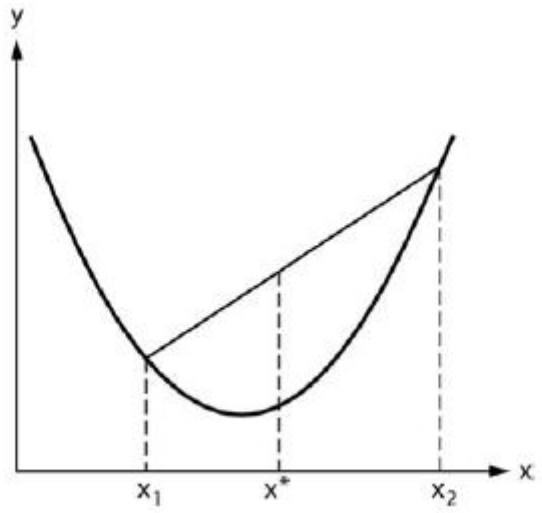
mientras que la segunda quedaría:

$$\mathbf{z}'\nabla^2 f(\mathbf{x}^*)\mathbf{z} = \mathbf{z}'\mathbf{H}(\mathbf{x}^*)\mathbf{z} < 0$$

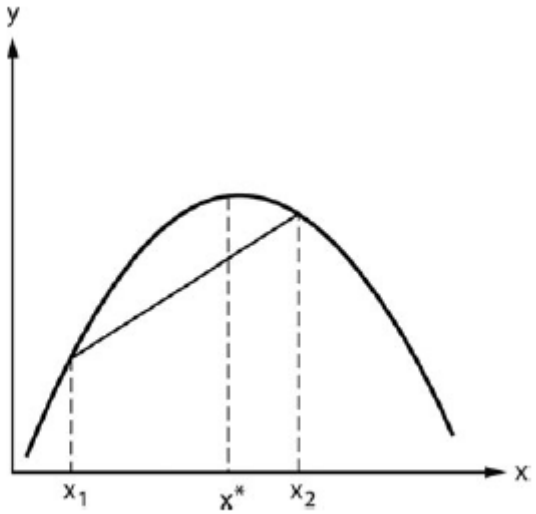
para todo vector **z** diferente a cero. Es decir, que la matriz hessiana sea definida negativa.

Una idea que es necesaria tener en mente es que los mínimos se dan en situaciones de convexidad, mientras que los máximos, se dan en situaciones de concavidad:

Convexidad



Concavidad



Optimización numérica

Muchas veces nos encontramos con funciones que no son cóncavas ni convexas y, por tanto, no se pueden encontrar sus máximos o mínimos igualando sus derivadas o gradientes a cero. Para solucionarlo, se usan métodos numéricos. La idea de estos métodos es usar el gradiente de la función, que determina la dirección de crecimiento de esta, e ir avanzando en esta dirección poco a poco hasta que llegue un punto donde no crezca más (convergencia). Es decir, se parte de un punto, y a la siguiente iteración, intentamos ir a un punto donde el valor de la

función se mejore, lo que viene determinado por el gradiente. Así, desde el punto x_{k-1}, nos moveremos al punto x_k, que viene determinado por:

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \lambda_k \mathbf{d}_k$$

donde λ_k es un escalar y $\mathbf{d}_k$ un vector que determina la dirección de nuestro movimiento. El problema es determinar estos λ_k y $\mathbf{d}_k$, sin embargo, ya hemos dicho que el gradiente es la dirección de incremento de la función, por lo que, si queremos encontrar un mínimo:

$$\mathbf{d}_k = -\nabla f(\mathbf{x}_k)$$

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \lambda_k \nabla f(\mathbf{x}_k)$$

que es la expresión del método del gradiente. Sin embargo, si hacemos

$$\mathbf{d}_k = -[\nabla^2 f(\mathbf{x})]^{-1} \nabla f(\mathbf{x})$$

estaremos hablando del método de Newton.

Para determinar λ_k existen numerosos métodos, como el de la razón de oro, pero estos métodos quedan fuera del alcance de este curso.

Optimización con restricciones

Como se comentó, lo habitual es encontrar problemas donde los valores de la variables están restringidos, bien por restricciones de igualdad, bien por restricciones de desigualdad:

Restricciones de igualdad:

$$\begin{aligned} &\max f(\mathbf{x}), \qquad \mathbf{x} \in \Re^n \\ &\text{subject to} \quad g_i(\mathbf{x}) = 0, \quad i = 1, \ldots, m \end{aligned}$$

Donde se intenta maximizar una función de n variables sujeta a m restricciones.

Restricciones de desigualdad:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t.} \quad g_i(\mathbf{x}) \leq 0 \qquad i = 1, \ldots, m \end{aligned}$$

Restricciones de igualdad y desigualdad:

$$\begin{aligned} &\min f(\mathbf{x}) \\ &\text{s.t.} \quad g_i(\mathbf{x}) \leq 0 \qquad i = 1, \ldots, m \\ &\qquad h_j(\mathbf{x}) = 0 \qquad j = 1, \ldots, J \end{aligned}$$

Este tipo de problemas se resuelve, generalmente, mediante el uso de multiplicadores de Lagrange o lagrangianos. La esencia es reemplazar la función objetivo por la función lagrangiana. Suponiendo que tenemos una sola restricción, esta función sería:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda g(\mathbf{x})$$

Con m restricciones, sería:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g(\mathbf{x})$$

$$\boldsymbol{\lambda} = [\lambda_1 \ \ldots \ \lambda_m]'$$

Y las condiciones de optimalidad que deben cumplir los puntos máximos o mínimos son:

$$\nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0}$$

Para ser máximo, la Hessiana deberá ser definida negativa y para ser mínimo, definida positiva.

En el caso de que el problema contenga tanto restricciones de igualdad como de desigualdad, se asignarán multiplicadores diferentes a cada tipo de restricción y estos, a su vez, deberán cumplir una serie de condiciones. Resumiendo, para un problema de minimización, la condición para que un punto **x*** sea máximo o mínimo es:

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\mathbf{x}^*) + \sum_{j=1}^{J} \mu_j^* \nabla h_j(\mathbf{x}^*) = 0$$
$$\lambda_i^* \geq 0, \qquad g_i(\mathbf{x}^*) \leq 0, \qquad \lambda_i^* g_i(\mathbf{x}^*) = 0 \qquad i = 1, \ldots, m$$
$$h_j(\mathbf{x}^*) = 0 \qquad j = 1, \ldots, J$$

donde g_i son restricciones de desigualdad y h_j de igualdad.

Nótese que el problema de maximizar $f(\mathbf{x})$ es equivalente a minimizar $-f(\mathbf{x})$

Con toda esta información no se pretende que podáis resolver manualmente problemas de optimización, pero sí que comprendás un poco la naturaleza de algunos algoritmos como el Suport Vector Machine, cuyos parámetros vienen determinados por problemas de maximización usando estos conceptos.

Dualidad en optimización

Me interesa mucho que comprendáis este concepto aún sin ahondar mucho en él o en sus repercusiones o formulación matemática.

Cada problema de optimización tiene un problema gemelo o doble. Bajo ciertas condiciones, la solución de uno conlleva la solución del otro. De manera intuitiva, supongamos que estamos asignando las cantidades disponibles de tres recursos, trabajo, capital y energía, por ejemplo, para dos productos. Dado el precio de los productos, el *problema primal* es encontrar la **cantidad** de cada recurso que se asignará a cada actividad para **maximiza**r una función

objetivo que puede ser una función de **beneficio**. Por otro lado, podemos plantear un *problema dual* al preguntar cómo podríamos calcular el **precio** de nuestros recursos para **minimizar** el **costo** de producir al menos un nivel dado de cada producto.
La importancia de la dualidad en economía es comprender el significado subyacente de un problema de optimización, en particular, el papel de los multiplicadores de Lagrange como óptimo de los precios de los recursos.