Notas de Optimización

Martin D. Maas, PhD

27 de marzo de 2025

Índice general

Prefacio

Estas son las notas de clase de la materia "Optimización" del Departamento de Matemática de la Facultad de Ciencias Exactas de la UBA. Se trata de un segunda materia de la temática, que se cursa a continuación de una primera parte que cubre temas importantes como programación lineal y métodos para algunos problemas de optimización discreta. Por ese motivo, la materia hará énfasis en los métodos de programación continua y no-lineal, centrándose en el desarrollo de métodos rigurosos y eficientes, para retomar la resolución de problemas discretos sobre el final del curso. Como también veremos, tal vez de forma anti-intutiva, los métodos continuos pueden muchas veces ayudar a resolver problemas discretos de grandes dimensiones.

Con el correr de los años, la optimización cada vez se aplica a problemas más desafiantes que surgen en el mundo de la ciencia y la tecnología. Por ejemplo, en aprendizaje automático, los modelos de redes neuronales profundas requieren optimizar miles de millones de parámetros simultáneamente, lo que implica resolver problemas de optimización en espacios de dimensiones muy grandes. En genómica, los algoritmos de optimización permiten alinear secuencias de ADN y predecir estructuras proteicas con una precisión impensada hace apenas una década. En logística global, las cadenas de suministro utilizan complejos modelos de optimización para minimizar costos y tiempos de entrega en redes que involucran cientos de miles de nodos y restricciones. En áreas como la geofísica, los algoritmos de optimización permiten reconstruir imágenes del subsuelo a partir de complejos datos de reflexión, combinando técnicas de inversión geofísica que procesan terabytes de información para mapear estructuras geológicas con sumo detalle. Las distintas áreas de la ingeniería utilizan métodos de optimización para diseñar componentes de máquinas, pequeños artefactos o edificios, donde encuentran configuraciones de materiales que maximizan la performance mientras minimizan el peso o el tamaño, generando diseños que desafían la intuición tradicional.

Como consecuencia de estos desafíos la optimización se ve obligada a ocuparse de sistemas de grandes dimensiones. Este salto cualitativo requiere desarrollar métodos que no solo exploren estos espacios, sino que lo hagan de manera eficiente y rigurosa, reduciendo la complejidad computacional y encontrando soluciones óptimas o cercanas al óptimo en tiempos razonables.

Un tema de recurrente discusión en la pedagogía de la matemática aplicada es cómo encontrar un buen balance entre la teoría y las aplicaciones. En el campo de la optimización, este dilema se torna particularmente difícil de resolver: por un lado, las aplicaciones son muy numerosas y provienen de prácticamente todas las áreas de la ciencia y la tecnología. Naturalmente, los métodos para resolver problemas de optimización también son variados, y en la aplicación práctica se combinan aspectos interdisciplinarios con métodos heurísticos, lo que dificulta el desarrollo de una teoría que englobe la totalidad de la disciplina y sus aplicaciones.

En principio, cabe mencionar que este apunte concentra los contenidos de las clases

teóricas, y por lo tanto naturalmente enfatiza el desarrollo de la teoría. Sin embargo, hemos acompañado el desarrollo de estas notas con una selección de ejercicios prácticos de aplicaciones. Hemos intentado que ejercicios sean lo suficientemente realistas como para dar una buena idea de los problemas reales de optimización que pueden encontrarse en la industria, sin abundar demasiado en detalles que van por fuera de la disciplina central que queremos desarrollar en este curso, que es la optimización. El resultado ha sido elegir problemas que de algún modo consideramos emblemáticos y sobre los cuales existen datasets considerados clásicos que permiten abordar su resolución sin tener que pasar tanto tiempo en preparativos.

Este apunte se elaboró en gran medida siguiendo las siguienes referencias:

- Dimitri P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 3rd Edition, 2016.
- Nocedal, Jorge and Wright, Stephen J, *Numerical optimization*, Springer, 1999.

Capítulo 1

Introducción

1.1. Algunos problemas

Tal vez en mayor medida que en otros temas de matemática, la optimización es una disciplina interesante en la medida en la que los problemas que resuelve son interesantes. Es por eso que es conveniente comenzar por justamente por allí, por el origen de algunos problemas de optimización.

Una de las fuentes más grande de problemas de optimización es la estadística. Los estadísticos buscan sistemáticamente minimizar errores, maximizar verosimilitudes y construir modelos que capturen la estructura subyacente de los fenómenos observados. A efectos prácticos, cuando se trata de ajustar los modelos con los datos, emergen los problemas de optimización.

1.1.1. Máxima verosimilitud

Consideremos uno de los métodos estadícos más clásicos: la estimación de máxima verosimilitud. Dentro de este método, los modelos probabilísticos que se propusieron para modelar determinados fenómenos, usualmente dependen de una cantidad finita de parámetros que deben ser estimados a partir de los datos, y para ello se propone maximizar la función de verosimilitud, que es la función que mide la probabilidad de haber observado los datos, en función de los parámetros del modelo.

Veamos un ejemplo. La distribución gamma es una variable aleatoria que permite modelar fenómenos que poseen cierta asimetría positiva. Un buen ejemplo de esto son los volúmenes de lluvia, que tienden a tener valores bajos frecuentes y algunos eventos de lluvia intensa menos comunes. La función de verosimilitud para eventos de lluvia modelados con una distribución gamma se construye como el producto de las probabilidades individuales de cada observación.

Supongamos que tenemos un conjunto de mediciones de precipitación $x_1, x_2, ..., x_n$, y queremos estimar los parámetros de forma (k) y escala (θ) de la distribución gamma. La función de densidad de probabilidad de la distribución gamma para cada observación x_i será:

$$f(x_i; k, \theta) = \frac{1}{\Gamma(k)\theta^k} x_i^{k-1} e^{-x_i/\theta}$$

La función de verosimilitud $L(k, \theta)$ será entonces el producto de estas densidades:

$$L(k,\theta) = \prod_{i=1}^{n} \frac{1}{\Gamma(k)\theta^k} x_i^{k-1} e^{-x_i/\theta}$$

El objetivo de la estimación de máxima verosimilitud será encontrar los valores de k y θ que maximicen esta función, es decir, calcular los estimadores $(\hat{k}, \hat{\theta})$ a partir de la resolución del siguiente problema de optimización:

$$(\hat{k}, \hat{\theta}) = \text{argmax}_{(k,\theta)}\{L(k,\theta)\}$$

donde definimos la notación 'argmax' como los argumentos que realizan el máximo de la función.

Dado que no existe una expresión analítica cerrada para la solución de este problema, el mismo debe abordarse mediante métodos computacionales como los que estudiaremos en esta materia.

1.1.2. Clasificación binaria

El enfoque de parametrizar una función de probabilidad y obtener los parámetros mediante técnicas de optimización es muy poderoso. Comencemos por modelar un problema de regresión más complejo que el caso de nuestra variable de lluvia, como puede ser el problema de la clasificación binaria, que permite abordar situaciones donde la variable de respuesta es dicotómica, como determinar si un estudiante aprueba un examen, si una persona sobrevive al desastre del Titanic o si un paciente padece una enfermedad específica.

Uno de los modelos para resolver la clasificación binaria es la llamada regresión logística. Para formular el modelo de regresión logística desde una perspectiva probabilística, se parte de la suposición de que la variable respuesta $y^{(i)}$ de cada observación sigue una distribución binomial (específicamente, una distribución de Bernoulli en el caso binario). Es decir, para cada i:

$$y^{(i)} \sim \text{Bernoulli}(p^{(i)}),$$

donde $p^{(i)}$ es la probabilidad de que la observación pertenezca a la clase 1. El modelo establece que esta probabilidad se obtiene al aplicar la función sigmoide a una combinación lineal de las variables predictoras:

$$p^{(i)} = \sigma(z^{(i)}) = \frac{1}{1+\exp(-z^{(i)})}, \quad \text{con } z^{(i)} = \mathbf{w}^\top \mathbf{x}^{(i)} + b.$$

En este marco, la función de masa de probabilidad para cada observación es:

$$P(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}, b) = [\sigma(z^{(i)})]^{y^{(i)}} \, [1-\sigma(z^{(i)})]^{1-y^{(i)}}.$$

Para un conjunto de N observaciones, la función de verosimilitud del modelo se obtiene multiplicando las probabilidades individuales:

$$\mathcal{L}(\mathbf{w}, b) = \prod_{i=1}^{N} [\sigma(z^{(i)})]^{y^{(i)}} \, [1-\sigma(z^{(i)})]^{1-y^{(i)}}.$$

Dado que trabajar con productos de probabilidades puede resultar numéricamente inestable, se toma el logaritmo natural de la verosimilitud para simplificar la optimización. Esto conduce a la función de log-verosimilitud:

$$\ell(\mathbf{w}, b) = \sum_{i=1}^{N} \left[y^{(i)} \log(\sigma(z^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(z^{(i)})) \right].$$

El objetivo es encontrar los parámetros $\mathbf{w}$ y b que maximizan esta log-verosimilitud, lo que equivale a ajustar el modelo para que se acomode lo mejor posible a los datos observados. En la práctica, se minimiza la pérdida negativa, conocida como "log loss.º pérdida logarítmica:

$$\mathcal{J}(\mathbf{w}, b) = -\ell(\mathbf{w}, b) = -\sum_{i=1}^{N} \left[y^{(i)} \log(\sigma(z^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(z^{(i)})) \right].$$

Esta formulación probabilística no solo dota al modelo de una interpretación basada en la probabilidad, sino que, lo que nos importa en esta materia, establece un marco claro para el uso de técnicas de optimización.

1.1.3. Clasificación multiclase

Consideremos ahora problemas donde la variable de respuesta puede tomar más de dos valores, como en el caso del dataset Iris (con datos sobre tres especies de flores y mediciones de algunas de sus características morfológicas). Para ello, extenderemos el modelo de regresión logística a la *regresión softmax*.

Figura 1.1: Algunas flores de ejemplo. El dataset Iris contiene mediciones de varias características como longitudes de pétales o sépalos de distintas especies de flores.

En este enfoque, se supone que cada observación pertenece a una de K clases, y la probabilidad de que la observación i pertenezca a la clase j se define mediante:

$$p_j^i(z) = \text{softmax}(z) = \frac{\exp(z_j^{(i)})}{\sum_{k=1}^{K} \exp(z_k^{(i)})}, \quad \text{con } z_j^{(i)} = \mathbf{w}_j^\top \mathbf{x}^{(i)} + b_j.$$

Aquí, cada clase tiene asociado un vector de pesos $\mathbf{w}_j$ y un sesgo b_j. Suponiendo que la variable de respuesta sigue una distribución multinomial, la función de verosimilitud para todo el conjunto de datos se construye a partir de la probabilidad de la clase observada

en cada caso. Al tomar el logaritmo, se obtiene la función de log-verosimilitud, que se minimiza en forma de la pérdida de entropía cruzada:

$$\mathcal{J}(\{\mathbf{w}_j, b_j\}_{j=1}^K) = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_j^{(i)} \log\left(p_j^i\right),$$

donde $y_j^{(i)}$ es la codificación one-hot de la etiqueta de la observación i. Este planteamiento probabilístico, que generaliza el modelo binario, sienta las bases para aplicar técnicas de optimización en la búsqueda de los parámetros que mejor ajusten el modelo a los datos.

1.1.4. Redes Neuronales para Problemas de Clasificación

La clasificación multiclase de la sección anterior tiene una limitación en cuanto a la complejidad de las relaciones funcionales que puede capturar entre los parámetros de entrada y las clases de salida. Para resolver problemas más complejos puede no alcanzar con una combinación lineal y la función softmax.

Consideremos por ejemplo el problema de clasificación de dígitos manuscritos, uno de los problemas más emblemáticos del aprendizaje automático. Un dataset clásico para este problema es conocido como MNIST, que consta de 70,000 imágenes en escala de grises de 28x28 píxeles, cada una etiquetada con un dígito del 0 al 9. Aquí, el objetivo es entrenar una red neuronal para que, dado un dígito manuscrito, prediga correctamente la clase a la que pertenece.

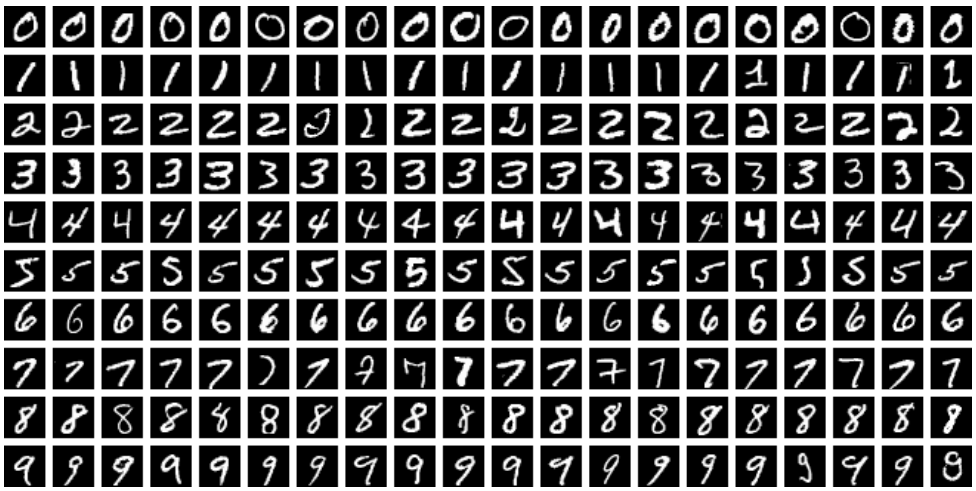

Figura 1.2: Algunas imagenes de ejemplo del dataset MNIST, que consta de 70.000 de estas imágenes, de 28 x 28 pixeles cada una

Para construir una red neuronal en principio mantenemos la solución anterior como una "capa de salida", que utiliza la función softmax para convertir los valores lineales en probabilidades normalizadas. Es decir, el objetivo será transformar los vectores de entrada x que teníamos antes mediante una nueva función que procese directamente los pixels de las imágenes en cuestión y que permita aislar "características" adecuadas, a ser procesadas por esta capa de salida.

La innovación metodológica en el estudio de este problema (por ejemplo, el del reconocimiento de dígitos manuscritos) fue que las características no estaban determinadas

de forma ad-hoc para cada problema mediante algoritmos expertos que por ejemplo identificaban que ciertos dibujos tenían esquinas, eran suaves, tenían un segmento, etc, sino que se ajustaban directamente a partir de una gran cantidad de datos etiquetados.

Para esto, en el modelo de redes neuronales multicapa, se propone reemplazar este paso de extracción de características con una cantidad N de capas ocultas, que aplican, cada una de ellas, una combinación lineal seguida de una función de activación no lineal. Es decir, para cada capa i, se define la transformación:

$$f_i(\mathbf{z}) = \phi^{(i)}\Big(\mathbf{W}^{(i)}\mathbf{z} + \mathbf{b}^{(i)}\Big).$$

y para la capa de salida utilizamos la función softmax como antes:

$$f_{\text{out}}(\mathbf{z}) = \text{softmax}(\mathbf{W}_{\text{out}}\,\mathbf{z} + \mathbf{b}_{\text{out}}).$$

La red completa se puede escribir como la composición de la función de salida con las capas ocultas:

$$NN(x) = \big(f_{\text{out}} \circ f_{L-1} \circ \cdots \circ f_1\big)(\mathbf{x}).$$

La función de pérdida empleada es la misma que utilizamos antes:

$$\mathcal{J} = -\sum_{i=1}^{N}\sum_{j=0}^{9} y_{ij} \log\left(NN(x^{(i)})\right).$$

Hemos arribado a nuestro problema de optimización, que consiste en ajustar los parámetros $W^{(i)}, b^{(i)}, W^{(out)}, b^{(out)}$ para $i = 1, \ldots, N$ para minimizar la funcional J. En el lenguaje de las redes neuronales, al proceso de resolver este problema de optimización (aunque sea de forma aproximada), se lo conoce como "entrenar" la red.

1.2. Visión 3D a partir de imágenes

Consideremos una colección de fotografías tomadas por turistas del Coliseo, con centenares o miles de fotografías de la misma estructura... ¿Podemos usar estos datos (y únicamente estos datos) para reconstruir la estructura 3D del Coliseo? La respuesta nos la da el problema del "bundle adjustment", que es un problema ya clásico de la visión por computadora y fotogrametría, disciplinas con una larga tradición.

La resolución de este problema consta de varias etapas. En la primera de ellas, el objetivo es identificar en diferentes imágenes aquellos puntos que corresponden a la misma posición en el espacio 3D, pese a haber sido captados desde diversas perspectivas.

Es decir, queremos poder obtener valores de x_{ij} las coordenadas observadas del punto j en la imagen i.

Para ello, se utilizan algoritmos como SIFT, SURF o ORB para identificar puntos de interés (o "keypoints") en cada imagen. Estos algoritmos buscan regiones con cambios significativos de intensidad o patrones únicos que puedan distinguirse de su entorno. Posteriormente, se calculan descriptores para cada punto, que son vectores que resumen la información local alrededor del keypoint, y que poseen propiedades como ser invariantes por cambios de escala y rotación. Finalmente, se comparan estos descriptores entre las diferentes imágenes para encontrar correspondencias. Este es un problema de búsqueda que

Figura 1.3: El problema de la reconstrucción de modelos 3D a partir de fotografías 2D es parte del área conocida como structure-from-motion (SfM) dentro de la visión por computadora

se conoce como "búsqueda de vecinos más cercanos", donde para cada descriptor se identifica el descriptor más similar (por ejemplo, minimizando la distancia euclidiana) en otro conjunto. Algoritmos como FLANN abordan este problema de forma eficiente, ofreciendo soluciones aproximadas que permiten emparejar características en grandes volúmenes de datos y en espacios de alta dimensión.

1.2.1. Bundle adjustment

Una vez obtenidos los puntos x_{ij}, es decir, las coordenadas observadas del punto j en la imagen i, podemos continuar con el segundo paso para resolver este problema, que es un problema de optimización no-lineal conocido como el problema de "bundle adjustment".

El objetivo principal del Bundle Adjustment es refinar conjuntamente la estimación de la estructura 3D de una escena y la configuración de las cámaras que la capturaron, mediante la minimización del error de reproyección. En términos simples, se trata de ajustar los parámetros de cada cámara (su posición, orientación y características internas) y las coordenadas 3D de los puntos de la escena, de manera que cuando se proyecten esos puntos en las imágenes, la diferencia con las ubicaciones observadas sea mínima.

Es decir, tenemos que ajustar los parámetros de las cámaras (determinando dónde estaban ubicadas, en qué dirección miraban y cuáles eran sus características internas en el momento de la toma) y las coordenadas de los puntos 3D de los puntos que conforman la estructura del Coliseo.

La función objetivo de Bundle Adjustment se expresa matemáticamente como:

$$\min_{\{\mathbf{P}_i, \mathbf{X}_j\}} \sum_{i=1}^{M} \sum_{j=1}^{N} \left\| \mathbf{x}_{ij} - \text{proj}(\mathbf{P}_i, \mathbf{X}_j) \right\|^2,$$

donde:

1. $\mathbf{P}_i$ representa los parámetros de la cámara i.
2. $\mathbf{X}_j$ es la posición 3D del punto j de la escena.
3. $\mathbf{x}_{ij}$ son las coordenadas observadas del punto j en la imagen i.
4. $\text{proj}(\mathbf{P}_i, \mathbf{X}_j)$ es la función que proyecta el punto 3D $\mathbf{X}_j$ en la imagen i según el modelo de cámara.

Nuevamente, hemos obtenido un problema de optimización no-lineal de grandes dimensiones, que es un caso particular de la familia de problemas conocida como cuadrados mínimos no-lineales.

1.3. Problemas de optimización con restricciones

Un tema importante en la optimización es la presencia de restricciones, que pueden ser lineales, no lineales o incluso definidas a través de ecuaciones diferenciales. Estas restricciones son cruciales, ya que delimitan el conjunto de soluciones factibles y condicionan la viabilidad de las soluciones óptimas.

Una fuente rica de problemas no lineales proviene del ámbito de la economía y las finanzas. Los modelos financieros, por ejemplo, buscan maximizar el rendimiento o minimizar el riesgo, pero la complejidad inherente a los mercados—con sus incertidumbres y dinámicas de comportamiento—hace que tanto la función objetivo como las restricciones sean, en muchos casos, no lineales.

Otra fuente importante de ejemplos es la optimización de redes. En la investigación operativa clásica se estudian problemas lineales como el camino mínimo, el problema de asignación, el flujo máximo o el problema de transporte. Sin embargo, en redes reales es frecuente enfrentar interacciones y dinámicas que generan relaciones no lineales. Por ejemplo, en el flujo de potencia en sistemas eléctricos las relaciones entre tensiones, ángulos de fase y potencias se describen mediante funciones trigonométricas. Estos casos requieren la extensión de los modelos lineales clásicos a formulaciones que capturan mejor la complejidad de las interacciones en la red.

Por último, es fundamental mencionar aquellos problemas en los que las restricciones están definidas por ecuaciones diferenciales. En la teoría de control, por ejemplo, se busca optimizar el desempeño de sistemas dinámicos, como minimizar el consumo de combustible de un cohete que debe maniobrar para alcanzar una órbita determinada.

Consideremos algunos de estos ejemplos en detalle.

1.3.1. Optimización de portafolios

Una fuente de ejemplos de problemas no-lineales son la economía y las finanzas. Por ejemplo, el modelo clásico de optimización de portafolios, conocido como el modelo de Markowitz, se basa en la idea de balancear el rendimiento esperado y el riesgo (medido a través de la varianza) de un portafolio. En su forma más sencilla, se plantea como el siguiente problema cuadrático:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \mathbf{w}^\top \Sigma \, \mathbf{w} \\ \text{sujeto a} \quad & \mathbf{w}^\top \mu \geq R, \\ & \mathbf{w}^\top \mathbf{1} = 1, \\ & w_i \geq 0, \quad \forall i, \end{aligned}$$

donde:

1. $\mathbf{w} = (w_1, w_2, \ldots, w_N)$ son los pesos asignados a cada uno de los N activos.
2. Σ es la matriz de covarianza de los retornos de los activos.
3. μ es el vector de rendimientos esperados.

4. R es el rendimiento mínimo deseado.

5. La restricción $\mathbf{w}^\top \mathbf{1} = 1$ asegura que la totalidad del capital se invierta.

Por supuesto, existen muchas variantes que pueden volver al problema más complejo, como la utilización de otras medidas de riesgo, etc.

1.3.2. Flujo de Potencia en Sistemas Eléctricos

En redes eléctricas, el objetivo es determinar el estado operativo óptimo (tensiones, ángulos y potencias) que minimice el costo de generación o las pérdidas, sujeto a las ecuaciones no lineales que describen el flujo de potencia según las leyes de Kirchhoff. La formulación típica es:

$$
\begin{aligned}
\min_{V,\theta,P_G,Q_G} \quad & \sum_{i\in\mathcal{G}} C_i(P_{Gi}) \\
\text{sujeto a} \quad & P_i - P_{Di} = \sum_{j\in\mathcal{N}(i)} V_i V_j \left(G_{ij}\cos(\theta_i-\theta_j) + B_{ij}\sin(\theta_i-\theta_j)\right), \quad \forall i, \\
& Q_i - Q_{Di} = \sum_{j\in\mathcal{N}(i)} V_i V_j \left(G_{ij}\sin(\theta_i-\theta_j) - B_{ij}\cos(\theta_i-\theta_j)\right), \quad \forall i, \\
& V_i^{\text{mín}} \le V_i \le V_i^{\text{máx}}, \quad \forall i, \\
& \text{otras restricciones operativas (límites de generadores, líneas, etc.).}
\end{aligned}
$$

Aquí, V_i y θ_i son la magnitud y el ángulo de la tensión en el bus i, P_{Gi} y Q_{Gi} son las potencias generadas, mientras que P_{Di} y Q_{Di} representan las demandas. G_{ij} y B_{ij} son, respectivamente, las conductancias y susceptancias de la línea que conecta los buses i y j.

1.3.3. Optimización en Redes de Comunicaciones Inalámbricas

En el ámbito de las redes inalámbricas, uno de los objetivos es maximizar la capacidad total o la tasa de transmisión, ajustando la potencia de transmisión de cada nodo y gestionando la interferencia entre ellos. Un modelo típico se basa en la fórmula de Shannon y se puede plantear de la siguiente manera:

$$
\begin{aligned}
\max_{\mathbf{p}} \quad & \sum_{i=1}^{N} \log\left(1 + \frac{h_{ii}p_i}{\sigma_i^2 + \sum_{j\neq i} h_{ij}p_j}\right) \\
\text{sujeto a} \quad & 0 \le p_i \le P_i^{\text{máx}}, \quad \forall i,
\end{aligned}
$$

donde:

1. p_i es la potencia de transmisión asignada al nodo i,

2. h_{ij} representa el canal o ganancia de la señal del transmisor j al receptor i,

3. σ_i^2 es la potencia de ruido en el receptor i,

4. $P_i^{\text{máx}}$ es la potencia máxima disponible para el nodo i.

Esta formulación es intrínsecamente no lineal y, en general, no convexa, lo que requiere el uso de métodos avanzados o aproximaciones para encontrar soluciones que maximicen la eficiencia espectral de la red.

1.3.4. Optimización de Redes de Distribución de Agua

En sistemas de distribución de agua, el objetivo es garantizar un suministro adecuado manteniendo niveles de presión y caudales óptimos en toda la red, mientras se minimizan los costos operativos o se maximizan criterios de eficiencia. Este problema involucra restricciones de conservación de caudal y ecuaciones no lineales que describen las relaciones entre diferencias de presión y caudales en cada tubería, derivadas de la hidráulica y la dinámica de fluidos.

La formulación típica es:

$$
\begin{aligned}
\min_{q,H} \quad & \sum_{i\in\mathcal{N}} \left(H_i - H_i^{\text{ref}}\right)^2 \quad \text{(o bien minimizar el costo energético de bombeo)} \\
\text{sujeto a} \quad & \sum_{j:(i,j)\in\mathcal{A}} q_{ij} - \sum_{j:(j,i)\in\mathcal{A}} q_{ji} = d_i, \quad \forall i \in \mathcal{N}, \\
& H_i - H_j = r_{ij}\,|q_{ij}|^{\gamma}, \quad \forall (i,j) \in \mathcal{A}, \\
& q_{ij}^{\text{mín}} \le q_{ij} \le q_{ij}^{\text{máx}}, \quad \forall (i,j) \in \mathcal{A}, \\
& H_i^{\text{mín}} \le H_i \le H_i^{\text{máx}}, \quad \forall i \in \mathcal{N}.
\end{aligned}
$$

Aquí:

1. H_i es la cabeza hidráulica (presión) en el nodo i y H_i^{ref} es el valor de presión deseado.
2. q_{ij} representa el caudal en la tubería que conecta los nodos i y j.
3. d_i es la demanda (o aporte, si es negativa) en el nodo i.
4. r_{ij} y γ son parámetros que capturan las características físicas de la tubería (por ejemplo, según el modelo de Hazen-Williams).
5. $\mathcal{N}$ es el conjunto de nodos y $\mathcal{A}$ el conjunto de tuberías.

Esta formulación es intrínsecamente no lineal debido a la ecuación de pérdida de carga $H_i - H_j = r_{ij}\,|q_{ij}|^{\gamma}$, y requiere el uso de métodos de optimización no lineales para encontrar soluciones factibles que cumplan con todas las restricciones operativas y de seguridad.

1.4. Problemas con restricciones dadas por ODEs

Los problemas de la sección anterior eran sistemas en condiciones de operación estáticas, es decir que no cambian con el tiempo. En muchas áreas de la ingeniería se suele lidiar con sistemas dinámicos, cuya evolución temporal se suele modelar con ecuaciones diferenciales. Veamos algunos problemas de ejemplo

1.4.1. Transferencia de Órbita de Cohetes

En mecánica orbital, la transferencia de órbita consiste en diseñar la trayectoria óptima que permita a un cohete pasar de una órbita inicial a una órbita final, minimizando el consumo de combustible (o el delta-V) y respetando las dinámicas del sistema. Este problema se formula como un problema de control óptimo, donde la señal de control $u(t)$

(por ejemplo, la aceleración aplicada) influye en la evolución del estado del cohete, que incluye posición y velocidad.

La formulación matemática típica es:

$$\begin{aligned} \min_{u(t)} \quad & \int_{t_0}^{t_f} \|u(t)\|\, dt \quad \text{(minimizar el consumo de combustible)} \\ \text{sujeto a} \quad & \dot{\mathbf{x}}(t) = f\big(\mathbf{x}(t), u(t)\big), \quad t \in [t_0, t_f], \\ & \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_f) = \mathbf{x}_f, \\ & u(t) \in \mathcal{U}, \quad \forall t \in [t_0, t_f], \end{aligned}$$

donde:

1. $\mathbf{x}(t)$ es el vector de estado (por ejemplo, posición y velocidad),
2. $f(\mathbf{x}, u)$ describe la dinámica del sistema (basada en las leyes de Newton y la gravitación),
3. $\mathcal{U}$ es el conjunto de controles admisibles.

La solución de este problema se obtiene mediante técnicas de control óptimo, como el principio del máximo de Pontryagin o métodos numéricos basados en la discretización del tiempo.

1.4.2. Optimización de Administración de Fármacos

El diseño de esquemas de administración de fármacos busca determinar el protocolo óptimo de dosificación que maximice la eficacia terapéutica y minimice los efectos secundarios, considerando la dinámica no lineal de absorción, distribución, metabolización y excreción del fármaco. Este problema se aborda como un problema de control óptimo en el que la señal de control $u(t)$ representa la dosis administrada a lo largo del tiempo.

La formulación matemática puede expresarse de la siguiente manera:

$$\begin{aligned} \min_{u(t)} \quad & \int_0^T L\big(x(t), u(t)\big)\, dt \\ \text{sujeto a} \quad & \dot{x}(t) = f\big(x(t), u(t)\big), \quad t \in [0, T], \\ & x(0) = x_0, \quad x(T) \text{ deseado (o restricciones en } x(t)), \\ & u(t) \in \mathcal{U}, \quad \forall t \in [0, T], \end{aligned}$$

donde:

1. $x(t)$ representa la concentración del fármaco (o estados relacionados) en el organismo,
2. $f(x, u)$ es el modelo farmacocinético/farmacodinámico (PK/PD) que describe la evolución de $x(t)$ en función de $u(t)$,
3. $L(x, u)$ es la función de costo que penaliza tanto concentraciones excesivamente altas (para minimizar efectos secundarios) como concentraciones demasiado bajas (para garantizar la eficacia),
4. $\mathcal{U}$ define los límites y restricciones en la dosificación.

Esta formulación permite encontrar el régimen de dosificación óptimo, utilizando técnicas numéricas de control óptimo que pueden incluir métodos basados en gradientes o algoritmos de optimización directa.

1.4.3. Optimización de Procesos Químicos

En muchos procesos químicos es crucial optimizar la operación dinámica de reactores, asegurando que las concentraciones de reactivos y productos evolucionen de manera óptima en el tiempo. Por ejemplo, en un reactor continuo de tanque agitado (CSTR) para una reacción de primer orden, el objetivo puede ser maximizar la conversión o minimizar el tiempo de residencia, optimizando condiciones como el caudal de alimentación o la temperatura.

Una formulación típica es:

$$\begin{aligned} \min_{u(t)} \quad & \int_0^T L\big(C_A(t), u(t)\big)\, dt \\ \text{sujeto a} \quad & \frac{dC_A}{dt} = \frac{F}{V}\big(C_{A0} - C_A(t)\big) - k\, C_A(t), \quad t \in [0, T], \\ & C_A(0) = C_{A0}, \\ & u(t) \in \mathcal{U}. \end{aligned}$$

Aquí:

1. $C_A(t)$ es la concentración del reactivo A en el reactor,
2. C_{A0} es la concentración de A en la corriente de entrada,
3. F es el caudal volumétrico, V el volumen del reactor y k la constante de velocidad,
4. $u(t)$ representa la variable de control (por ejemplo, la temperatura o el caudal ajustable),
5. $L(C_A(t), u(t))$ es la función de costo que penaliza condiciones indeseadas (como bajas conversiones o altos costos energéticos).

Esta formulación integra ecuaciones diferenciales para capturar la dinámica del proceso y se resuelve mediante técnicas de control óptimo, siguiendo un enfoque similar a otros problemas dinámicos en ingeniería.

1.5. Problemas con restricciones dadas por PDEs

Más allá de las ecuaciones ordinarias, que modelan la evolución temporal de un sistema puntual, muchos problemas en ciencia e ingeniería se modelan mediante sistemas distribuidos, es decir, donde la dependencia espacial es un factor no-trivial. Sobre estos sistemas también pueden plantearse problemas de optimización.

Capítulo 2

Condiciones de optimalidad

2.1. Definiciones

En la optimización sin restricciones se minimiza una función objetivo que depende de varias variables reales, sin ninguna restricción sobre los valores de estas. A este problema, lo denotaremos

$$\min_x f(x),$$

donde $x \in \mathbb{R}^n$ es un vector real con $n \geq 1$ componentes y $f : \mathbb{R}^n \to \mathbb{R}$ es una función suave.

Generalmente se desea encontrar un minimizador global de f. Una definición formal es la siguiente:

Definición 1. *Un punto x^* es un minimizador global si $f(x^*) \leq f(x)$ para todo x, donde x recorre todo $\mathbb{R}^n$ (o al menos el dominio de interés).*

El minimizador global puede ser difícil de encontrar, ya que nuestro conocimiento de f suele ser únicamente local. Es más, dado que nuestro algoritmo no visita muchos puntos (o al menos eso esperamos), generalmente no tenemos una buena imagen de la forma global de f, y nunca podemos estar seguros de que la función no presente una caída abrupta en alguna región que no ha sido explorada por el algoritmo. Por este motivo, la mayoría de los algoritmos son capaces de encontrar solo un minimizador local, que es un punto que alcanza el valor más bajo de f en un entorno. Formalmente, decimos:

Definición 2. *Un punto x^* es un minimizador local si existe un entorno N de x^* tal que $f(x^*) \leq f(x)$ para todo $x \in N$.*

Podemos distinguir los minimizadores donde vale una condición más fuerte (la desigualdad estricta) con una nueva definición:

Definición 3. *Un punto x^* es un minimizador local estricto (también llamado minimizador local fuerte) si existe un entorno N de x^* tal que*

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

para todo $x \in N$ con $x \neq x^*$.

Un tipo algo más exótico de minimizador local se define de la siguiente manera:

Definición 4. *Un punto x^* es un minimizador local aislado si existe un entorno N de x^* tal que x^* es el único minimizador local en N.*

Algunos minimizadores locales estrictos no son aislados, como lo ilustra la función

$$f(x) = x^4 \cos\left(\frac{1}{x}\right) + 2x^4, \quad f(0) = 0,$$

la cual es dos veces continuamente diferenciable y tiene un minimizador local estricto en $x^* = 0$. Sin embargo, existen minimizadores locales estrictos en muchos puntos cercanos, que podemos etiquetar como x_j, de modo que $x_j \to 0$ cuando $j \to \infty$. (Ejercicio)

Si bien los minimizadores locales estrictos no siempre son aislados, es cierto que todos los minimizadores locales aislados son estrictos.

A partir de las definiciones anteriores, podría parecer que la única forma de determinar si un punto x^* es un mínimo local es examinar todos los puntos en su entorno para asegurarse de que ninguno de ellos tiene un valor de función menor. Sin embargo, cuando la función f es suave, existen métodos más eficientes y prácticos para identificar mínimos locales. En particular, si f es dos veces continuamente diferenciable, es posible determinar que x^* es un minimizador local (y posiblemente un minimizador local estricto) examinando únicamente el gradiente $\nabla f(x^*)$ y el Hessiano $\nabla^2 f(x^*)$.

La herramienta matemática utilizada para estudiar los minimizadores de funciones suaves es el teorema de Taylor. Dado que este teorema es central en nuestro análisis, lo enunciamos a continuación. Su demostración se puede encontrar en cualquier libro de cálculo.

Teorema 1 (Teorema de Taylor)**.** *Supongamos que $f : \mathbb{R}^n \to \mathbb{R}$ es continuamente diferenciable y que $p \in \mathbb{R}^n$. Entonces se tiene que*

$$f(x+p) = f(x) + \nabla f(x+tp)^T p, \text{paraalgun} t \in (0,1).$$

Además, si f es dos veces continuamente diferenciable, se tiene que

$$\nabla f(x+p) = \nabla f(x) + \int_0^1 \nabla^2 f(x+tp)\, p\, dt,$$

y que

$$f(x+\alpha d) = f(x) + \alpha \nabla f(x)^T d + \frac{\alpha^2}{2} d^T \nabla^2 f(x) d + o(\alpha^2),$$

Las condiciones necesarias para la optimalidad se derivan asumiendo que x^* es un minimizador local y luego demostrando propiedades acerca de $\nabla f(x^*)$ y $\nabla^2 f(x^*)$.

2.2. Condiciones necesarias o suficientes

Teorema 2 (Condiciones Necesarias de Optimalidad)**.** *Sea x^* un mínimo local sin restricciones de $f : \mathbb{R}^n \to \mathbb{R}$, y supongamos que f es continuamente diferenciable en*

un conjunto abierto S que contiene a x^*. Entonces

$$\nabla f(x^*) = 0 \quad \textit{(Condición Necesaria de Primer Orden).}$$

Si además f es dos veces continuamente diferenciable en S, entonces

$$\nabla^2 f(x^*) \textit{ es semidefinida positiva.} \quad \textit{(Condición Necesaria de Segundo Orden).}$$

Demostración. Fijemos un $d \in \mathbb{R}^n$. Entonces, utilizando la regla de la cadena para derivar la función $g(\alpha) = f(x^* + \alpha d)$ con respecto al escalar α, tenemos

$$0 \leq \lim_{\alpha \to 0} \frac{f(x^* + \alpha d) - f(x^*)}{\alpha} = \frac{dg(0)}{d\alpha} = d' \nabla f(x^*),$$

donde la desigualdad se sigue de la suposición de que x^* es un mínimo local. Como d es arbitrario, la misma desigualdad se mantiene al reemplazar d por $-d$. Por lo tanto, $d' \nabla f(x^*) = 0$ para todo $d \in \mathbb{R}^n$, lo que demuestra que $\nabla f(x^*) = 0$.

Supongamos ahora que f es dos veces continuamente diferenciable, y sea d un vector arbitrario en $\mathbb{R}^n$. Para todo $\alpha \in \mathbb{R}$, la expansión de segundo orden da

$$f(x^* + \alpha d) - f(x^*) = \alpha \nabla f(x^*)' d + \frac{\alpha^2}{2} d' \nabla^2 f(x^*) d + o(\alpha^2).$$

Utilizando la condición $\nabla f(x^*) = 0$ y la optimalidad local de x^*, vemos que existe un $\varepsilon > 0$ suficientemente pequeño tal que para todo α con $\alpha \in (0, \varepsilon)$,

$$0 \leq \frac{f(x^* + \alpha d) - f(x^*)}{\alpha^2} = \frac{1}{2} d' \nabla^2 f(x^*) d + \frac{o(\alpha^2)}{\alpha^2}.$$

Tomando el límite cuando $\alpha \to 0$ y usando el hecho de que $\lim_{\alpha \to 0} \frac{o(\alpha^2)}{\alpha^2} = 0$, obtenemos que $d' \nabla^2 f(x^*) d \geq 0$, lo que muestra que $\nabla^2 f(x^*)$ es semidefinida positiva. □

Ahora describimos condiciones suficientes, que son condiciones sobre las derivadas de f en el punto x^* que garantizan que x^* es un mínimo local.

Teorema 3 (Condiciones Suficientes de Optimalidad de Segundo Orden)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ dos veces continuamente diferenciable sobre un conjunto abierto S. Supongamos que un vector $x^* \in S$ satisface las condiciones*

$$\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) : \textit{definida positiva.}$$

Entonces, x^* es un mínimo local estricto sin restricciones de f. En particular, existen escalares $\gamma > 0$ y $\varepsilon > 0$ tales que

$$f(x) \geq f(x^*) + \frac{\gamma}{2} \|x - x^*\|^2, \quad \forall \ \ x \textit{ con } \|x - x^*\| < \varepsilon. \tag{2.1}$$

Demostración. Denotemos por λ el menor autovalor de $\nabla^2 f(x^*)$. Por el Teorema 11 del Apéndice A, λ es positivo, ya que $\nabla^2 f(x^*)$ es definida positiva. Además, por el Teorema 10 del Apéndice A,

$$d \cdot \nabla^2 f(x^*) d \geq \lambda \|d\|^2, \quad \forall \ \ d \in \mathbb{R}^n.$$

Utilizando esta relación, la hipótesis $\nabla f(x^*) = 0$ y una expansión de segundo orden, tenemos para todo d

$$\begin{aligned} f(x^* + d) - f(x^*) &= \nabla f(x^*) \cdot d + \frac{1}{2} d \cdot \nabla^2 f(x^*) d + o(\|d\|^2) \\ &\geq \frac{\lambda}{2}\|d\|^2 + o(\|d\|^2) \\ &= \left(\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2}\right)\|d\|^2. \end{aligned}$$

Elijamos cualquier $\varepsilon > 0$ y $\gamma > 0$ tales que

$$\frac{\lambda}{2} + \frac{o(\|d\|^2)}{\|d\|^2} \geq \frac{\gamma}{2}, \quad \forall d \text{ con } \|d\| < \varepsilon.$$

Entonces, se satisface la Ecuación (2.1) □

2.3. Funciones convexas

Teorema 4. *Sea f convexa, entonces cualquier minimizador local x^* es un minimizador global de f. Además, si f es diferenciable, entonces cualquier punto estacionario x^* es un minimizador global de f.*

Demostración. Supongamos, para llegar a una contradicción, que x^* es un minimizador local pero no global. Entonces, existe algún punto $z \in \mathbb{R}^n$ tal que

$$f(z) < f(x^*).$$

Consideremos el segmento de recta que une x^* con z; es decir, definamos

$$x = \lambda z + (1 - \lambda)x^*, \quad \text{para algún } \lambda \in (0, 1]. \tag{2.7}$$

Por la propiedad de convexidad de f, se tiene

$$f(x) \leq \lambda f(z) + (1 - \lambda) f(x^*) < f(x^*). \tag{2.8}$$

Cualquier entorno N de x^* contiene un tramo del segmento (2.7), por lo que siempre habrá puntos en N en los que el valor de la función es menor que $f(x^*)$. Esto contradice la hipótesis de que x^* es un minimizador local.

Para la segunda parte del teorema, supongamos que x^* no es un mínimo global y elijamos z como se hizo anteriormente. Entonces, a partir de la convexidad, se tiene

$$\begin{aligned} \frac{d}{d\lambda} f\Big(x^* + \lambda(z - x^*)\Big)\Big|_{\lambda=0} &= \lim_{\lambda \downarrow 0} \frac{f(x^* + \lambda(z - x^*)) - f(x^*)}{\lambda} && (2.2) \\ &\leq \lim_{\lambda \downarrow 0} \frac{\lambda f(z) + (1 - \lambda) f(x^*) - f(x^*)}{\lambda} && (2.3) \\ &= \lim_{\lambda \downarrow 0} \frac{f(z) - f(x^*)}{\lambda} < 0. && (2.4) \end{aligned}$$

Por lo tanto,

$$\nabla f(x^*)^T (z - x^*) < 0.$$

De ello se deduce que $\nabla f(x^*) \neq 0$, y por lo tanto x^* no es un punto estacionario. □

Capítulo 3

Métodos de gradiente

Consideramos métodos iterativos de la forma

$$x_{k+1} = x_k + \alpha_k d_k \tag{3.1}$$

donde $d_k \in \mathbb{R}^n$ es una dirección, y α_k un tamaño de paso.

$$f(x_{k+1}) = f(x_k) + \alpha_k \nabla f(x_k) \cdot d_k + o(\alpha_k) \tag{3.2}$$

3.1. Elección de la dirección

Consideremos ahora varios métodos para elegir la dirección de descenso d_k.

Definición 5. *Llamaremos a $d_k = -\nabla f_k$ la dirección de descenso más rápido.*

La dirección de descenso más rápido $-\nabla f_k$ es la elección más obvia para la dirección de búsqueda. Es intuitiva; entre todas las direcciones en las que podríamos movernos desde x_k, es aquella a lo largo de la cual f decrece más rápidamente (de manera local). Para verificar esta afirmación, recurrimos nuevamente al teorema de Taylor, el cual nos dice que para cualquier dirección de búsqueda p y un parámetro de longitud de paso α, tenemos

$$f(x_k + \alpha p) = f(x_k) + \alpha p^T \nabla f_k + \frac{1}{2}\alpha^2 p^T \nabla^2 f(x_k) p, \tag{3.3}$$

para algún $t \in (0, \alpha)$ (ver (2.6)). La tasa de cambio de f a lo largo de la dirección p en x_k es simplemente el coeficiente de α, es decir, $p^T \nabla f_k$. Por lo tanto, la dirección unitaria p de descenso más rápido es la solución del problema

$$\text{mín } p^T \nabla f_k, \quad \text{sujeto a } \|p\| = 1.$$

Dado que $p^T \nabla f_k = \|p\| \|\nabla f_k\| \cos\theta = \|\nabla f_k\| \cos\theta$, donde θ es el ángulo entre p y ∇f_k, es fácil ver que el mínimo se alcanza cuando $\cos\theta = -1$ y

$$p = -\frac{\nabla f_k}{\|\nabla f_k\|},$$

como se afirmaba.

Esta dirección es ortogonal a las curvas de nivel de la función. El hecho se demuestra fácilmente. Supongamos que tenemos una función diferenciable $f(x, y)$ y una curva de

nivel definida por $f(x, y) = c$ (con c constante). Si parametrizamos la curva de nivel mediante $\mathbf{r}(t) = (x(t), y(t))$ de modo que $f(x(t), y(t)) = c$ para todo t, al derivar respecto a t se obtiene:

$$\frac{d}{dt} f(x(t), y(t)) = \nabla f(x(t), y(t)) \cdot \mathbf{r}'(t) = 0.$$

La ecuación anterior significa que el producto escalar entre el gradiente ∇f y el vector tangente $\mathbf{r}'(t)$ a la curva es cero. Esto implica que ∇f es ortogonal al vector tangente a la curva de nivel en cada punto, es decir, el gradiente es perpendicular a las curvas de nivel.

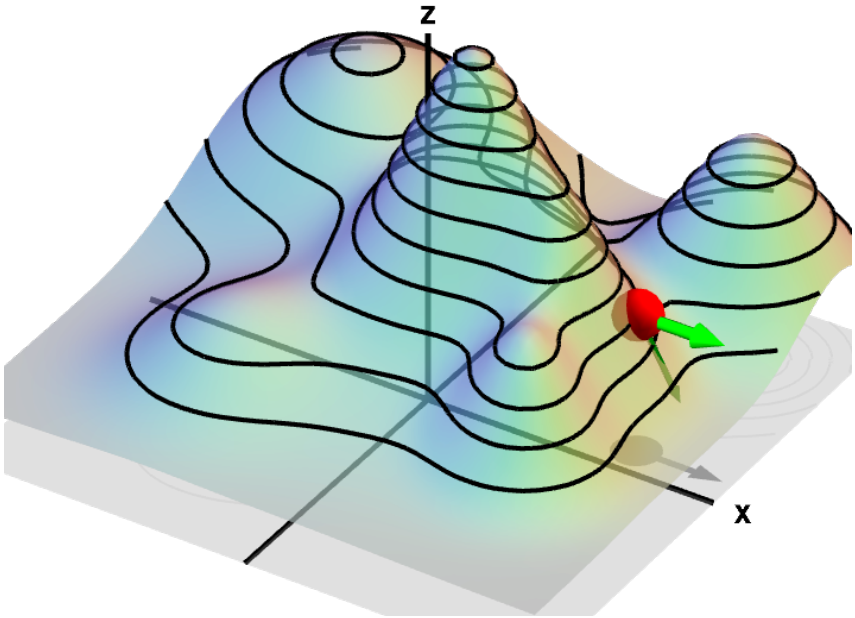

Figura 3.1: La dirección del gradiente es ortogonal a las curvas de nivel de una función.

El método de descenso más rápido es un método de descenso en el que en cada paso se mueve en la dirección $d_k = -\nabla f_k$. La longitud de paso α_k puede elegirse de diversas maneras, como se discutirá a continuación. Una ventaja de la dirección de descenso más rápido es que solo requiere calcular el gradiente ∇f_k, pero no las segundas derivadas. Sin embargo, puede ser extremadamente lento en problemas difíciles.

Los métodos de gradiente pueden usar direcciones de búsqueda distintas a la dirección de descenso más rápido. En general, cualquier dirección de descenso—una que forme un ángulo estrictamente menor que $\pi/2$ radianes con $-\nabla f_k$—garantiza una disminución de f, siempre que la longitud de paso es suficientemente pequeña. Esto motiva la siguiente definición:

Definición 6. *Decimos que* $\{d_k \geq 0\}$ *es una dirección* ***de descenso o relacionada con el gradiente*** *si*

$$\nabla f(x_k) \cdot d_k < 0 \quad \forall k \tag{3.4}$$

Lema 1. *En los métodos de gradiente, cualquier dirección de descenso* (3.4) *garantiza que* $f(x_k + \alpha d_k) < f(x_k)$ *para valores positivos pero suficientemente pequeños de* α.

Demostración. Podemos verificar esta afirmación utilizando el teorema de Taylor. A partir de (3.3), tenemos que

$$f(x_k + \alpha d_k) = f(x_k) + \alpha d_k^T \nabla f_k + O(\alpha^2).$$

Cuando d_k es una dirección de descenso, el ángulo θ_k entre d_k y ∇f_k satisface $\cos\theta_k < 0$, por lo que

$$d_k^T \nabla f_k = \|d_k\| \|\nabla f_k\| \cos\theta_k < 0.$$

De esto se deduce que $f(x_k + \alpha d_k) < f(x_k)$ para valores positivos pero suficientemente pequeños de α. □

Otra dirección de búsqueda importante es la dirección de Newton. Esta dirección se deriva de la aproximación de segundo orden de la serie de Taylor para $f(x_k + p)$, que es

$$f(x_k + p) \approx f_k + p^T \nabla f_k + \frac{1}{2} p^T \nabla^2 f_k p = m_k(p).$$

Asumiendo por el momento que $\nabla^2 f_k$ es definida positiva, obtenemos la dirección de Newton al encontrar el vector p que minimiza $m_k(p)$. Simplemente igualamos a cero el gradiente de $m_k(p)$, para lo que recordamos que, dada

$$f(p) = \frac{1}{2} p^T \nabla^2 f_k p$$

tenemos

$$\nabla f(p) = p^T \nabla^2 f_k$$

por lo que obtenemos

$$\nabla m_k(p) = \nabla f_k + p^T \nabla^2 f_k$$

y entonces, igualando a cero y despejando, obtenemos

$$d_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k.$$

Definición 7. *La dirección del método de Newton puro está dada por* $d_k^N = -(\nabla^2 f_k)^{-1} \nabla f_k$.

La dirección de Newton es confiable cuando la diferencia entre la función real $f(x_k+p)$ y su modelo cuadrático $m_k(p)$ no es demasiado grande. Haciendo una expansión de Taylor de orden superior, vemos que la única diferencia entre la función y su aproximación cuadrática es que la matriz $\nabla^2 f(x_k + tp)$ ha sido reemplazada por $\nabla^2 f_k$. Si $\nabla^2 f$ es suficientemente suave, esta diferencia introduce una perturbación de solo $O(\|p\|^3)$ en la expansión, por lo que cuando $\|p\|$ es pequeña, la aproximación $f(x_k + p) \approx m_k(p)$ es bastante precisa.

Lema 2. *Suponiendo que* $\nabla^2 f_k$ *es definida positiva, la dirección de Newton pura es una dirección de descenso o relacionada con el gradiente, es decir, satisface* (3.4).

Demostración. Para demostrar esto, observemos que la dirección de Newton pura, satisface

$$d_k = -H_k^{-1} \nabla f_k$$

Multiplicando a ambos lados por ∇f_k^t,

$$\nabla f_k^t d_k = -\nabla f_k^t H_k^{-1} \nabla f_k$$

Y como H_k^{-1} es definida-positiva, siempre que $\nabla f_k \neq 0$, tendremos $\nabla f_k^T H_k^{-1} \nabla f_k > 0$. Por lo tanto,

$$\nabla f_k^t d_k < 0$$

□

A diferencia de la dirección de descenso más rápido, la dirección de Newton tiene una longitud de paso "natural" de 1. La mayoría de las implementaciones de búsqueda en línea del método de Newton utilizan el paso unitario $\alpha = 1$ siempre que sea posible y ajustan α solo cuando este no produce una reducción satisfactoria en el valor de f.

Cuando $\nabla^2 f$ no es definida positiva, la dirección de Newton puede no estar definida, ya que $(\nabla^2 f_k)^{-1}$ podría no existir. Incluso cuando está definida, puede no satisfacer la propiedad de descenso, en cuyo caso $\nabla f_k^T d_k^N$ no sería negativo, volviéndola inadecuada como dirección de búsqueda. En estas situaciones, los métodos de búsqueda en línea modifican la definición de d_k para garantizar que satisfaga la condición de descenso, mientras retienen la ventaja de la información de segundo orden contenida en $\nabla^2 f_k$. Estas modificaciones las describiremos más adelante.

Los métodos que utilizan la dirección de Newton tienen una tasa rápida de convergencia local, típicamente cuadrática. Una vez que se alcanza un entorno de la solución, la convergencia con alta precisión suele lograrse en solo unas pocas iteraciones. La principal desventaja de la dirección de Newton es la necesidad de calcular el Hessiano $\nabla^2 f(x)$. El cálculo explícito de esta matriz de derivadas segundas puede ser un proceso complejo, propenso a errores y costoso. Las técnicas de diferenciación automática y diferencias finitas, descritas más adelante, pueden ser útiles para evitar la necesidad de calcular las segundas derivadas manualmente.

Una alternativa atractiva al método de Newton nos la proporcionan las direcciones de búsqueda de quasi-Newton, que no requieren el cálculo del Hessiano y aun así logran una tasa de convergencia superlineal. En lugar del verdadero Hessiano $\nabla^2 f_k$, se puede utilizar una aproximación B_k, que se actualiza después de cada paso para incorporar el conocimiento adicional adquirido durante la iteración. Las actualizaciones aprovechan el hecho de que los cambios en el gradiente ∇f proporcionan información sobre la segunda derivada de f a lo largo de la dirección de búsqueda.

3.2. Elección del tamaño del paso

Definición 8. *La regla de Minimización selecciona α_k mediante la resolución de un problema de minimización unidimensional. En detalle, se toma α_k tal que*

$$f(x_k + \alpha_k d_k) = \min_{\alpha \geq 0} f(x_k + \alpha_k d_k)$$

Definición 9. *La regla de Minimización limitada añade una restricción a la regla de minimización, definiendo un valor máximo s, y selecciona α_k mediante*

$$f(x_k + \alpha_k d_k) = \min_{\alpha \in [0,s]} f(x_k + \alpha_k d_k)$$

Las reglas de minimización pueden incurrir en elevados costos computacionales. Para mitigarlos, una manera es considerar un valor inicial s, y si este no permite conseguir una reducción en el valor de la función objetivo (y por ejemplo, lo empeora), entonces se procede a reducir el valor de s, multiplicando por un factor fijo. Si bien esto suele funcionar en la práctica, no es teóricamente sólido, ya que existen contra-ejemplos que muestran que la solución puede no converger.

Para resolver este inconveniente, se puede introducir la regla de Armijo, que es una versión rigurosa de esta misma idea.

Definición 10. *La regla de Armijo, dados valores de s, $0 < \sigma, \beta < 1$, selecciona $\alpha_k^m = \beta^m s$ mediante el test*

$$f(x_k) - f(x_k + \alpha_k^m d_k) \geq -\sigma \alpha_k^m \nabla f(x_k) \cdot d_k,$$

donde m es el primer entero que satisface el test.

Definición 11. *La regla del paso constante, dado un valor de s, elije $\alpha_k = s$.*

La regla constante es muy simple, pero si el valor elegido es muy grande el método diverge, y si es muy pequeña, la convergencia puede ser muy lenta. Entonces la regla del paso constante es útil para problemas donde podemos determinar este valor fácilmente.

Por ejemplo, en el caso donde f es una función convexa, hay métodos que permiten determinar el valor de s de manera automática. (Ver ejercicios). En este método, el valor del paso se reduce hasta que eventualmente permanece constante. Sin embargo, la convergencia de un método de descenso con paso constante require que el gradiente ∇f satisfaga una condición de Lipschitz. Por el contrario, las reglas de minimización o Armijo no tienen esta restricción.

Definición 12. *La regla del paso descendiente considera sucesiones tales que $\alpha_k \to 0$, de modo que $\sum_k a_k = \infty$. Una condición relacionada (Robins-Monro) requiere $\sum_k a_k = \infty$ y $\sum_k a_k^2 < \infty$.*

La restricción de que la suma de los pasos diverja se toma para que la regla no converga artificialmente a un punto no-estacionario, o, equivalentemente, que tenga la capacidad de alcanzar cualquier punto. Esta regla tiene buenas garantías de convergencia. Sin embargo la tasa de convergencia suele ser lenta, con lo cual, esta regla se utiliza en casos donde la convergencia lenta es inevitable, como son algunos problemas singulares o cuando los gradientes se calculan con errores considerables, como es el caso cuando existe ruido aleatorio en la evaluación de los gradientes.

3.3. Teoría de convergencia

Los principales teoremas de convergencia que veremos a continuación son de la forma "todo punto de acumulación es estacionario". Estos teoremas no garantizan de por sí la convergencia de los métodos de descenso, pero, en conjunto con hipótesis extras sobre las funciones objetivo, sí permiten obtener resultados de convergencia global (convergencia a un mínimo local, pero a partir de cualquier punto inicial x_0).

Más adelante veremos también un resultado de convegencia local, es decir, que dado $x_0 \in I$ para cierto dominio I que contiene un mínimo, podremos garantizar la convergencia a dicho mínimo.

La primera hipótesis que podemos dar para garantizar convergencia para todo valor inicial x_0 está relacionada con el Teorema de Weierstrass.

Lema 3. *Si la función objetivo f es continua y tiene un conjunto de nivel acotado para algún x_0, es decir, el conjunto $\{x \mid f(x) \leq f(x_0)\}$ es acotado, y si el algoritmo es un método de descenso (es decir, $f(x_{k+1}) \leq f(x_k)$ para todo k), entonces la secuencia $\{x_k\}$ generada por el algoritmo estará acotada y por lo tanto existe una subsucesión $\{x_{kj}\}$ convergente, es decir $x_{kj} \to x^*$.*

Demostración. Esto se debe a que todos los iterandos permanecerán dentro del conjunto de nivel inicial, que por hipótesis es acotado. La existencia de una subsucesión convergente está garantizada por el Teorema de Weierstrass. □

Funciones que satisfacen estas hipótesis son, por ejemplo, toda función que satisfaga $f(x) \to \infty$ cuando $x \to \infty$. Estas funciones se conocen como funciones coercivas.

Para nuestro primer teorema de convergencia, vamos a solicitar que las direcciones de descenso cumplan con dos propiedades adicionales: la siguiente condición del ángulo mínimo

$$\cos(\theta_k) \geq \delta > 0 \tag{3.5}$$

donde θ_k es el ángulo entre d_k y $\nabla f(x_k)$, y la existencia de una cota uniforme para la direcciones d_k

$$\|d_k\| \leq M \quad \forall M \tag{3.6}$$

Ambas condiciones son muy razonables y fáciles de implementar en la práctica. Por ejemplo, para la segunda condición alcanza con normalizar las direcciónes d_k y/o imponer una cota.

Veamos un pequeño lema

Lema 4. *Sea $\{x_k\}$ una sucesión generada por un método de gradiente $x_{k+1} = x_k + \alpha_k d_k$ que contiene al menos un punto de acumulación, y supongamos que $\{d_k\}$ es una dirección de descenso* (3.4)*. Entonces se tiene*

$$f(x_k) - f(x_{k+1}) \to 0. \tag{3.7}$$

Demostración. Dado que $\{f(x_k)\}$ es monótonamente no creciente, $\{f(x_k)\}$ o bien converge a un valor finito o diverge a $-\infty$. Sea $\bar{x}$ un punto de acumulación de $\{x_k\}$. Como f es continua, $f(\bar{x})$ es un punto de acumulación de $\{f(x_k)\}$. Por lo que tanto, no puede diverger, y entonces toda la sucesión $\{f(x_k)\}$ converge a $f(\bar{x})$. □

Teorema 5 (Los puntos de acumulación de los métodos de gradiente son estacionarios)**.** *Sea $\{x_k\}$ una sucesión generada por un método de gradiente $x_{k+1} = x_k + \alpha_k d_k$, y supongamos que $\{d_k\}$ es una dirección de descenso (ecuación* (3.4)*) y que α_k se elige mediante la regla de minimización, la regla de minimización limitada o la regla de Armijo. Entonces, todo punto de acumulación de $\{x_k\}$ satisface $\nabla f(x_k) = 0$*

Demostración. Consideremos primero la regla de Armijo. Por la definición de la regla de Armijo, se tiene que

$$f(x_k) - f(x_{k+1}) \geq -\sigma\alpha_k \nabla f(x_k) \cdot d_k \geq 0 \tag{3.8}$$

donde la última desigualdad se debe a que d_k es una dirección de descenso. Por lo tanto, el lado derecho de la relación anterior tiende a 0.

Procedamos por el absurdo. Consideremos una subsucesión $\{x_k\}_{k\in\mathcal{K}}$ que converge a $\bar{x}$, y supongamos que $\bar{x}$ no es estacionario. Por la condición del ángulo mínimo (3.5), se tiene que

$$\limsup_{k\to\infty, k\in\mathcal{K}} \nabla f(x_k) \cdot d_k < 0,$$

y por lo tanto, de las ecuaciones (3.7) y (3.8), se obtiene

$$\{\alpha_k\}_{\mathcal{K}} \to 0.$$

Por lo tanto, por la definición de la regla de Armijo, debe existir un índice $\bar{k} \geq 0$ tal que

$$f(x_k) - f(x_k + \alpha_k d_k) < -\sigma \alpha_k \nabla f(x_k) \cdot d_k, \quad \forall k \in \mathcal{K}, k \geq \bar{k}. \tag{3.9}$$

es decir, el tamaño de paso inicial s se reducirá al menos una vez para todo $k \in \mathcal{K}, k \geq \bar{k}$. Como $\{d_k\}$ está acotada, existe una subsucesión $\{d_k\}_{\bar{\mathcal{K}}}$ tal que $\{d_k\}_{\bar{\mathcal{K}}} \to \bar{d}$.

Ahora bien, de la ec. (3.9), se tiene que

$$\frac{f(x_k) - f(x_k + \alpha_k d_k)}{\alpha_k} < -\sigma \nabla f(x_k) \cdot d_k, \quad \forall k \in \bar{\mathcal{K}}, k \geq \bar{k}. \tag{3.10}$$

Utilizando el teorema del valor medio, esta relación se escribe como

$$-\nabla f(x_k + \bar{\alpha}^k d_k) \cdot d_k < -\sigma \nabla f(x_k) \cdot d_k, \quad \forall k \in \bar{\mathcal{K}},\ k \geq \bar{k},$$

donde $\bar{\alpha}^k$ es un escalar en el intervalo $[0, \alpha_k]$. Tomando límites en la relación anterior, obtenemos

$$-\nabla f(\bar{x}) \cdot \bar{d} \leq -\sigma \nabla f(\bar{x}) \cdot \bar{d},$$

o

$$0 \leq (1 - \sigma) \nabla f(\bar{x}) \cdot \bar{d}.$$

Dado que $\sigma < 1$, se deduce que

$$0 \leq \nabla f(\bar{x}) \cdot \bar{d}, \tag{3.11}$$

lo cual contradice la suposición de que $\{d_k\}$ es una dirección de descenso. Esto demuestra el resultado para la regla de Armijo.

El resultado se sigue fácilmente para cualquier regla que proporcione una mayor reducción en el costo en cada iteración que la regla de Armijo. En la ecuación (3.9), podemos intercalar

$$f(x_k) - f(x_{k+1}) \geq f(x_k) - f(\bar{x}^{k+1}) \geq -\sigma \bar{\alpha}^k \nabla f(x_k) \cdot d_k. \tag{3.12}$$

donde $\bar{x}^{k+1}$ es el punto generado a partir de x_k mediante la regla de Armijo, y $\bar{\alpha}^k$ el correspondiente tamaño de paso. A partir de aquí, podemos repetir los mismos argumentos que antes. □

Lema 5 (del descenso)**.** *Sea $f : \mathbb{R}^n \to \mathbb{R}$ una función continuamente diferenciable, y sean x e y dos vectores en $\mathbb{R}^n$. Supongamos que*

$$\|\nabla f(x + t\, y) - \nabla f(x)\| \leq L\, t\, \|y\|, \quad \forall\, t \in [0, 1], \tag{3.13}$$

donde L es un escalar. Entonces,

$$f(x + y) \leq f(x) + y \cdot \nabla f(x) + \frac{L}{2} \|y\|^2.$$

Demostración. Sea t un parámetro escalar y defina

$$g(t) = f(x + t\,y).$$

Por la regla de la cadena se tiene que

$$\frac{dg}{dt}(t) = y \cdot \nabla f(x + t\,y).$$

Ahora,

$$\begin{aligned} f(x+y) - f(x) &= g(1) - g(0) \\ &= \int_0^1 \frac{dg}{dt}(t)\,dt \\ &= \int_0^1 y \cdot \nabla f(x + t\,y)\,dt. \end{aligned}$$

Sumando y restando $y \cdot \nabla f(x)$, obtenemos

$$f(x+y) - f(x) = \int_0^1 y \cdot \nabla f(x)\,dt + \int_0^1 y \cdot \Big(\nabla f(x + t\,y) - \nabla f(x)\Big)\,dt.$$

Utilizando la desigualdad de Cauchy-Schwarz en el segundo término, obtenemos

$$\begin{aligned} \int_0^1 y \cdot \Big(\nabla f(x + t\,y) - \nabla f(x)\Big)\,dt &\leq \int_0^1 \|y\| \cdot \|\nabla f(x + t\,y) - \nabla f(x)\|\,dt \\ &\leq \int_0^1 \|y\| \cdot (L\,t\,\|y\|)\,dt \\ &= L\,\|y\|^2 \int_0^1 t\,dt \\ &= \frac{L}{2}\|y\|^2. \end{aligned}$$

Asimismo, el primer término es simplemente

$$\int_0^1 y \cdot \nabla f(x)\,dt = y \cdot \nabla f(x).$$

Por lo tanto,

$$f(x+y) \leq f(x) + y \cdot \nabla f(x) + \frac{L}{2}\|y\|^2.$$

□

Teorema 6 (Tamaño de Paso Constante)**.** *Sea $\{x_k\}$ una sucesión generada por un método de gradiente $x_{k+1} = x_k + \alpha_k d_k$, donde $\{d_k\}$ es una dirección de descenso. Supongamos que f cumple la condición de Lipschitz* (3.13) *y que para todo k se tiene $d_k \neq 0$ y*

$$\varepsilon \leq \alpha_k \leq (2 - \varepsilon)\frac{|\nabla f(x_k) \cdot d_k|}{L\|d_k\|^2}, \tag{3.14}$$

y $\varepsilon \in (0, 1]$ es un escalar fijo. Entonces, todo punto límite de $\{x_k\}$ es un punto estacionario de f.

Demostración. Usando el lema 5, obtenemos

$$\begin{aligned} f(x_k) - f(x_k + \alpha_k d_k) &\geq -\alpha_k \nabla f(x_k) \cdot d_k - \frac{1}{2}(\alpha_k)^2 L \|d_k\|^2 \\ &= \alpha_k \big(|\nabla f(x_k) \cdot d_k| - \frac{1}{2}\alpha_k L \|d_k\|^2 \big) \end{aligned} \tag{3.15}$$

El lado derecho de la ec. (3.14) nos da

$$\begin{aligned} \alpha_k L \|d_k\|^2 &\leq (2-\varepsilon)|\nabla f_k \cdot d_k| \\ \frac{1}{2}\alpha_k L \|d_k\|^2 + \frac{\varepsilon}{2}|\nabla f_k \cdot d_k| &\leq |\nabla f_k \cdot d_k| \\ \frac{\varepsilon}{2}|\nabla f_k \cdot d_k| &\leq |\nabla f_k \cdot d_k| - \frac{1}{2}\alpha_k L \|d_k\|^2 \end{aligned} \tag{3.16}$$

Y utilizando esta desigualdad en (3.15),

$$f(x_k) - f(x_k + \alpha_k d_k) \geq \frac{1}{2}\alpha_k \varepsilon |\nabla f(x_k) \cdot d_k| \geq \frac{1}{2}\varepsilon^2 |\nabla f(x_k) \cdot d_k|$$

Ahora, suponiendo que existe un punto de acumulación de $\{x_k\}$ (si no existe, el teorema vale trivialmente), y usando el lema 4, tenemos

$$f(x_k) - f(x_{k+1}) \to 0.$$

La relación anterior implica que $|\nabla f(x_k) \cdot d_k| \to 0$. Junto con la condición del ángulo mínimo, necesariamente se tiene que $\nabla f(x_k) \to 0$. Por lo tanto, para un punto límite $\bar{x}$ de $\{x_k\}$, se tendrá $\nabla f(\bar{x}) = 0$. □

Teorema 7 (Tamaño de paso decreciente)**.** *Sea $\{x_k\}$ una sucesión generada por un método de gradiente*

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

Supóngase que se cumple la condición de Lipschitz (1.23) y que existen escalares positivos c_1, c_2 tales que, para todo k, se tiene

$$c_1 \|\nabla f(x_k)\|^2 \leq -\nabla f(x_k) \cdot d_k, \quad \|d_k\|^2 \leq c_2 \|\nabla f(x_k)\|^2. \tag{1.28}$$

Supóngase además que

$$\alpha_k \to 0, \quad \sum_{k=0}^{\infty} \alpha_k = \infty.$$

Entonces, o bien $f(x_k) \to -\infty$, o la sucesión $\{f(x_k)\}$ converge a un valor finito y $\nabla f(x_k) \to 0$. Además, todo punto límite de $\{x_k\}$ es un punto estacionario de f.

Demostración. Combinando las Ecuaciones (1.25) y (1.28), obtenemos

$$f(x_{k+1}) \leq f(x_k) + \alpha_k \left(\frac{1}{2}\alpha_k L \|d_k\|^2 - \nabla f(x_k) \cdot d_k \right)$$

$$\leq f(x_k) - \alpha_k \left(c_1 - \frac{1}{2}\alpha_k c_2 L \right) \|\nabla f(x_k)\|^2.$$

Dado que el término lineal en α_k domina al término cuadrático en α_k para α_k suficientemente pequeños, y puesto que $\alpha_k \to 0$, existe una constante positiva c y un índice $\bar{k}$ tal que, para todo $k \geq \bar{k}$,

$$f(x_{k+1}) \leq f(x_k) - \alpha_k c \|\nabla f(x_k)\|^2. \tag{1.29}$$

A partir de esta relación, se observa que para $k \geq \bar{k}$ la sucesión $\{f(x_k)\}$ es monótonamente decreciente, de modo que o bien $f(x_k) \to -\infty$ o $\{f(x_k)\}$ converge a un valor finito. En este último caso, sumando la Ecuación (1.29) para todos los $k \geq \bar{k}$ obtenemos

$$c \sum_{k=\bar{k}}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 \leq f(x^{\bar{k}}) - \lim_{k\to\infty} f(x_k) < \infty.$$

Esto implica que no puede existir un $\varepsilon > 0$ tal que

$$\|\nabla f(x_k)\|^2 > \varepsilon$$

para todo k mayor que algún $\bar{k}$, pues ello contradeciría la suposición $\sum_{k=0}^{\infty} \alpha_k = \infty$. Por lo tanto, debemos tener

$$\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0.$$

Para demostrar que, en efecto, $\nabla f(x_k) \to 0$, supongamos lo contrario; es decir,

$$\limsup_{k\to\infty} \|\nabla f(x_k)\| \geq \varepsilon > 0. \tag{1.30}$$

Sean $\{m_j\}$ y $\{n_j\}$ sucesiones de índices tales que

$$m_j < n_j < m_{j+1}, \quad \frac{\varepsilon}{3} < \|\nabla f(x_k)\| \quad \text{para } m_j \leq k < n_j, \tag{1.31}$$

$$\|\nabla f(x_k)\| \leq \frac{\varepsilon}{3} \quad \text{para } n_j \leq k < m_{j+1}. \tag{1.32}$$

Sea además $\bar{j}$ suficientemente grande tal que

$$\sum_{k=m_{\bar{j}}}^{\infty} \alpha_k \|\nabla f(x_k)\|^2 < \frac{\varepsilon^2}{9L\sqrt{c_2}}. \tag{1.33}$$

Para cualquier $j \geq \bar{j}$ y para cualquier m con $m_j \leq m \leq n_j - 1$, se tiene:

$$\|\nabla f(x^{n_j}) - \nabla f(x^m)\| \leq \sum_{k=m}^{n_j-1} \|\nabla f(x_{k+1}) - \nabla f(x_k)\|$$

$$\leq L \sum_{k=m}^{n_j-1} \|x_{k+1} - x_k\| = L \sum_{k=m}^{n_j-1} \alpha_k \|d_k\|$$

$$\leq L\sqrt{c_2} \sum_{k=m}^{n_j-1} \alpha_k \|\nabla f(x_k)\|$$

$$\leq \frac{3L\sqrt{c_2}}{\varepsilon} \sum_{k=m}^{n_j-1} \alpha_k \|\nabla f(x_k)\|^2$$
$$\leq \frac{3L\sqrt{c_2}}{\varepsilon} \cdot \frac{\varepsilon^2}{9L\sqrt{c_2}} = \frac{\varepsilon}{3}.$$

(donde las dos últimas desigualdades se obtienen usando las Ecuaciones (1.31) y (1.33)). Así, se tiene

$$\|\nabla f(x^m)\| \leq \|\nabla f(x^{n_j})\| + \frac{\varepsilon}{3} \leq \frac{2\varepsilon}{3}, \quad \forall\, j \geq \bar{j},\ m_j \leq m \leq n_j - 1.$$

Por consiguiente, usando también la Ecuación (1.32), para todo $m \geq m_j$ se cumple que

$$\|\nabla f(x^m)\| \leq \frac{2\varepsilon}{3}.$$

Esto contradice la Ecuación (1.30), lo que implica que

$$\lim_{k\to\infty} \nabla f(x_k) = 0.$$

Finalmente, si $\bar{x}$ es un punto límite de la sucesión $\{x_k\}$, entonces $f(x_k)$ converge al valor finito $f(\bar{x})$. Por lo tanto, tenemos que $\nabla f(x_k) \to 0$, lo que implica que $\nabla f(\bar{x}) = 0$. □

Teorema 8 (Teorema de Captura)**.** *Sea f continuamente diferenciable y sea $\{x_k\}$ una sucesión que satisface*

$$f(x_{k+1}) \leq f(x_k) \quad \textit{para todo } k,$$

generada por un método de gradiente

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

y supongamos que es convergente en el sentido de que todo punto límite de las sucesiones que genera es un punto estacionario de f. Además, supongamos que existen escalares $s > 0$ y $c > 0$ tales que para todo k se tiene

$$\alpha_k \leq s, \quad \|d_k\| \leq c\,\|\nabla f(x_k)\|.$$

Sea x^* un mínimo local de f, el cual es el único punto estacionario de f dentro de algún conjunto abierto. Entonces existe un conjunto abierto S que contiene a x^* tal que si $x_k \in S$ para algún $k \geq 0$, entonces $x_k \in S$ para todo $k \geq k$ y $\{x_k\}$ converge a x^*. Además, dado cualquier escalar $\bar{\varepsilon} > 0$, el conjunto S puede escogerse de modo que $\|x - x^*\| < \bar{\varepsilon}$ para todo $x \in S$.

Demostración. Supongamos que existe $\rho > 0$ tal que

$$f(x^*) < f(x), \quad \forall\, x \text{ con } \|x - x^*\| \leq \rho.$$

Definimos, para $t \in [0, \rho]$,

$$\phi(t) = \text{mín}\{f(x) - f(x^*) \mid t \leq \|x - x^*\| \leq \rho\}.$$

Notamos que ϕ es una función monótonamente no decreciente en t y que $\phi(t) > 0$ para todo $t \in (0, \rho]$. Dado cualquier $\varepsilon \in (0, \rho]$, elige $r \in (0, \varepsilon)$ tal que

$$\|x - x^*\| < r \implies \|x - x^*\| + s\, c\, \|\nabla f(x)\| < \varepsilon. \tag{1.34}$$

Considera el conjunto abierto

$$S = \{x \mid \|x - x^*\| < \varepsilon,\ f(x) < f(x^*) + \phi(r)\}.$$

Afirmamos que si $x_k \in S$ para algún k, entonces $x_{k+1} \in S$. En efecto, si $x_k \in S$, por la definición de ϕ y S se tiene

$$\phi(\|x_k - x^*\|) \leq f(x_k) - f(x^*) < \phi(r).$$

Dado que ϕ es monótona no decreciente, esto implica que $\|x_k - x^*\| < r$. Entonces, por (1.34),

$$\|x_{k+1} - x^*\| + s\, c\, \|\nabla f(x_k)\| < \varepsilon.$$

Además, utilizando las hipótesis $\alpha_k \leq s$ y $\|d_k\| \leq c\, \|\nabla f(x_k)\|$, tenemos

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \|\alpha_k\, d_k\| \leq \|x_k - x^*\| + s\, c\, \|\nabla f(x_k)\|,$$

de donde se deduce que $\|x_{k+1} - x^*\| < \varepsilon$. Además, puesto que $f(x_{k+1}) < f(x_k)$, se tiene que

$$f(x_{k+1}) - f(x^*) < \phi(r),$$

lo que implica que $x_{k+1} \in S$.

Por inducción, si existe algún $\bar{k}$ tal que $x^{\bar{k}} \in S$, entonces $x_k \in S$ para todo $k \geq \bar{k}$. Sea $\bar{S}$ el cierre de S. Dado que $\bar{S}$ es compacto, la sucesión $\{x_k\}$ tendrá al menos un punto límite, el cual, por hipótesis, debe ser un punto estacionario de f. Pero el único punto estacionario de f en $\bar{S}$ es x^* (puesto que $\|x - x^*\| \leq \rho$ para todo $x \in \bar{S}$). Por ello, $x_k \to x^*$. Finalmente, dado cualquier $\bar{\varepsilon} > 0$, podemos elegir $\varepsilon \leq \bar{\varepsilon}$ de modo que $\|x - x^*\| < \bar{\varepsilon}$ para todo $x \in S$. □

Apéndice A

Matrices simétricas y definidas positivas

Teorema 9. *Let A be a symmetric $n \times n$ matrix. Then:*

(a) The eigenvalues of A are real.

(b) The matrix A has a set of n mutually orthogonal, real, and nonzero eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

(c) Suppose that the eigenvectors in part (b) have been normalized so that $\|\mathbf{x}_i\| = 1$ for each i. Then

$$A = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i',$$

where λ_i is the eigenvalue corresponding to $\mathbf{x}_i$.

Teorema 10. *Sea A una matriz simétrica de $n \times n$, sean $\lambda_1 \leq \cdots \leq \lambda_n$ sus valores propios (reales), y sean $x_1, \ldots, x_n$ los vectores propios ortogonales asociados, normalizados de modo que $\|x_i\| = 1$ para todo i. Entonces:*

(a) $\|A\| = \rho(A) = \max\{|\lambda_1|, |\lambda_n|\}$, donde $\|\cdot\|$ es la norma de la matriz inducida por la norma euclidiana.

(b) $\lambda_1 \|y\|^2 \leq y \cdot Ay \leq \lambda_n \|y\|^2$ para todo $y \in \mathbb{R}^n$.

Demostración. (a) Sabemos que $\|A\| \geq \rho(A)$, por lo que debemos demostrar la desigualdad inversa. Sea un vector arbitrario $y \in \mathbb{R}^n$, el cual podemos expresar en la forma

$$y = \sum_{i=1}^{n} \xi_i x_i,$$

donde cada ξ_i es un escalar. Usando la ortogonalidad de los vectores x_i y el teorema de Pitágoras, tenemos

$$\|y\|^2 = \sum_{i=1}^{n} |\xi_i|^2 \|x_i\|^2.$$

Aplicando nuevamente el teorema de Pitágoras, se obtiene

$$\|Ay\|^2 = \left\| \sum_{i=1}^{n} \lambda_i \xi_i x_i \right\|^2 = \sum_{i=1}^{n} |\lambda_i|^2 |\xi_i|^2 \|x_i\|^2.$$

Dado que máx$\{|\lambda_1|, |\lambda_n|\} = \rho(A)$, se tiene que

$$\|Ay\|^2 \leq \rho^2(A) \sum_{i=1}^{n} |\xi_i|^2 \|x_i\|^2 = \rho^2(A)\|y\|^2.$$

Esto implica que $\|A\| \leq \rho(A)$. Como ya se sabía que $\|A\| \geq \rho(A)$, se concluye que

$$\|A\| = \rho(A) = \text{máx}\{|\lambda_1|, |\lambda_n|\}.$$

(b) De manera similar a la parte (a), expresamos un y genérico en $\mathbb{R}^n$ como

$$y = \sum_{i=1}^{n} \xi_i x_i.$$

Entonces,

$$y \cdot Ay = \sum_{i=1}^{n} \lambda_i |\xi_i|^2 \|x_i\|^2 = \sum_{i=1}^{n} \lambda_i |\xi_i|^2,$$

y

$$\|y\|^2 = \sum_{i=1}^{n} |\xi_i|^2 \|x_i\|^2 = \sum_{i=1}^{n} |\xi_i|^2.$$

Dado que $\lambda_1 \leq \lambda_i \leq \lambda_n$ para cada i, se sigue que

$$\lambda_1 \|y\|^2 = \lambda_1 \sum_{i=1}^{n} |\xi_i|^2 \leq \sum_{i=1}^{n} \lambda_i |\xi_i|^2 \leq \lambda_n \sum_{i=1}^{n} |\xi_i|^2 = \lambda_n \|y\|^2.$$

Estas relaciones prueban el resultado deseado. □

Teorema 11. *(a) Para cualquier matriz A de $n \times n$, la matriz $A \cdot A$ es simétrica y semidefinida positiva. La matriz $A \cdot A$ es definida positiva si y solo si A tiene rango n. En particular, si $m = n$, $A \cdot A$ es definida positiva si y solo si A es no singular.*

(b) Una matriz cuadrada simétrica es semidefinida positiva (respectivamente, definida positiva) si y solo si todos sus valores propios son no negativos (respectivamente, positivos).

(c) La inversa de una matriz simétrica definida positiva es simétrica y definida positiva.

Demostración. (a) La simetría es evidente. Para cualquier vector $x \in \mathbb{R}^n$, se tiene

$$x \cdot A \cdot Ax = \|Ax\|^2 \geq 0,$$

lo cual establece que $A \cdot A$ es semidefinida positiva. La definida positividad se obtiene si y solo si la desigualdad es estricta para todo $x \neq 0$, lo que ocurre si y solo si $Ax \neq 0$ para todo $x \neq 0$. Esto es equivalente a que A tenga rango n.

(b) Sea λ y x un valor propio y un vector propio real no nulo asociado, respectivamente, de una matriz simétrica semidefinida positiva A. Entonces

$$0 \leq x \cdot Ax = \lambda\, x \cdot x = \lambda\, \|x\|^2,$$

lo que prueba que $\lambda \geq 0$. Para la recíproca, sea y un vector arbitrario en $\mathbb{R}^n$. Sean $\lambda_1, \ldots, \lambda_n$ los valores propios de A, asumidos no negativos, y sean $x_1, \ldots, x_n$ un conjunto

correspondiente de vectores propios reales, no nulos y ortogonales. Expresamos y en la forma

$$y = \sum_{i=1}^{n} \xi_i x_i.$$

Entonces,

$$y \cdot Ay = \left(\sum_{i=1}^{n} \xi_i x_i \right) \cdot A \left(\sum_{i=1}^{n} \xi_i x_i \right) = \sum_{i=1}^{n} \lambda_i \xi_i^2,$$

y

$$\|y\|^2 = \sum_{i=1}^{n} \xi_i^2.$$

Dado que cada $\lambda_i \geq 0$, se tiene

$$\sum_{i=1}^{n} \lambda_i \xi_i^2 \geq 0,$$

lo que prueba que A es semidefinida positiva. La demostración para el caso de matrices definidas positivas es similar.

(c) Los valores propios de A^{-1} son los recíprocos de los valores propios de A (Prop. A.13(e)), de modo que, aplicando la parte (b), se concluye que A^{-1} es simétrica y definida positiva. □

Apéndice B

Resultados de Análisis Matemático

Recordemos las definiciones formales del gradiente y el Hessiano, en términos de sus propiedades de aproximación de una función dada.

Definición 13. *Sea $f : U \to \mathbb{R}$ una función definida en un conjunto abierto $U \subset \mathbb{R}^n$ que es diferenciable en un punto $x \in U$. El* ***gradiente*** *de f en x, denotado por $\nabla f(x) \in \mathbb{R}^n$, es el único vector que cumple*

$$\lim_{h \to 0} \frac{\left| f(x+h) - f(x) - \nabla f(x) \cdot h \right|}{\|h\|} = 0.$$

Esta definición expresa que la función f puede aproximarse linealmente cerca de x mediante la aplicación del producto escalar $\nabla f(x) \cdot h$, siendo el error de esta aproximación despreciable en comparación con $\|h\|$ cuando h se aproxima a cero.

Definición 14. *Si f es dos veces diferenciable en x, el* ***hessiano*** *de f en x, denotado por $H_f(x)$ (o simplemente $\nabla^2 f(x)$), es la única forma bilineal simétrica (o, de forma equivalente, la única matriz simétrica) que satisface*

$$\lim_{h \to 0} \frac{\left| f(x+h) - f(x) - \nabla f(x) \cdot h - \frac{1}{2}\, h^T H_f(x)\, h \right|}{\|h\|^2} = 0.$$

En esta definición:

1. $\nabla f(x) \cdot h$ representa la aproximación lineal (primer orden) de f en x,
2. $\frac{1}{2}\, h^T H_f(x)\, h$ representa la aproximación cuadrática (segundo orden),
3. y el límite garantiza que el error de aproximar $f(x+h)$ mediante esta expansión cuadrática es despreciable en comparación con $\|h\|^2$ cuando h tiende a cero.

Teorema 12. *(a) Una sucesión acotada de vectores en $\mathbb{R}^n$ converge si y sólo si tiene un único punto límite.*

(b) Una sucesión en $\mathbb{R}^n$ converge si y sólo si es una sucesión de Cauchy.

(c) Toda sucesión acotada en $\mathbb{R}^n$ tiene al menos un punto límite.

(d) Sea $\{z_k\}$ una sucesión escalar. Si $\limsup_{k\to\infty} z_k$ ($\liminf_{k\to\infty} z_k$) es finito, entonces es el mayor (respectivamente, el menor) punto límite de $\{z_k\}$.

Teorema 13 ((Teorema de Weierstrass)). *Sea X un subconjunto no vacío de $\mathbb{R}^n$ y sea $f : X \to \mathbb{R}$ semicontinua inferiormente en todos los puntos de X. Asúmase que se cumple una de las tres condiciones siguientes:*

1. *X es compacto.*
2. *X es cerrado y f es coerciva.*
3. *Existe un escalar γ tal que el conjunto de nivel*

$$\{x \in X \mid f(x) \leq \gamma\}$$

es no vacío y compacto.

Entonces, el conjunto de mínimos de f sobre X es no vacío y compacto.

Demostración. Asúmase la condición (1). Sea $\{z_k\} \subset X$ una sucesión tal que

$$\lim_{k\to\infty} f(z_k) = \inf_{z\in X} f(z).$$

Dado que X es acotado, esta sucesión tiene al menos un punto límite x [Prop. A.5(c)]. Dado que X es cerrado, x pertenece a X, mientras que la semicontinuidad inferior de f implica que

$$f(x) \leq \lim_{k\to\infty} f(z_k) = \inf_{z\in X} f(z).$$

Por lo tanto, debemos tener $f(x) = \inf_{z\in X} f(z)$. El conjunto de todos los mínimos de f sobre X es el conjunto de nivel

$$\{x \in X \mid f(x) \leq \inf_{z\in X} f(z)\},$$

el cual es cerrado por la semicontinuidad inferior de f [Prop. A.7(e)], y por lo tanto compacto, dado que X es acotado.

Asúmase la condición (2). Considérese una sucesión $\{z_k\}$ como en la demostración de la parte (a). Dado que f es coerciva, $\{z_k\}$ debe ser acotada, y la demostración procede como la demostración de la parte (a).

Asúmase la condición (3). Si el γ dado es igual a $\inf_{z\in X} f(z)$, el conjunto de mínimos de f sobre X es

$$\{x \in X \mid f(x) \leq \gamma\},$$

y puesto que por hipótesis este conjunto es no vacío, hemos terminado. Si $\gamma > \inf_{z\in X} f(z)$, considérese una sucesión $\{z_k\}$ como en la demostración de la parte (a). Entonces, para todo k suficientemente grande, z_k debe pertenecer al conjunto

$$\{x \in X \mid f(x) \leq \gamma\}.$$

Dado que este conjunto es compacto, $\{z_k\}$ debe ser acotada, y la demostración procede como la demostración de la parte (a). □