FUNDAMENTO DE OPTIMIZACIÓN 1S-2024
RESUMEN DE CLASES

Clase 1 (18/03/2024) - Introducción al curso y repaso

Para introducir el curso comenzamos con una presentación hecha por el prof. Marcelo Fiori donde se muestran diversos ejemplos de optimización en problemas de ingeniería (para ver lo slides puede hacer click AQUI).

Luego comenzamos con un repaso general de algunos conceptos que se ven mayoritariamente en Cálculo 2 como ser:

(1) El diferencial df_p, la matriz Jacobiana $\mathrm{Jf}(p)$, el gradiente $\nabla f(p)$, derivadas parciales $\frac{\partial f}{\partial x_i}(p)$ y direccionales $\frac{\partial f}{\partial v}(p)$. Recordamos algunas importantes propiedades de las derivadas direccionales:
   - (a) $\frac{\partial f}{\partial v}(p) > \frac{\partial f}{\partial w}(p)$ entonces para algún $\epsilon > 0$ se cumple $f(p+vt) > f(p+wt)$ para todo $t \in (0,\epsilon)$.
   - (b) Si $\nabla f(p) \neq 0$, $v_0 = \frac{\nabla f(p)}{||\nabla f(p)||}$ entonces $\frac{\partial f}{\partial v_0}(p) > \frac{\partial f}{\partial v}(p)$, $\forall v : ||v|| = 1$ (se lee: el gradiente es la dirección de mayor crecimiento).
   - (c) Si $\nabla f(p) \neq 0$, $v_0 = -\frac{\nabla f(p)}{||\nabla f(p)||}$ entonces $\frac{\partial f}{\partial v_0}(p) < \frac{\partial f}{\partial v}(p)$, $\forall v : ||v|| = 1$ (se lee: el gradiente opuesto es la dirección de mayor decrecimiento).
   - (d) El signo de $\frac{\partial f}{\partial v}(p)$ indica crecimiento/decrecimiento (si es positivo/negativo entonces la función en p crece/decrece en la dirección v).

(2) Condición necesaria de optimalidad: $\nabla f(x^*) = 0$ para x^* mínimo local.

(3) Ejemplo: Minimizar $f^2_{\mathbb{R}} \to R$ dada por $f(x,y) = x^2 + y^2$
   - sin restricciones (aprovechamos para repasar conceptos como curvas de nivel y vimos que el gradiente es perpedicular a las curvas de nivel)
   - con una restricción del tipo $g(x,y) = 0$ (método de los multiplicadores de Lagrange).

Comentario: El método de los multiplicadores de Lagrange sirve para minimizar una función $f(x)$ sujeto a la restricción del tipo $g(x) = 0$ (donde $f, g : \mathbb{R}^n \to \mathbb{R}$ son de clase C^1). Se considera la función Lagrangiana (o Lagrangiano) $F(x,\lambda) = f(x) - \lambda g(x)$. El Teorema de los multiplicadores de Lagrange dice que si f tiene mínimo en $x = x^*$ y $\nabla g(x^*) \neq 0$ entonces existe un λ^* que verifica $\nabla F(x^*, \lambda^*) = 0$. Entonces la idea es encontrar todos los (x,λ) para los cuales $\nabla F = 0$ y chequear esos posibles x cual minimiza f (debemos estar seguros que f tenga mínimo absoluto en la región considerada sino solo vamos a encontrar un mínimo relativo).

Lectura introductoria recomendada sobre el tema:

- Unas notas (en las páginas 1-3 ya da una idea básica del método)
- Wikipedia

Clase 2 (22/03/2024) - Repaso y más ejemplos

Repasamos algunas propiedad de la Jacobiana (similares a las propiedades que cumple la derivada para funciones de una variable): $J(f+g)(p) = Jf(p) + Jg(p)$, $J(fg)(p) = f(p)Jg(p) + Jf(p)g(p)$ y $J(f \circ g)(p) = Jf(g(p)) \cdot Jg(p)$ (cuando las operaciones tienen sentido y las funciones son diferenciables en todos los lados donde se necesita).

Norma y productos internos en espacios vectoriales. Como ejemplos importantes vimos:

- El producto interno usual de $\mathbb{R}^n$: $\langle x, y\rangle = x^t y$, que da lugar a la norma euclídea (denotada por $||x||_2$ o simplemente por $||x||$ cuando no hay peligro de ambigüedad).
- El producto interno en $\mathbb{R}^{n\times m}$: $\langle A, B\rangle = \mathrm{tr}(A^t \cdot B)$, que da lugar a la norma de Frobenius $||A||_F = \sum_{i,j} A_{i,j}^2$ (juega un papel importante en aplicaciones prácticas).
- La norma espectral: $||A||_2 = \max\{Ax : ||x|| = 1\}$ que cumple la propiedad $||Ax||_2 \leq ||A||_2 \cdot ||x||_2$ (esa propiedad la hace muy útil para probar resultados teóricos). Esta norma también se denote simplemente por $||A||$ si no hay peligro de ambigüedad y la propiedad mencionada se puede escribir como $||Ax|| \leq ||A|| \cdot ||x||$. Es importante observar que $||A||$ es un número real que no depende de x.

Luego vimos ejemplos varios de cálculos de gradiente, mostrando dos formas de cálculo:

- Usando la unicidad de la Jacobiana y la propiedad $||Ax|| \leq ||A|| \cdot ||x||$ para acotar el resto.
- Usando la regla de la cadena (quedó como ejercicio completar detalles).

Los ejemplos fueron:

- Cálculo del gradiente de una forma cuadrática $q : \mathbb{R}^n \to \mathbb{R}$, $q(x) = x^t Ax$.
- Cálculo del gradiente de una función $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = ||Ax||^2$.

Para finalizar mostramos una forma de chequeo para el cálculo del gradiente, aproximando las derivadas parciales usando la definición de límite.

Comentario 1: el contenido corresponde al video 1 del 2023, más un poco de contenido extra como multiplicadores de Lagrange, la norma espectral y vimos con un poco más de detalle el cálculo del gradiente usando la unicidad del Jacobiano y la propiedad de la norma espectral, no solo para el ejemplo de mínimos cuadrados sino también para formas cuadráticas. Para el ejemplo de mínimos cuadrados también mostramos como calcular el gradiente usando la regla de la cadena para Jacobianos.

Comentario 2: Para aquellos estudiantes que no pudieron asistir a clases y están utilizando los videos del curso anterior recomiendo escoger la opción 1 del obligatorio 1 para el cual alcanza con lo visto en el video 1 del 2023.

Clase 3 (01/04/2024)
Clasificación de métodos iterativos y optimización en una variable

Definimos método iterativo y los clasificamos según su velocidad de convergencia en: lineales, superlineales y cuadráticos (convergencia cuadrática implica convergencia superlineal que a su vez implica convergencia lineal).

Vimos algunos métodos de optimización en una variable como el Grid Search, el método de bisección (para raices) y el método Golden Section.

Video del 2023 relacionado: video 2.

Clase 4 (05/04/2024)
Optimización en una variable y optimización convexa

Optimización en una variables (continuación):

- Vimos un ejemplo explícito del Golden Section Search para hallar el mínimo de la función $f(x) = x^3 + 6x^2 - 3x$ en el intervalo $[0, 1]$ (mostramos como construir una tablita de los $\{x_n, y_n, z_n\}$ para las primeras iteraciones del método). Si llamamos $p_n = \frac{x_n+z_n}{2}$ probamos que la convergencia $p_n \to x^*$ es lineal (más concretamente probamos que $\frac{e(p_{n+1})}{e(p_n)} = 1/\Phi$ donde Phi es el número áureo).
- Para finalizar vimos el método de Newton (para f de clase C^3 y $f''(x^*) \neq 0$) donde se construye la secuencia siguiendo el método de la parábola tangente, obteniendo $x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$. Probamos que si el método comienza de un punto x_0 suficientemente cercano al mínimo x^* el método tiene convergencia (al menos) cuadrática.

Optimización convexa:

- Definimos conjuntos convexos y vimos varios ejemplos: los intevalos reales, las bolas en $\mathbb{R}^n$ (con cualquier norma), los conjuntos afines (subespacios vectoriales transladados), semiespacios, intersección de convexos es convexo, poliedros convexos (por ser intersección finitas de semiespacios), la envolente convexa de un subconjunto de puntos de $\mathbb{R}^n$.
- Definimos funciones convexas, vimos la interpretación geométrica y dejamos como ejercicio probar que una función es convexa si y solo si su epigrafo es un conjunto convexo. Probamos el Teorema que establece que un mínimo relativo de una función convexa debe ser también un mínimo absoluto.

Videos del 2023 relacionados: videos 2 y 3.

Clase 5 (08/04/2024)
Optimización convexa

Funciones convexas y estrictamente convexas, como distringuilas por su epigrafo. Funciones cóncavas. Varios ejemplos de funciones convexas o cóncavas. Construcciones: combinación lineal convexa de funciones convexas es convexa, máximo de funciones convexas es convexa y $f(Ax+b)$ es convexa si $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ lo es. Restricción de funciones convexas a segmentos sigue siendo convexa (en el caso que el dominio de f sea un espacio afín esto vale también para restricción a rectas). El subconjunto de nivel $H_c = \{x \in \Omega : f(x) \leq c\}$ es un conjunto convexo si $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}$ es una función convexa (el recíproco no vale). Convexidad y diferenciabilidad: recordamos que era que una función fuese de clase C^k para $k \geq 1$. Vimos que una función diferenciable y convexa verifica $f(z) \geq f(x) + \nabla^t f(x) \cdot (z-x)$. Probamos que en el caso de funciones convexas y diferenciable vale que x^* es mínimo local si y solo si $\nabla f(x^*) = 0$.

Videos del 2023 relacionados: videos 3 y 4.

Clase 6 (12/04/2024)
Optimización convexa y métodos de descenso por gradiente

Enunciamos: si $f \in C^2(\Omega)$ entonces si la matriz Hessiana $\nabla^2 f(x)$ es semi-definida positiva (definida positiva) para todo $x \in \Omega$ entonces f es convexa (convexa estricta). Repaso: cuando f es de clase C^2 la matriz Hessiana es simétrica y por el teorema espectral sus valores propios son todos positivos entonces puede clasificarse en 5 casos: definida positiva/negativa, semi-definida positiva/negativa e indefinida. Vimos como expresar esa condición en relación a su forma cuadrática asociada. Métodos iterativos: cuándo decimos que es un método de descenso, cuándo decimos que es un método de descenso por gradiente. En métodos de descenso por gradiente $x_{k+1} = x_k + \alpha_k d_k$ hace falta definir tres parámetros: elección de dirección d_k (se toman direcciones de decrecimiento, o sea, aquellas direcciones cuyas derivadas direccionales $\frac{\partial f}{\partial d_k}(x_k) = \nabla^t f(x_k) \cdot d_k$ son negativas). Las formas comunes de conseguirlo es tomando $d_k = -D_k \nabla f(x)$ donde D_k es una matriz simétrica definida positiva. Vimos algunos casos particulares: Steepest descent tomando $D_k = I$ (vimos que es bastante simple de implementar pero a veces tiene el problema del zigzagueo por causa de la curvatura). Newton tomando $D_k = (\nabla^2 f(x_k))^{-1}$ (esto soluciona el problema de la curvatura pero tiene el inconveniente que a veces es demasiado costoso). Diagonal scaling tomando $D_k = \mathrm{diag}(d_{k1}, \ldots, d_{kn})$ matriz diagonal $n \times n$ (para que sea una dirección descenso se debe cumplir $\sum_{i=1}^n d_{ki} (\frac{\partial f}{\partial x_i}(x))^2 < 0$).

Videos del 2023 relacionados: video 4.

Clase 7 (15/04/2024)
Métodos de desenso por gradiente, la norma espectral

Continuamos viendo posibilidades para la elección de la dirección:

- Diagonal scaling: una elección común es tomar D_k como la inversa de la matriz diagonal que surge de borrar los ceros de $\nabla^2 f(x_k)$ fuera de la diagonal principal (se usa en la práctica, pero no verifica en general ser direcciones de decrecimiento y a veces ni siquiera está bien definida como por ejemplo cuando $\nabla^2 f(x_k)$ tiene ceros o valores negativos en la diagonal principal. Se requieren ciertas condiciones para que esto funcione, por ejemplo si $\nabla^2 f(x^*)$ tiene la diagonal principal positiva y comenzamos muy cerca de x^*, la matriz Hessiana de puntos cercanos va a tener diagonal positiva usando que f es de clase C^2). Esta forma de elegir se la conoce como "aproximación Diagonal del método de Newton".
- Método de Newton aproximado: Tomar $D_k = (\nabla^2 f(x_0))^{-1}$ para todo k (o ir actualizando este valor cada cierta cantidad $p > 1$ de pasos).

Elección del paso α_k:

- Line search: Elegimos α_k que minimize $f(x_k + \alpha d_k)$ con $\alpha > 0$.
- Limited line search: Idem pero fijamos $s > 0$ y buscamos el mínimo con $\alpha \in (0, s]$.
- Regla de Armijo: Fijar $c \in (0, 1)$, también $\beta \in (0, 1)$ and $s > 0$. Tomamos m el menor natural tal que $\alpha = s\beta^m$ verifica $f(x_k + \alpha d_k) < f(x_k) + c\alpha\nabla f(x_k)^t d_k$. Probamos que existe $\epsilon > 0$ tal que eso se verifica para todo $\alpha \in (0, \epsilon)$.
- Paso fijo $\alpha_k = \alpha$
- Paso decreciente $\alpha_k \to 0$ decreciente y con serie divergente.

Criterio de parada: Usualmente $||\nabla(x_k)|| \leq \epsilon$ o la versión normalizada.

Estudiamos el caso simple de steepest descent con paso fijo cuando el operador $T(x) = x - \alpha\nabla f(x)$ resulta una contracción (veremos que eso pasa por ejemplo en el caso de formas cuadráticas definidas positivas para una buena elección de α).

Antes de pasar al análisis del caso cuadrático vimos un poco de norma espectral:

- Propiedades básicas $||Ax|| \leq ||A||||x||$ para todo $x \in \mathbb{R}^n$ y $||AB|| \leq ||A|| \cdot ||B||$ para todo $A, B \in \mathbb{R}^{n\times n}$(en $\mathbb{R}^n$ se considera la norma euclideana y en el espacio de matrices la norma espectral).
- Probamos que $||A|| = \sqrt{\lambda}$ donde λ es el mayor valor propio de A^tA. En el caso que $A = A^t$ entonces $||A|| = \max\{|\lambda| : \lambda \in \mathrm{Spec}(A)\}$ y si además A es semidefinida positiva entonces $||A||$ es el mayor propio de A.
- Si $A = \alpha I + \beta M$ con $M^t = M$ entonces $||A|| = \max\{|\alpha + \beta\lambda| : \lambda \in \mathrm{Spec}(M)\}$.
- Si A es simétrica y definida positiva, vimos que su número de condición $\kappa(A) = ||A|| \cdot ||A^-1|| = \frac{\lambda_{max}}{\lambda_{min}}$, donde λ_{max} y λ_{min} son el mayor y menor valor propio de A, respectivamente.

Clase 8 (19/04/2024)
Análisis del caso cuadrático y el método Heavy-Ball

- Hicimos el análisis del caso cuadrático $f(x) = x^t Ax$ donde $A \in \mathrm{Sym}^n_{++}$ usando Steepest Descent con paso fijo $\alpha > 0$. Vimos que para cierto rango de valores de α el método converge al óptimo $x^* = 0$, calculamos el valor de α que minimiza la tasa de convergencia obteniendo un valor de $\frac{\kappa-1}{\kappa+1}$.
- Vimos el método Heavy-Ball que depende de dos parámetros α y β. Para valores óptimos (y explícitos) de esos parámetros vimos que obtenemos una mejora en la tasa de convergencie de $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Es es el primer método acelerado que vimos en el curso.

Clase 9 (26/04/2024)
Método de Nesterov y Método de Gradiente Conjugado

- Hicimos una comparación de los dos métodos que vimos la clase anterior, en función del número de iteraciones.
- Vimos un segundo método acelerado: el método de Nesterov, comentamos sobre las ventajas y desventajas del mètodo.
- Vimos el método de Gradiente Conjugado para minimizar una función de la forma $f(x) = x^t Ax - b^t x$ donde A es una matriz simétrica definida positiva. Discutimos dos formas de implementar el método: via subespacios de Krilov y via direcciones conjugadas (i.e. ortogonales con respecto al producto interno inducido por A). Discutimos sobre las ventajas y desventajas del método. Mostramos un pseudo-código para implementar el método (con respecto al segundo abordaje). Hicimos un ejemplo gráfico para comparar este método con Steepest Descent.

Receso por primeros parciales del 27/04 al 08/05

Clase 10 (10/05/2024)
Descenso coordenado, otros

- Vimos el método de descenso coordenado (coordinate descent) para minimizar funciones $f : \prod X_i \to \mathbb{R}$ con $X_i \subseteq \mathbb{R}$ (existe una generalización llamado descenso coordenado con bloques donde los $X_i \subseteq \mathbb{R}^{n_i}$). Describimos el método y mostramos un ejemplo. Para el caso que los X_i son cerrados convexos y $f \in C^1$, vimos condiciones que nos garantizan que los puntos límites de $\{x_k\}$ son puntos estacionarios (Bertsekas Prop. 2.7.1). Comentamos ideas para mejorar la implementación (ej. aleatorizar el orden).
- Optimización con restricciones: consideramos $f : \mathbb{R}^n \to \mathbb{R}$ diferenciable y $\mathcal{X} \subseteq \mathbb{R}^n$ cerrado y convexo. Repasamos los conceptos de mínimo local, mínimo local estricto y puntos factibles. Enunciamos (y probamos) una generalización de la condición necesaria de optimalidad: $\nabla f(x^*)\cdot(x-x^*) \geq 0, \forall x \in \mathcal{X}$. Además, si la f es convexa entonces la condición anterior es también suficiente.
- Definimos la proyección a un convexo cerrado $\mathcal{X} \subseteq \mathbb{R}^n$ como $P_{\mathcal{X}} : \mathbb{R}^n \to \mathcal{X}$ definida por la propiedad $||z - P_{\ddagger}|| \leq ||z - x||, \forall x \in \mathcal{X}$ (o equivalentemente, $P_{\mathcal{X}}(z) = \text{argmin}_{x\in\mathcal{X}} ||z - x||^2$). Vimos que la proyección está bien definida.
- Para $\mathcal{X}$ convexo y cerrado probamos que son equivalentes:
  - (Condición necesaria de optimalidad) $\nabla f(x^*) \cdot (x - x^*) \geq 0, \forall x \in \mathcal{X}$.
  - (Condición de estacionaridad) $x^* = P_{\mathcal{X}}(x^* - \alpha\nabla f(x^*)), \forall \alpha \geq 0$.

Clase 11 (13/05/2024)
Métodos del gradiente condicional (Frank-Wolfe), del gradiente proyectado (normal y acelerado a lo Nesterov)

- Método del gradiente condicional: Dado $x_k \in \mathcal{X}$ (compacto, convexo) buscamos una dirección factible $d_k = \overline{x_k} - x_k$ de forma de minimizar $\frac{\partial f}{\partial d_k}(x_k) = \nabla f(x_k) \cdot (\overline{x_k} - x_k)$ (problema de optimización lineal con restricción $\overline{x_k} \in \mathcal{X}$). Probamos que si x_k no verifica la condición necesaria de optimalidad (restringida a $\mathcal{X}$) entonces d_k es una dirección de descenso y factible. El método tiene garantía de convergencia sublineal $e_k = O(1/k)$. Para iterar definimos como siempre $x_{k+1} = x_k + \alpha_k d_k$, donde para la elección del paso podíamos tomar $\alpha_k \in [0, 1]$ que minimize $f(x_k + \alpha_k d_k)$ (usando Armijo por ejemplo) o simplemente tomar $\alpha_k = \frac{2}{k+2}$.
- Método del gradiente proyectado: Dado $x_k \in \mathcal{X}$ (compacto, convexo) consideramos la dirección factible $d_k = \overline{x_k} - x_k$ donde $\overline{x_k} = P_{\mathcal{X}}(x_k - \nabla f(x_k))$ (en otras variantes se puede tomar $\overline{x_k} = P_{\mathcal{X}}(x_k - s\nabla f(x_k))$ con $s > 0$ fijo). Para iterar definimos como siempre $x_{k+1} = x_k + \alpha_k d_k$, donde para la elección del paso podíamos tomar $\alpha_k \in [0, 1]$ que minimize $f(x_k + \alpha_k d_k)$ (usando Armijo por ejemplo) o simplemente tomar $\alpha_k \in (0, 1]$ fijo ($\alpha = 1$ es una elección común, pero no siempre óptima).
- Vimos casos donde la proyección $P_{\mathcal{X}}$ podía calcularse de forma explícita (ej: bolas, subespacios afines e intersección de semiespacios).
- Vimos una versión acelerada del gradiente proyectado:

$$\begin{cases} x_{k+1} = P_{\mathcal{X}}(y_k - s\nabla f(y_k)) \\ y_{k+1} = x_{k+1} + \beta_{k+1}(x_{k+1} - x_k) \end{cases}$$

  donde $\beta_k = \frac{k-1}{k+2}$.

Clases del 17/05 y del 20/05 suspendidas por enfermedad

Clase 12 (24/05/2024)
Otros tópicos

- Para funciones no diferenciables: subgradiente, subdiferencial, otra generalización de la condición necesaria de optimalidad, ejemplos y extensión de los métodos vistos en el curso (en general funcionan pero se pierde velocidad de convergencia).
- Métodos proximales: el operador proximal, propiedades y uso.
- ADMM (Alternating directions methods of multiplayers): para problemas de la forma $\min_{Ax+Bz=C} f(x) + g(z)$. Explicación del método y ejemplos de uso (por ejemplo para LASSO).
- Descenso por gradiente estocástico: explicación del método, aplicaciones modernas (por ejemplo en redes neuronales), comentarios sobre resultados de convergencia. Optimización en variedades. Algoritmos de punto interior. Combinaciones de métodos.

Clase 13 (27/05/2024)
Teoría de Lagrange y dualidad

- Recordando Lagrange, Regresión Ridge, restricciones duras y suaves.
- Consideramos un problema primal de la forma $\min f_0(x)$ con restricciones $f_i(x) \leq 0$ para $i = 1 : m$ y $h_i(x) = 0$ para $i = 1 : p$; con valor óptimo p^* y dominio $D \subseteq \mathbb{R}^n$. Definimos el Lagrangiano $L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ dado por $L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p \mu_i h_i(x)$ y la función dual $g(\lambda, \mu) = \inf_{x\in D} L(x, \lambda, \mu)$.
- Vimos que g es una función cóncava (por ser ínfimo de funciones lineales, que son cóncavas) y definimos el problema dual: $\max g(\lambda, \mu)$ con restricción $\lambda \geq 0$ con valor óptimo d^*. Probamos que siempre $d^* \leq p^*$ (dualidad débil) y definimos la brecha de dualidad $g = p^* - d^*$. Varios comentarios sobre cuando un problema tiene dualidad fuerte (i.e. cuando $g = 0$). Ejemplo. Complementary slackness y las condiciones de Karush-Kuhn-Tucker (condiciones necesarias para que haya dualidad fuerte). En problemas convexos hay dualidad fuerte bajo condiciones bastantes generales (Condición de Slater).

Clase 14 (31/05/2024)
Notebooks y Obligatorio Final

En esta clase vimos algunos notebooks usando el paquete CVXPY con múltiples ejemplos (disponible en el eva del curso). También discutimos sobre los ejercicios de ambas propuestas para la tarea final del curso.