INSTITUTO DE INFORMÁTICA
Universidade Federal de Goiás

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
TÓPICOS ESPECIAIS EM FUNDAMENTOS DE COMPUTAÇÃO – MATEMÁTICA E ESTATÍSTICA PARA CIÊNCIA DE DADOS
PROF. DR. ROMMEL MELGAÇO BARBOSA

SEMINÁRIOS - CAPÍTULO 07

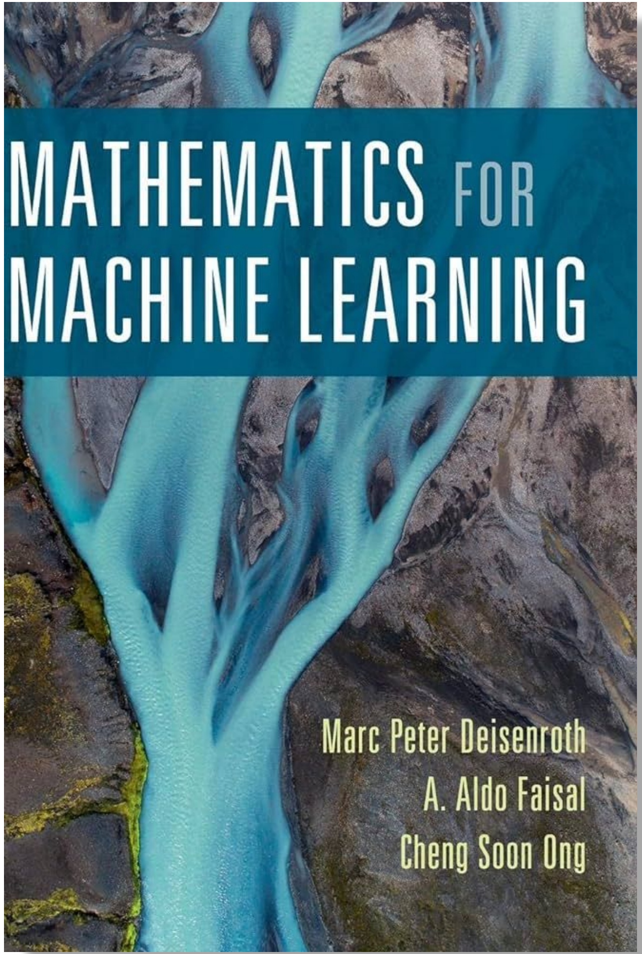
OTIMIZAÇÃO CONTÍNUA

André Rodrigues Coimbra
Rayane Araujo Lima
Renan Rodrigues de Oliveira

Junho/2024

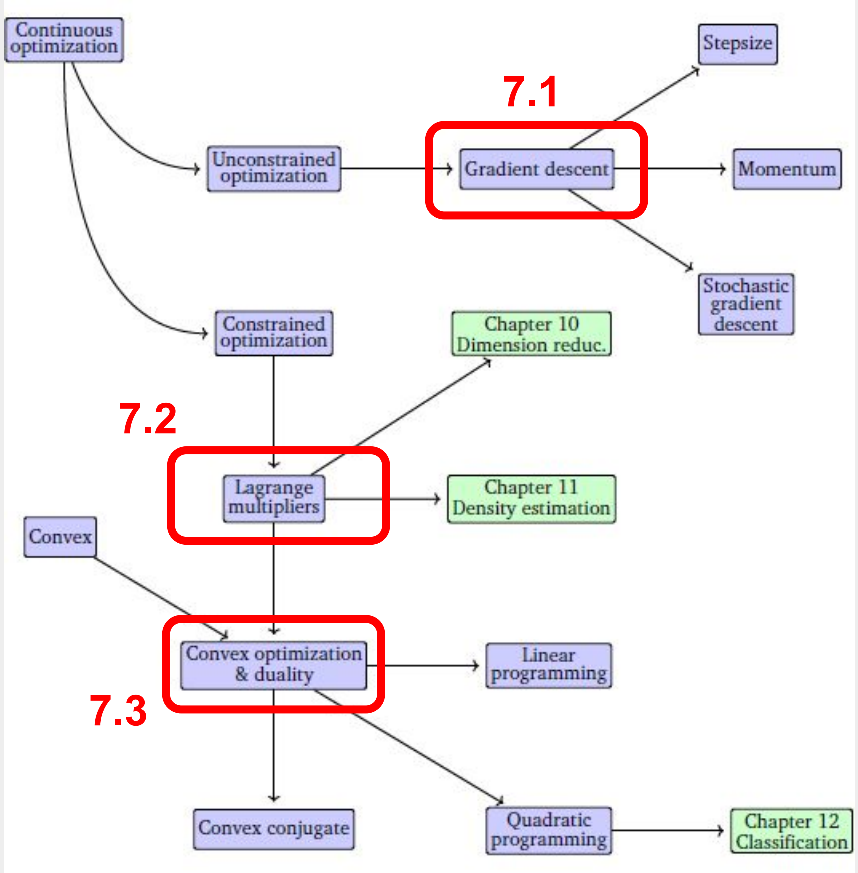
INSTITUTO DE INFORMÁTICA
UFG

UFG
UNIVERSIDADE FEDERAL DE GOIÁS



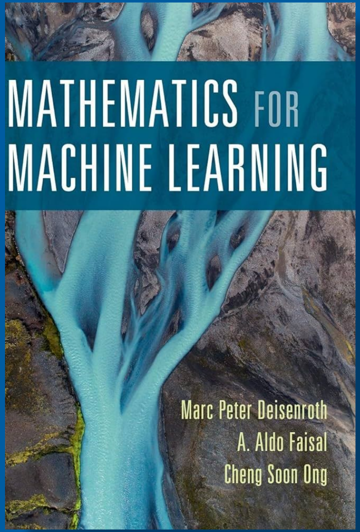
Capítulo 07

Otimização Contínua



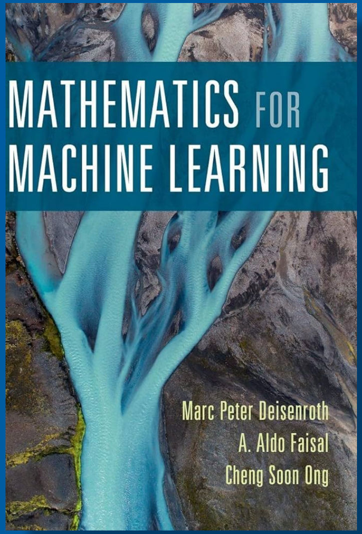
Capítulo 07

Otimização Contínua



7.1 Otimização usando Gradiente Descendente

- Tamanho do Passo
- Momentum
- SGD



Otimização usando Gradiente Descendente

- Considere o problema para encontrar o valor que minimiza uma função

$$\min_{x} f(x),$$

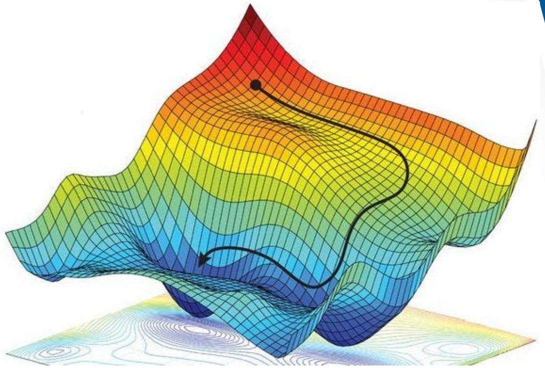
onde $f : \mathbb{R}^d \rightarrow \mathbb{R}$ é uma função objetivo que reflete um problema de otimização.

- Assumimos que a função f é diferenciável e não é possível encontrar analiticamente uma solução direta utilizando operações elementares.
  - Neste caso, o algoritmo gradiente descendente pode ser utilizado para encontrar mínimos locais de funções.

Gradiente e Taxa de Aumento da Função

- O gradiente de uma função multivariável é um vetor que contém todas as derivadas parciais da função.
  - O gradiente aponta na direção da maior taxa de aumento da função e tem magnitude que indica quão rápido a função aumenta nessa direção.
  - Em contexto de minimização, o gradiente descendente deve caminhar na direção oposta ao gradiente da função no ponto atual.

- Considere a superfície da função e uma bola no ponto inicial x_0 .
  - Se a bola for solta, ela naturalmente rolará para baixo, seguindo a trajetória da descida mais íngreme.
  - O valor de $f(x_0)$ diminui mais rapidamente movendo o ponto x_0 na direção oposta ao gradiente da função nesse ponto.

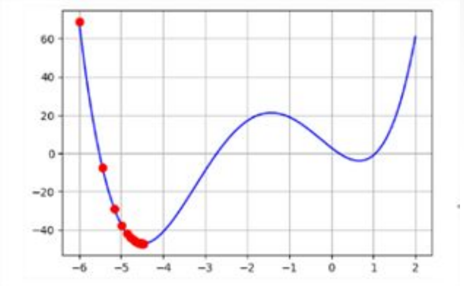


Algoritmo Gradiente Descendente

- Portanto, temos se

$$x_1 = x_0 - \gamma((\nabla f)(x_0))^\top$$

para um tamanho de passo pequeno $\gamma \geq 0$, então $f(x_1) \leq f(x_0)$.



- Dessa forma, para encontrar um ótimo local o algoritmo gradiente descendente deve iterar de acordo com

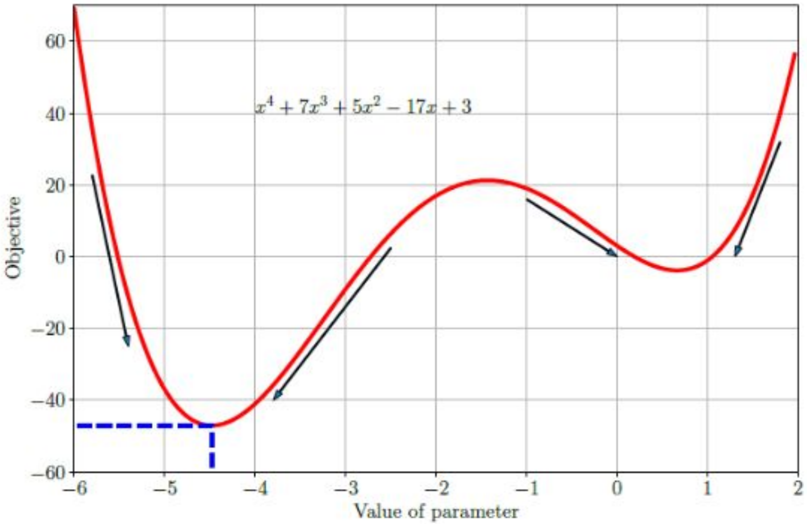
$$x_{i+1} = x_i - \gamma((\nabla f(x_i))^\top .$$

  - Para um tamanho de passo adequado γ_i , a sequência $f(x_0) \geq f(x_1) \geq \ldots \geq f(x_n)$ converge para um mínimo local.

Gradiente Descendente com Função Univariada

- Considere a seguinte função polinomial de uma única variável.

$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$



Gradiente de Funções de uma Única Variável

- Para funções de uma única variável, o gradiente da função é a derivada da função com relação à sua única variável.
  - A derivada de uma função polinomial pode ser encontrada aplicando a regra da potência a cada termo da função.

$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$

- Neste caso, a derivada $f'(x)$ da função $f(x)$ é

$$f'(x) = 4x^3 + 21x^2 + 10x - 17$$

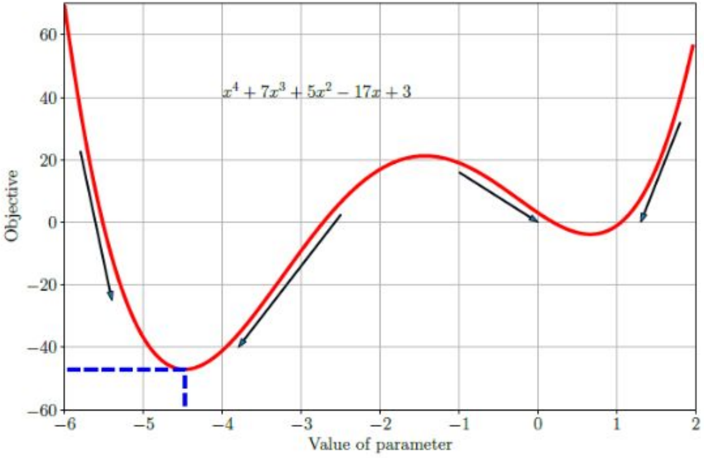
Comportamento da Função

- O valor de $f'(x)$ indica a direção de maior crescimento da função.

Em funções de uma única variável:	
$f'(x) > 0$	Indica que a função está aumentando à medida que *x* aumenta.
$f'(x) < 0$	Indica que a função está diminuindo à medida que *x* aumenta.
$f'(x) = 0$	Indica os pontos críticos da função.

Exemplo

$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$



$$f'(x) = 4x^3 + 21x^2 + 10x - 17$$

```
for i in np.arange(-6, 2.5, 0.5):
  print(f"f'({i}) = {grad_f(i)}")
```

```
f'(-6.0) = -185.0
f'(-5.5) = -102.25
f'(-5.0) = -42.0
f'(-4.5) = -1.25
f'(-4.0) = 23.0
f'(-3.5) = 33.75
f'(-3.0) = 34.0
f'(-2.5) = 26.75
f'(-2.0) = 15.0
f'(-1.5) = 1.75
f'(-1.0) = -10.0
f'(-0.5) = -17.25
f'(0.0) = -17.0
f'(0.5) = -6.25
f'(1.0) = 18.0
f'(1.5) = 58.75
f'(2.0) = 119.0
```

- Na busca de um ponto onde a função atinge um valor mínimo (local ou global):
  - O método do gradiente descendente deve mover-se na direção de $-f'(x)$.

Pontos Estacionários

- Os pontos estacionários são as raízes reais da derivada, ou seja, pontos que possuem gradiente zero.
    - Para verificar se um ponto estacionário é mínimo ou máximo, precisamos calcular a derivada uma segunda vez.

$$f'(x) = 4x^3 + 21x^2 + 10x - 17$$

- Neste caso, a derivada $f''(x)$ da função $f(x)$ é

$$f''(x) = 12x^2 + 42x + 10$$

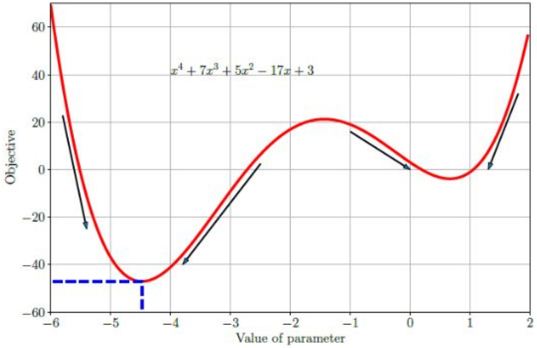
Pontos Estacionários

- Neste caso, $f''(c)$ indica se c é um valor de máximo ou mínimo.

Em funções de uma única variável:	
$f''(c) > 0$	Indica que a função tem concavidade voltada para cima no ponto c. ○ Neste caso, *c* é um ponto de mínimo local da função f.
$f''(c) < 0$	Indica que a função tem concavidade voltada para baixo no ponto c. ○ Neste caso, *c* é um ponto de máximo local da f.

Exemplo

- Como $f'(x)$ uma equação cúbica, em geral esta função possui três soluções quando definida como zero.
    - Na gráfico existem dois pontos são mínimos (em torno de -4.5 e 0.7) e um que é máximo (em torno de -1.4).



$$f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$$
$$f'(x) = 4x^3 + 21x^2 + 10x - 17$$
$$f''(x) = 12x^2 + 42x + 10$$

- Substituindo os valores de c em $f''(c)$.

$$f''(-4.5) = 64 \implies f''(c) > 0$$
$$f''(0.7) = 45.28 \implies f''(c) > 0$$
$$f''(-1.4) = -25.28 \implies f''(c) < 0$$

- Portanto, o ponto médio é um máximo, pois $f''(-1.4) < 0$. Os outros dois pontos estacionários são mínimos.

Direção do Gradiente para Polinômios de Alta Ordem

- Para polinômios de alta ordem, torna-se impraticável analisar o comportamento local da função de maneira analítica.
  - Em tais situações, é necessário iniciar com um valor inicial, por exemplo e seguir a direção do gradiente.

- A direção do gradiente orienta para onde devemos nos mover na busca pelo ponto mínimo ou máximo.
  - No entanto, a direção do gradiente não especifica a magnitude do deslocamento, conhecida como “tamanho do passo”.

Exemplo

- Execução de três passos do método do gradiente descendente para uma função de uma única variável.

Configuração Inicial:

- Função:
  - ◇ $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$
- Gradiente da função:
  - ◇ $\nabla f(x) = 4x^3 + 21x^2 + 10x - 17$
- Ponto Inicial:
  - ◇ $x_0 = -6$
- Taxa de Aprendizagem:
  - ◇ $\gamma = 0.001$

Método do Gradiente Descendente:

- Passo 1:
  - ◇ Calcular o gradiente em $x_0 = -6$:
    $\nabla f(-6) = 4(-6)^3 + 21(-6)^2 + 10(-6) - 17$
    $= 4(-216) + 21(36) - 60 - 17$
    $= -864 + 756 - 60 - 17$
    $= -185$
  - ◇ Atualizar x:
    $x^{Novo} = -6 - 0.001 \times (-185)$
    $\approx -6 + 0.185$
    ≈ -5.815
- Passo 2:
  - ◇ Calcular o gradiente em $x_1 = -5.815$:
    $\nabla f(-5.815) = 4(-5.815)^3 + 21(-5.815)^2 + 10(-5.815) - 17$
    $\approx 4(-197.753) + 21(33.834) - 58.15 - 17$
    $\approx -791.012 + 710.514 - 58.15 - 17$
    ≈ -155.648
  - ◇ Atualizar x:
    $x^{Novo} = -5.815 - 0.001 \times (-155.648)$
    $\approx -5.815 + 0.155648$
    ≈ -5.659

Exemplo

- Execução de três passos do método do gradiente descendente para uma função de uma única variável.

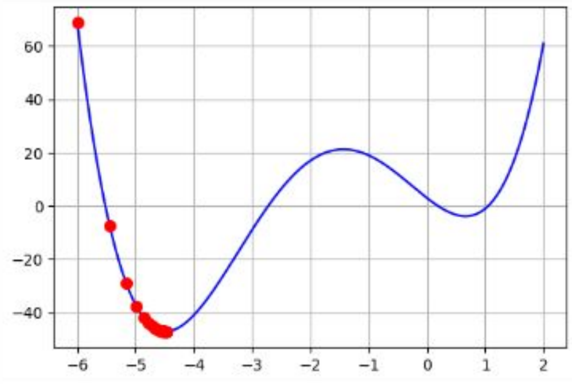
Configuração Inicial:

- Função:
  - $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$
- Gradiente da função:
  - $\nabla f(x) = 4x^3 + 21x^2 + 10x - 17$
- Ponto Inicial:
  - $x_0 = -6$
- Taxa de Aprendizagem:
  - $\gamma = 0.001$

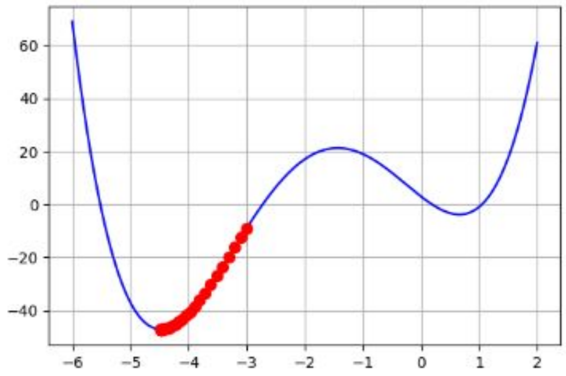
- Passo 3:
  - Calcular o gradiente em $x_2 = -5.659$:
    $\nabla f(-5.659) = 4(-5.659)^3 + 21(-5.659)^2 + 10(-5.659) - 17$
    $\approx 4(-181.225) + 21(32.024) - 56.59 - 17$
    $\approx -724.9 + 672.504 - 56.59 - 17$
    ≈ -125.986
  - Atualizar x:
    $x^{Novo} = -5.659 - 0.001 \times (-125.986)$
    $\approx -5.659 + 0.125986$
    ≈ -5.533

- Passo $4, 5, \ldots, n$:
  - Repetindo o mesmo procedimento. O método deve continuar até que uma condição de parada seja atingida.

Exemplo



(a) O valor inicial foi definido como $x_0 = -6$, com a aproximação do valor do mínimo global após 30 iterações.



(b) O valor inicial foi definido como $x_0 = -3$, com a aproximação do valor do mínimo global após 40 iterações.

Figura 7.3: Exemplo do método do gradiente descendente para a função $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$. O algoritmo segue a direção do gradiente negativo na busca pelo ponto mínimo, com a condição de parada quando $|\nabla f(x)| < \varepsilon$, onde $\varepsilon = 0.1$ e $\gamma = 0.003$.

Gradiente Descendente com Função Multivariada

- Considere a seguinte função quadrática de duas variáveis, x_1 e x_2, que é definida em um espaço bidimensional.

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 5 \\ 3 \end{bmatrix}^{\mathrm{T}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

- A função em sua forma matricial pode ser representada da forma geral como

$$f(x) = \frac{1}{2}x^{\mathrm{T}}Ax - b^{\mathrm{T}}x, \qquad \text{onde } x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, A = \begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} \text{ e } b = \begin{bmatrix} 5 \\ 3 \end{bmatrix}.$$

Forma Algébrica de $f(x)$

$$f(x) = \frac{1}{2}x^T A x - b^T x, \quad \text{onde } x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, A = \begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} \text{ e } b = \begin{bmatrix} 5 \\ 3 \end{bmatrix}.$$

- Cálculo de Ax:
  - $Ax = \begin{bmatrix} 2 & 1 \\ 1 & 20 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \times x_1 + 1 \times x_2 \\ 1 \times x_1 + 20 \times x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 20x_2 \end{bmatrix}$
- Cálculo de $x^T Ax$:
  - $x^T Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2x_1 + x_2 \\ x_1 + 20x_2 \end{bmatrix} = x_1(2x_1 + x_2) + x_2(x_1 + 20x_2)$
    $= 2x_1^2 + x_1x_2 + x_1x_2 + 20x_2^2 = 2x_1^2 + 2x_1x_2 + 20x_2^2$
- Cálculo de $\frac{1}{2}x^T Ax$:
  - $\frac{1}{2}x^T Ax = \frac{1}{2}(2x_1^2 + 2x_1x_2 + 20x_2^2) = x_1^2 + x_1x_2 + 10x_2^2$
- Cálculo de $b^T x$:
  - $b^T x = \begin{bmatrix} 5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 5x_1 + 3x_2$

- Cálculo de $\frac{1}{2}x^T Ax - b^T x$:
  - $\frac{1}{2}x^T Ax - b^T x = (x_1^2 + x_1x_2 + 10x_2^2) - (5x_1 + 3x_2)$
    $= x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$

Dessa forma, temos que $f(x)$
é algebricamente equivalente a

$$f(x_1, x_2) = x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$$

Gradiente de Funções de Várias Variáveis

- Para funções de várias variáveis, o gradiente é o vetor de derivadas parciais da função em relação a cada uma das variáveis.
  - Além de indicar a direção de maior crescimento da função, o vetor de gradientes também representa a taxa de variação da função em cada dimensão do espaço.

- De forma geral, temos que

$$\nabla f(x) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1}, \quad \frac{\partial f(\mathbf{x})}{\partial x_2}, \quad \cdots \quad \frac{\partial f(\mathbf{x})}{\partial x_n},\right]^{\mathrm{T}}$$

Gradiente de Funções de Várias Variáveis

$$f(x_1, x_2) = x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$$

- Cálculo da derivada parcial em relação a x_1:
  - $\frac{\partial}{\partial x_1}(x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2) = 2x_1 + x_2 - 5$
- Cálculo da derivada parcial em relação a x_2:
  - $\frac{\partial}{\partial x_2}(x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2) = x_1 + 20x_2 - 3$
- Portanto, temos que o gradiente de $f(x_1, x_2)$ é

$$\nabla f(x_1, x_2) = \begin{bmatrix} 2x_1 + x_2 - 5 \\ x_1 + 20x_2 - 3 \end{bmatrix}$$

Exemplo

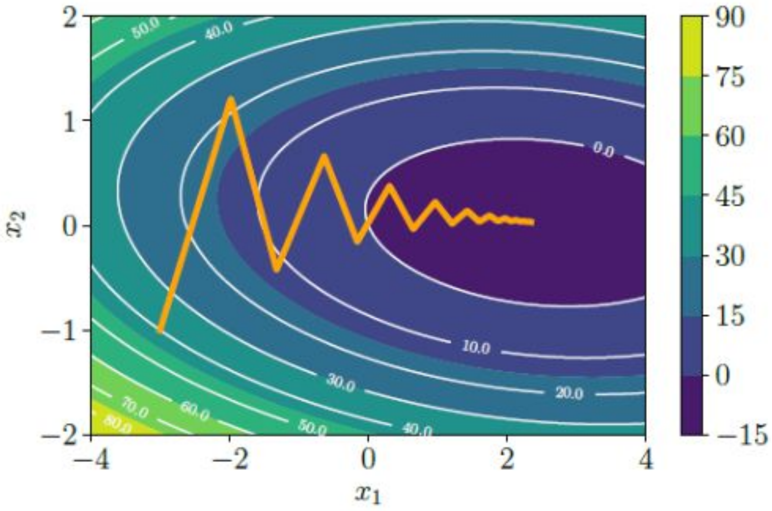


Figura 7.4: Exemplo do método do gradiente descendente para a função $f(x_1, x_2) = x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$ em uma superfície quadrática bidimensional com $\gamma = 0.085$.

Exemplo

- Execução de três passos do método do gradiente descendente para uma função de duas variáveis.

Configuração Inicial:

- Função:
  - $f(x_1, x_2) = x_1^2 + x_1 x_2 + 10x_2^2 - 5x_1 - 3x_2$
- Gradiente da função:
  - $\nabla f(x_1, x_2) = \begin{bmatrix} 2x_1 + x_2 - 5 \\ x_1 + 20x_2 - 3 \end{bmatrix}$
- Ponto Inicial:
  - $x_0 = [-3, -1]$
- Taxa de Aprendizagem:
  - $\gamma = 0.085$

Método do Gradiente Descendente:

- Passo 1:
  - Calcular o gradiente em $x_0 = [-3, -1]$:
    $\frac{\partial}{\partial x_1} = 2(-3) - 1 - 5 = -12$
    $\frac{\partial}{\partial x_2} = -3 + 20(-1) - 3 = -26$
  - Atualizar x:
    $x_1^{Novo} = -3 - 0.085 \times (-12) = -3 + 1.02 \approx -1.98$
    $x_2^{Novo} = -1 - 0.085 \times (-26) = -1 + 2.21 \approx 1.21$
- Passo 2:
  - Calcular o gradiente em $x_1 = [-1.98, 1.21]$:
    $\frac{\partial}{\partial x_1} = 2(-1.98) + 1.21 - 5 = -8.75$
    $\frac{\partial}{\partial x_2} = -1.98 + 20(1.21) - 3 = 19.82$
  - Atualizar x:
    $x_1^{Novo} = -1.98 - 0.085 \times (-8.75) = -1.98 + 0.74375 \approx -1.23625$
    $x_2^{Novo} = 1.21 - 0.085 \times (19.82) = 1.21 - 1.6847 \approx -0.4747$

Exemplo

- Execução de três passos do método do gradiente descendente para uma função de duas variáveis.

Configuração Inicial:

- Função:
  - ◇ $f(x_1,x_2) = x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$
- Gradiente da função:
  - ◇ $\nabla f(x_1,x_2) = \begin{bmatrix} 2x_1 + x_2 - 5 \\ x_1 + 20x_2 - 3 \end{bmatrix}$
- Ponto Inicial:
  - ◇ $x_0 = [-3,-1]$
- Taxa de Aprendizagem:
  - ◇ $\gamma = 0.085$

- Passo 3:
  - ◇ Calcular o gradiente em $x_2 = [-1.23625, -0.4747]$:
    $\frac{\partial}{\partial x_1} = 2(-1.2362) - 0.4747 - 5 = -7.9472$
    $\frac{\partial}{\partial x_2} = -1.23625 + 20(-0.4747) - 3 = -11.73$
  - ◇ Atualizar x:
    $x_1^{Novo} = -1.23625 - 0.085 \times (-7.9472) = -1.236 + 0.675 \approx -0.561$
    $x_2^{Novo} = -0.4747 - 0.085 \times (-11.73) = 0.4747 + 0.997 \approx 1.4717$

- Passo $4, 5, \ldots, n$:
  - ◇ Repetindo o mesmo procedimento. O método deve continuar até que uma condição de parada seja atingida.

Exemplo:

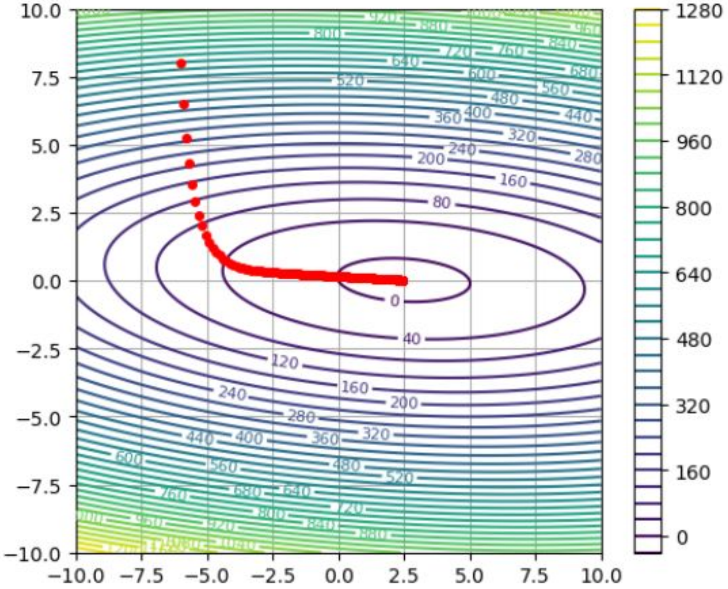
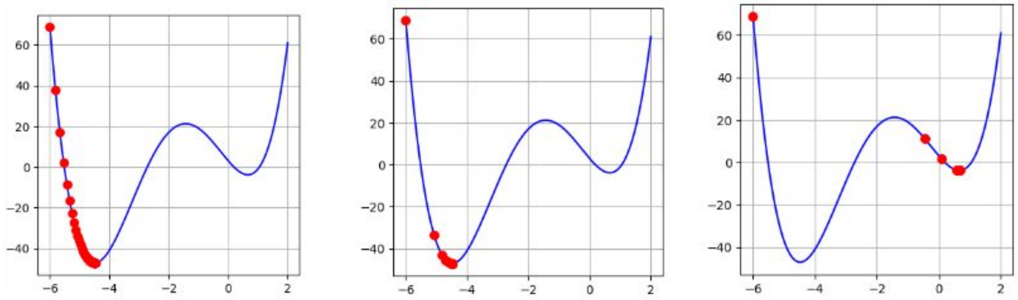


Figura 7.5: Exemplo do método do gradiente descendente para a função $f(x_1,x_2) = x_1^2 + x_1x_2 + 10x_2^2 - 5x_1 - 3x_2$. O valor inicial foi definido como $x_0 = [-7, 7.5]$, seguindo a direção do gradiente negativo na busca pelo ponto mínimo, com a condição de parada quando $|\nabla f(x_1,x_2)| < \varepsilon$, onde $\varepsilon = 0.1$ e $\gamma = 0.01$.

Tamanho do Passo

- O termo “tamanho do passo”' é frequentemente usado como sinônimo para “taxa de aprendizagem”.
  - A taxa de aprendizagem determina o tamanho do passo tomado em direção ao mínimo de uma função a cada iteração do algoritmo.



(a) Aproximação do valor do mínimo global após 96 iterações, com $\gamma = 0.001$

(b) Aproximação do valor do mínimo global após 16 iterações, com $\gamma = 0.005$

(c) Divergência do valor do mínimo global após 6 iterações, com $\gamma = 0.03$.

Figura 7.6: Exemplo do método do gradiente descendente para a função $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$. O valor inicial foi definido como $x_0 = -6$, com a condição de parada quando $|\nabla f(x)| < \varepsilon$, onde $\varepsilon = 0.1$ com variações do valor de γ.

Métodos Adaptativos para o Tamanho do Passo

- Os métodos de gradiente adaptativo redimensionam o tamanho do passo a cada iteração, dependendo das propriedades locais da função.
  - De forma geral, existem duas heurísticas simples [Toussain 2012]:

- Redução do tamanho do passo quando o valor da função aumenta.

Procedimento:

- Após uma iteração, deve-se verificar se o valor da função $f(x)$ aumentou.
- Neste caso, cancela-se a última atualização realizada (desfazendo o último passo) para diminuir o tamanho do passo. Isso é feito multiplicando a taxa de aprendizado atual por um fator $\delta < 1$.

- Aumento do tamanho do passo quando o valor da função diminui.

Procedimento:

- Após uma iteração, deve-se verificar se o valor da função $f(x)$ diminuiu.
- Neste caso, tenta-se aumentar o tamanho do passo. Isso pode ser feito multiplicando a taxa de aprendizado atual por um fator $\delta' > 1$.

Gradiente Descendente com *Momentum*

- O gradiente descendente com *momentum* é uma extensão da técnica tradicional de descida do gradiente.
  - O componente de momento é um termo extra que captura a essência do que ocorreu na iteração anterior.

- O conceito fundamental do *momentum* é simular o comportamento de uma partícula se movendo em uma superfície de erro com alguma inércia.
  - Em termos físicos, isso significa que a partícula (ou os parâmetros do modelo) não somente reage ao gradiente (força) local mas também mantém uma direção e velocidade consistentes baseadas em seu movimento anterior.

Gradiente Descendente com *Momentum*

- A implementação clássica do *momentum* incorpora uma fração do gradiente anterior.

$$m = \beta m - \gamma \nabla f(x_i),$$

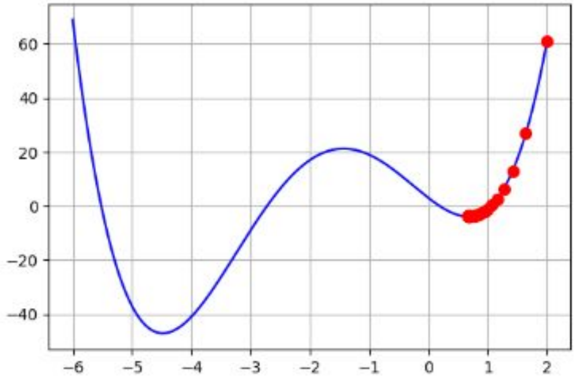
onde m é o vetor de momentum e β é um fator de decaimento que determina o quanto do movimento anterior será retido em comparação ao novo gradiente.

- A atualização do ponto x_{i+1} é caracterizado pelo valor de x_i adicionado ao vetor de *momentum*, que é definido por.

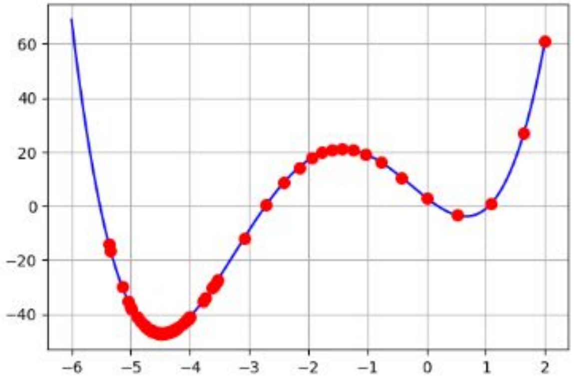
$$x_{i+1} = x_i + m,$$

onde o vetor de momentum contém uma combinação do gradiente atual (direção imediata de descida mais íngreme) e uma fração do vetor de *momentum* anterior, permitindo que o algoritmo mantenha a direção geral de movimento anterior enquanto ainda responde ao gradiente local.

Exemplo



(a) Implementação sem *momentum*, onde o algoritmo atinge um mínimo local após 41 iterações.



(b) Implementação com *momentum*, onde o algoritmo atinge uma aproximação do valor do mínimo global após 83 iterações.

Figura 7.7: Exemplo do método do gradiente descendente para a função $f(x) = x^4 + 7x^3 + 5x^2 - 17x + 3$. O valor inicial foi definido como $x_0 = 2$ com a condição de parada quando $|\nabla f(x)| < \varepsilon$, onde $\varepsilon = 0.1$, $\gamma = 0.003$ e $\beta = 0.92$.

Descida do Gradiente Estocástica

- No aprendizado de máquina, dado $n = 1, \ldots, N$ pontos de dados, frequentemente consideramos funções objetivo caracterizadas pela soma das perdas L_n incorridas por cada exemplo n. Em notação matemática, temos a forma

$$L(\theta) = \sum_{n=1}^{N} L_n(\theta),$$

onde θ é o vetor de parâmetros, ou seja, queremos encontrar θ que minimize L.

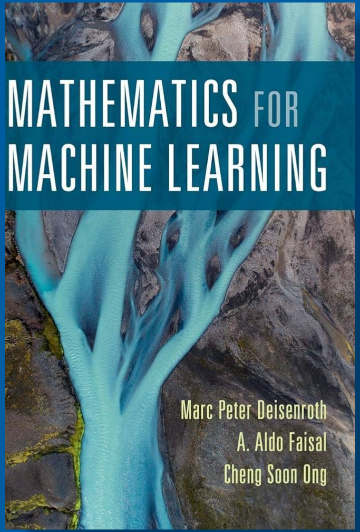
- A descida gradiente padrão, conforme introduzido anteriormente, é um método de otimização em “lote” .
    - A otimização é realizada usando o conjunto de treinamento completo.
    - Especialmente quando o conjunto de dados de treinamento é extenso e/ou as fórmulas para os gradientes não são diretas, o custo associado à avaliação dessas somas de gradientes pode ser proibitivo.

Descida do Gradiente Estocástica

- Em contraste com a descida gradiente em lote, opcionalmente podemos escolher aleatoriamente um subconjunto de L_n para descida gradiente de minilote.
  - A Descida de Gradiente Estocástica o (SGD) calcula o gradiente utilizando um subconjunto aleatoriamente de pontos de dados a cada iteração.
  - O termo ``estocástico" neste contexto significa que não temos conhecimento exato do gradiente, mas sim de uma estimativa aproximada.

- Um dos principais motivos para considerar o uso de gradientes aproximados são as restrições práticas de implementação.
  - Um tamanhos adequado de minilotes consegue fornecer estimativas precisas do gradiente.
  - Abordagens de minilote têm sido amplamente utilizados, sendo eficazes em problemas de aprendizado de máquina em grande escala.

7.2 Otimização Restrita e Multiplicadores de Lagrange



7.2. Otimização Restrita e Multiplicadores de Lagrange

- Continuaremos a analisar o problema de encontrar o mínimo de uma função $f : \mathbb{R}^D \rightarrow \mathbb{R}$, contudo considerando a imposição de **restrições.**

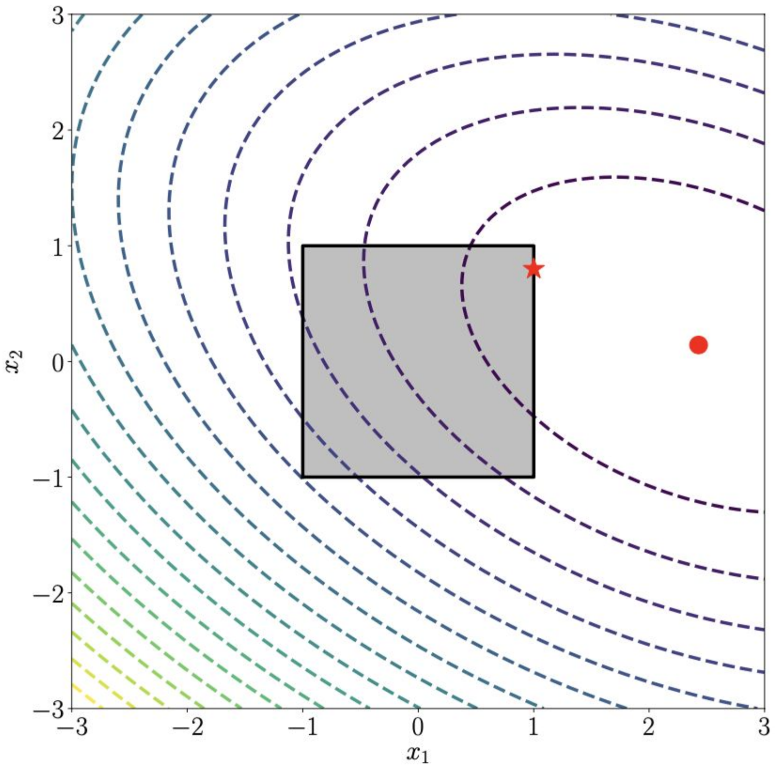
- **Restrições**: funções com valores reais $g_i : \mathbb{R}^D \rightarrow \mathbb{R}$ para $i = 1, \ldots, m$

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{sujeito a} \quad & g_i(x) \leq 0 \quad \text{para todo} \quad i = 1, \ldots, m. \end{aligned}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Problema <u>sem restrições</u>:
  - Linhas de Contorno;
  - Mínimo indicado pelo círculo;

- <u>Restrições</u>:
  - Caixa de restrições ($-1 \leqslant x_1 \leqslant 1$ e $-1 \leqslant x_2 \leqslant 1$);
  - Mínimo indicado pela estrela.



7.2. Otimização Restrita e Multiplicadores de Lagrange

- Uma maneira óbvia, mas não muito prática, de converter o problema restrito em um problema irrestrito é usar uma função indicadora

$$J(\boldsymbol{x}) = f(\boldsymbol{x}) + \sum_{i=1}^{m} \mathbf{1}(g_i(\boldsymbol{x})) ,$$

- onde $\mathbf{1}(z)$ é uma função degrau infinita

$$1(z) = \begin{cases} 0 & \text{se } z \leq 0 \\ \infty & \text{caso contrário} \end{cases} .$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Nestes problemas, utilizaremos os multiplicadores de Lagrange $\lambda_i \geq 0$ correspondentes a cada restrição de desigualdade respectivamente de modo que:

$$\begin{aligned}\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) &= f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{x}) \\ &= f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boldsymbol{g}(\boldsymbol{x})\end{aligned}$$

- Desta forma, substitui-se a função degrau por uma função linear.

7.2. Otimização Restrita e Multiplicadores de Lagrange

- **Problema Dual de Lagrange**

- **Dualidade** na otimização é a ideia de converter um problema de otimização em um conjunto de variáveis x (chamadas de variáveis *primárias*), em outro problema de otimização em um conjunto diferente de variáveis λ (chamadas de variáveis *duais*).

7.2. Otimização Restrita e Multiplicadores de Lagrange

- O problema de encontrar o mínimo da função $f : \mathbb{R}^D \rightarrow \mathbb{R}$ sujeito às restrições $g_i : \mathbb{R}^D \rightarrow \mathbb{R}$ para $i = 1, \ldots, m$, é considerado o **Problema Primordial (Primal)** em relação às **variáveis primárias** x

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{sujeito a} \quad & g_i(x) \leq 0 \quad \text{para todo} \quad i = 1, \ldots, m. \end{aligned}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- O **Problema Dual de Lagrange** relacionado é dado por

$$\begin{aligned} \boxed{\max_{\boldsymbol{\lambda} \in \mathbb{R}^m}} \quad & \mathfrak{D}(\boldsymbol{\lambda}) \\ \text{sujeito a} \quad & \boldsymbol{\lambda} \geqslant \mathbf{0}\,, \end{aligned}$$

- onde $\boldsymbol{\lambda}$ são as **variáveis duais** e $\mathfrak{D}(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$.

$$\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = f(\boldsymbol{x}) + \boldsymbol{\lambda}^\top \boxed{\boldsymbol{g}(\boldsymbol{x})}$$

$$g_i(x) \leq 0$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Considere uma solução viável $\tilde{x}$ para o Problema Primordial (PP)

$$\mathfrak{D}(\boldsymbol{\lambda}) = \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda})$$

$$\Downarrow$$

$$\mathfrak{L}(\tilde{\boldsymbol{x}}, \boldsymbol{\lambda}) = f(\tilde{\boldsymbol{x}}) + \sum_{i=1}^{m} \overbrace{\boxed{\lambda_i}}^{\lambda_i \geq 0} \underbrace{\boxed{g_i(\tilde{\boldsymbol{x}})}}_{g_i(x) \leq 0} \leqslant f(\tilde{\boldsymbol{x}})$$

- Consequentemente, considerando uma solução ótima p^* para o PP, temos que

$$\mathfrak{D}(\boldsymbol{\lambda}) \leqslant p^*$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Desta forma, temos que a função Dual de Lagrange funciona como um limite inferior para o valor ótimo de PP

$$\mathfrak{D}(\boldsymbol{\lambda}) \leqslant p^*$$

- Logo, nosso objetivo será encontrar o melhor limite inferior

$$\begin{array}{ll} \max\limits_{\boldsymbol{\lambda} \in \mathbb{R}^m} & \mathfrak{D}(\boldsymbol{\lambda}) \\ \text{sujeito a} & \boldsymbol{\lambda} \geqslant \mathbf{0}\,, \end{array}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- É importante conhecer a **Desigualdade** *minimax*, que diz que para qualquer função com dois argumentos $\varphi(x, y)$, o *maximin* é menor que o *minimax*, ou seja,

$$\max_{\boldsymbol{y}} \min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \min_{\boldsymbol{x}} \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y})$$

- Esta desigualdade pode ser provada considerando a desigualdade:

Para todo $\boldsymbol{x}, \boldsymbol{y}$ $\quad \min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y})$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Portanto, considere os pontos x_0 e y_0

$$\min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \varphi(x_0, y_0) \leqslant \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y})$$

$$\Downarrow$$

$$\min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y})$$

$$\Downarrow$$

$$\min_{\boldsymbol{x}} \varphi(\boldsymbol{x}, \boldsymbol{y}) \leqslant \quad \max_{\boldsymbol{y}} \varphi(\boldsymbol{x}, \boldsymbol{y})$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Além disso, é importante conhecer o conceito da **Dualidade Fraca**, para mostrar que os valores **primordiais** são sempre maiores ou iguais aos valores **duais**.

- Problema primordial: $J(\boldsymbol{x}) = \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \Rightarrow \min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$

- Portanto: $\min_{\boldsymbol{x} \in \mathbb{R}^d} \max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) \geqslant \boxed{\max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})}$

$$\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} \quad \mathfrak{D}(\boldsymbol{\lambda})$$
$$\text{sujeito a} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}\,,$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

$$\boxed{\max_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})}$$

- Ao contrário do problema de otimização original, que possuía restrições, $\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ é um problema de otimização irrestrito para um determinado valor de λ;

- Se resolver $\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ for fácil, então o problema geral será fácil de se resolver;

- Além disso, sabendo que $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ é afim em relação a λ, podemos afirmar que $\min_{\boldsymbol{x} \in \mathbb{R}^d} \mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ é um mínimo pontual de funções afins de λ;

7.2. Otimização Restrita e Multiplicadores de Lagrange

- $\mathfrak{D}(\boldsymbol{\lambda})$ é côncavo mesmo que $f(\cdot)$ e $g_i(\cdot)$ possam ser não-convexos;
- Logo, o problema externo, maximização sobre λ, é o máximo de uma função côncava e pode ser calculado com eficiência.
- Assumindo que $f(\cdot)$ e $g_i(\cdot)$ são diferenciáveis, encontramos o problema dual de Lagrange diferenciando o Lagrangiano em relação a x, definindo o diferencial como zero e resolvendo o valor ideal.

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Considere ainda se houvesse restrições adicionais de igualdade

$$\begin{array}{llll} \min\limits_{x} & f(x) & & \\ \text{sujeito a} & g_i(x) \leq 0 & \text{para todo} & i = 1, \ldots, m \\ \text{sujeito a} & h_j(x) = 0 & \text{para todo} & j = 1, \ldots, n. \end{array}$$

- É possível modelar restrições de igualdade utilizando duas restrições de desigualdade. Ou seja, cada restrição de igualdade $h_j(x) = 0$, seria substituída equivalentemente por duas restrições $h_j(x) \leqslant 0$ e $h_j(x) \geqslant 0$.

Aplicação do Método

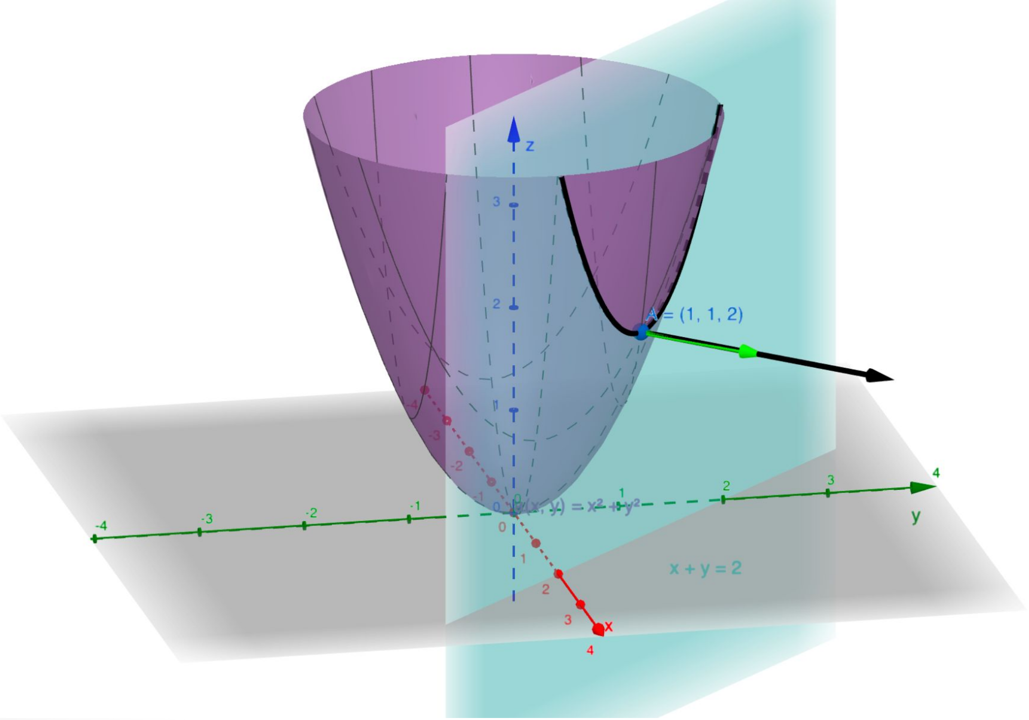
7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 1: minimizar a função $f(x,y) = x^2 + y^2$ considerando a restrição $g(x,y) = x + y = 2$
    - Análise Geométrica
    - Solução Algébrica
    - Solução Computacional

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 1: minimizar a função $f(x,y) = x^2 + y^2$ considerando a restrição $g(x,y) = x + y = 2$

- Análise Geométrica https://www.geogebra.org/calculator/ty5ysps4



7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 1: minimizar a função $\mathbf{f(x,y) = x^2 + y^2}$ considerando a restrição $\mathbf{g(x,y) = x + y = 2}$

$$\nabla f(x,y) = \lambda \, . \, \nabla g(x,y)$$

- Sendo λ é o multiplicador de Lagrange. Portanto, teremos que resolver um sistema com as seguintes equações:

$$\begin{cases} f_x = \lambda \, . \, g_x \\ f_y = \lambda \, . \, g_y \\ g(x,y) = k \end{cases}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 1: minimizar a função $\mathbf{f(x,y) = x^2 + y^2}$ considerando a restrição $\mathbf{g(x,y) = x + y = 2}$

$$\nabla f = (2x, 2y)$$
$$\nabla g = (1, 1)$$

$$\nabla f(x,y) = \lambda \,.\, \nabla g(x,y)$$
$$(2x, 2y) = \lambda \,.\, (1, 1)$$

$$\begin{cases} 2x = \lambda \,.\, 1 \\ 2y = \lambda \,.\, 1 \\ \boxed{x + y = 2} \end{cases} \Rightarrow \begin{matrix} x = \lambda/2 \\ y = \lambda/2 \\ x = y \end{matrix} \Rightarrow \begin{matrix} x + y = 2 \\ \lambda/2 + \lambda/2 = 2 \\ \lambda = 2 \end{matrix} \Rightarrow \begin{matrix} x = y = 1 \\ p_1 = (1, 1) \\ \boxed{f(1,1) = 2} \end{matrix}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 1: minimizar a função $\mathbf{f(x,y) = x^2 + y^2}$ considerando a restrição $\mathbf{g(x,y) = x + y = 2}$

  - Solução Computacional

    https://drive.google.com/file/d/1znSCDKThEP4-qIn2VS6yBc2krQw_tYjK/view?usp=sharing

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 2: minimizar a função $\mathbf{f(x,y) = x^2 + y}$ no disco $\mathbf{x^2 + y^2 \leq 9}$

Fronteira: $x^2 + y^2 = 9 \rightarrow g(x,y) = x^2 + y^2 - 9$

$$\nabla f = (2x, 1)$$
$$\nabla g = (2x, 2y)$$

$\Rightarrow$ $x^2 + y^2 = 9$, $0 + y^2 = 9$, $y = \pm 3$ $\Rightarrow$ $\boxed{p_1=(0, 3) \quad p_2=(0, -3)}$

$\Downarrow$

$$(2x, 1) = \lambda . (2x, 2y)$$
$$(2.0, 1) = \lambda . (2.0, 2.3)$$
$$(0, 1) = \lambda . (0, 6)$$
$$\lambda = 1/6$$

$$\nabla f(x,y) = \lambda . \boxed{\nabla g(x,y)}$$
$$(2x, 1) = \lambda . (2x, 2y)$$

$$\begin{cases} 2x = \lambda . 2x \\ 1 = \lambda . 2y \\ \boxed{x^2 + y^2 = 9} \end{cases} \Rightarrow \begin{array}{l} 2x = \lambda . 2x \\ x = 0 \mid \lambda = 1 \end{array}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 2: minimizar a função $\mathbf{f(x,y) = x^2 + y}$ no disco $\mathbf{x^2 + y^2 \leq 9}$

Fronteira: $x^2 + y^2 = 9 \rightarrow g(x,y) = x^2 + y^2 - 9$

$$\nabla f = (2x, 1)$$
$$\nabla g = (2x, 2y)$$

$$x^2 + y^2 = 9$$
$$0 + y^2 = 9$$
$$y = \pm 3$$

$$\boxed{\begin{array}{l} p_1 = (0, 3) \\ p_2 = (0, -3) \end{array}}$$

$$\nabla f(x,y) = \lambda \,.\, \nabla g(x,y)$$
$$(2x, 1) = \lambda \,.\, (2x, 2y)$$

$$\begin{cases} 2x = \lambda \,.\, 2x \\ \boxed{1 = \lambda \,.\, 2y} \\ \boxed{x^2 + y^2 = 9} \end{cases}$$

$$2x = \lambda \,.\, 2x$$
$$x = 0 \mid \lambda = 1$$

$$1 = \lambda \,.\, 2y$$
$$1 = 2y$$
$$y = ½$$

$$x^2 + y^2 = 9$$
$$x^2 + (½)^2 = 9$$
$$x = \pm\sqrt{35}/2$$

$$\boxed{\begin{array}{l} p_3 = (\sqrt{35}/2, ½) \\ p_4 = (-\sqrt{35}/2, ½) \end{array}}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 2: minimizar a função $\boxed{f(x,y) = x^2 + y}$ no disco $x^2 + y^2 \leq 9$

  Fronteira: $x^2 + y^2 = 9 \rightarrow g(x,y) = x^2 + y^2 - 9$

  - $p_1=(0, 3) \rightarrow f(p_1) = 0^2 + 3 = 3$
  - $p_2=(0, -3) \rightarrow f(p_2) = 0^2 - 3 = -3$ mínimo
  - $p_3=(\sqrt{35}/2, ½) \rightarrow f(p_3) = 37/4 = 9.25$
  - $p_4=(-\sqrt{35}/2, ½) \rightarrow f(p_4) = 37/4 = 9.25$ máximo

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 2: minimizar a função $\boxed{\mathbf{f(x,y) = x^2 + y}}$ no disco $\mathbf{x^2 + y^2 \leq 9}$

Dentro: $x^2 + y^2 \leq 9$

$$\nabla f = (2x, 1)$$

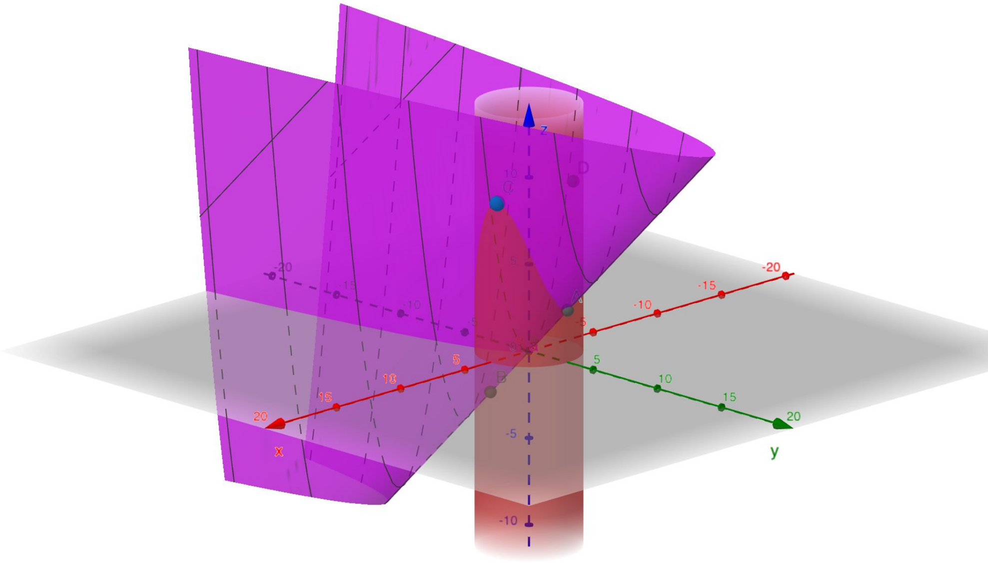
$$\begin{cases} 2x = 0 \\ \cancel{1 = 0} \end{cases}$$

$\boxed{p_2 = (0, -3) \rightarrow f(p_2) = -3}$ mínimo

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 2: minimizar a função $f(x,y) = x^2 + y$ no disco $x^2 + y^2 \leq 9$

- Análise Geométrica https://www.geogebra.org/calculator/gwxqj7na



7.2. Otimização Restrita e Multiplicadores de Lagrange

- <u>Exemplo 3</u>: minimizar a função $\mathbf{f(x,y) = x + 2y}$ sujeita às <u>restrições</u> $\mathbf{x + y + z = 1}$ e $\mathbf{y^2 + z^2 = 4}$

$$\nabla f(x,y) = \lambda \,.\, \nabla g(x,y) + \mu \,.\, \nabla h(x,y)$$

$$\begin{cases} f_x = \lambda \,.\, g_x + \mu \,.\, h_x \\ f_y = \lambda \,.\, g_y + \mu \,.\, h_y \\ g(x,y) = k \\ h(x,y) = m \end{cases}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 3: minimizar a função $\mathbf{f(x,y) = x + 2y}$ sujeita às restrições $\mathbf{x + y + z = 1}$ e $\mathbf{y^2 + z^2 = 4}$

$$g(x, y, z) = x + y + z - 1$$
$$h(x, y, z) = y^2 + z^2 - 4$$
$$f(x, y, z) = x + 2y$$

$$\nabla f = (1,2,0)$$
$$\nabla g = (1,1,1)$$
$$\nabla h = (0, 2y, 2z)$$

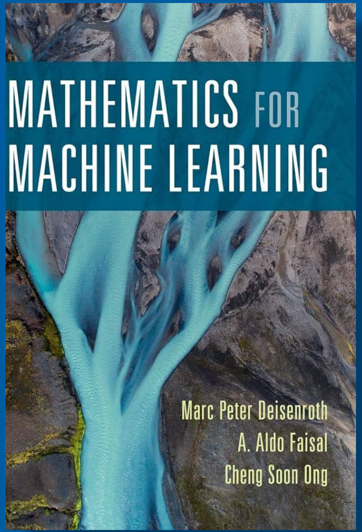
$$\begin{cases} 1 = \lambda \,.\, 1 + \mu \,.\, 0 \\ 2 = \lambda \,.\, 1 + \mu \,.\, 2y \\ 0 = \lambda \,.\, 1 + \mu \,.\, 2z \\ \boxed{x + y + z = 1} \\ \boxed{y^2 + z^2 = 4} \end{cases} \Rightarrow \begin{array}{l} \lambda = 1 \\ \mu \,.\, 2y = 1 \rightarrow \mu = 1/2y \\ \mu \,.\, 2z = -1 \rightarrow \mu = -1/2z \\ 1/2y = -1/2z \rightarrow z = -y \\ \rightarrow z^2 = y^2 \end{array} \Rightarrow \begin{array}{l} y^2 + z^2 = 4 \\ y^2 + y^2 = 4 \\ \boxed{\begin{array}{l} y = \pm\sqrt{2} \\ z = \pm\sqrt{2} \end{array}} \end{array} \Rightarrow \begin{array}{l} x + y + z = 1 \\ x + y - y = 1 \\ \boxed{x = 1} \end{array}$$

7.2. Otimização Restrita e Multiplicadores de Lagrange

- Exemplo 3: minimizar a função $\boxed{\mathbf{f(x,y) = x + 2y}}$ sujeita às restrições $\mathbf{x + y + z = 1}$ e $\mathbf{y^2 + z^2 = 4}$

    - $\boxed{p_1 = (1, \sqrt{2}, -\sqrt{2}) \rightarrow f(p_1) = 1 + 2\sqrt{2}}$ máximo
    - $\boxed{p_2 = (1, -\sqrt{2}, \sqrt{2}) \rightarrow f(p_2) = 1 - 2\sqrt{2}}$ mínimo

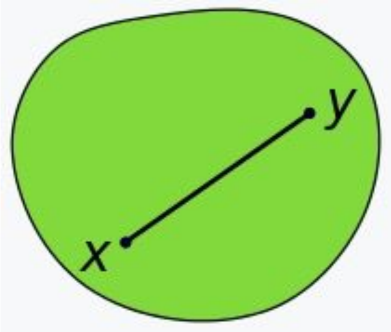
7.3 Otimização Convexa



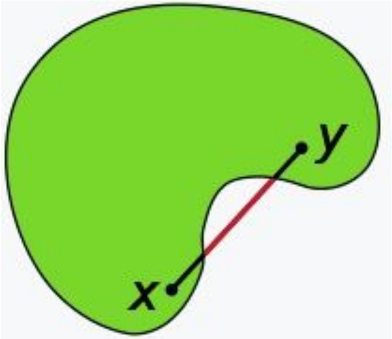
7.3 Otimização Convexa

Um conjunto C é convexo se para qualquer $x, y \in C$ e algum θ com $0 \leq \theta \leq 1$ nós temos

$$\theta x + (1-\theta) y \in C$$



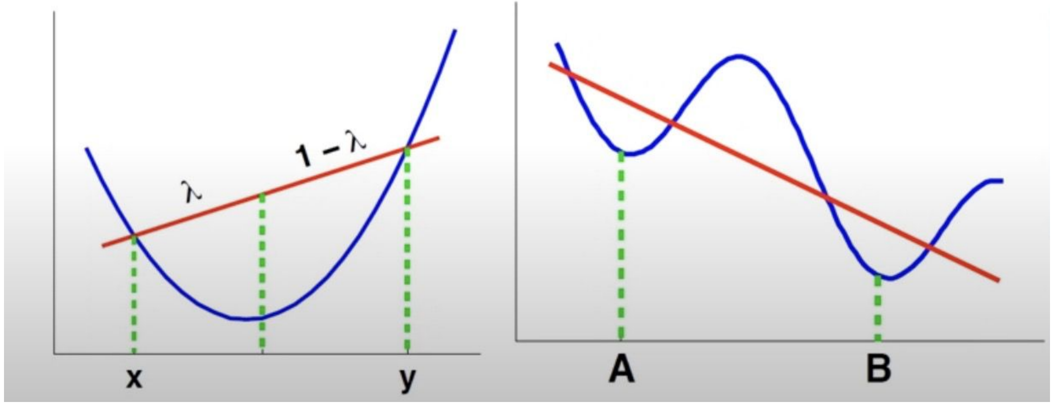
Conjunto convexo



Conjunto não convexo

Uma função $f : R^n \longrightarrow R$ é convexo se $\mathbf{dom}\, f$ é um conjunto convexo e se para todo $x, y \in \mathbf{dom}\, f$, e $0 \leq \theta \leq 1$, nós temos

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta) f(y)$$



Função convexa

Função não convexa

Observação: Uma função côncava é uma função convexa negativa.

Em resumo, um problema de otimização restrita é chamado de problema de otimização convexa se

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le 0, \quad i = 1, \ldots, m \\ & h_i(x) = 0, \quad i = 1, \ldots, p \end{array}$$

onde $f_0(x)$ e $f_i(x)$ são funções convexas e $h_i(x) = 0$ são conjuntos convexos e $x \in R^n$ são as variáveis de otimização e $f_0(x)$ é a função objetivo.

7.3.1 Programação Linear

Considere o caso especial quando todas as funções anteriores são lineares, tal que

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}} & c^T x \\ \text{subject to} & Ax \leq b \end{array}$$

Onde $A \in \mathbb{R}^{m \times d}$ e $b \in \mathbb{R}^m$. Conhecido como uma programação linear. Possui *d* variáveis e *m* restrições lineares. O Lagrangiano é dado por

$$\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \boldsymbol{c}^\top \boldsymbol{x} + \boldsymbol{\lambda}^\top (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$$

Derivando em relação a **x**

$$c + A^T \lambda = 0$$

Resultando o seguinte problema de otimização dual

$$\begin{aligned} \max_{\lambda \in \mathbb{R}^m} \quad & -b^T \lambda \\ \text{subject to} \quad & c + A^T \lambda = 0 \\ & \lambda \geq 0. \end{aligned}$$

Exemplo Programação Linear

$$\min_{x \in \mathbb{R}^2} \quad c^T x = 3x_1 + 2x_2$$

$$\text{subject to} \quad \begin{cases} x_1 + x_2 \geq 4 \\ x_1 - x_2 \geq 1 \\ x_1, x_2 \geq 0 \end{cases}$$

Formulando o Lagrangiano: Reescrevendo as restrições na forma $Ax \leq b$:

$$\begin{cases} -x_1 - x_2 \leq -4 \\ -x_1 + x_2 \leq -1 \\ x_1, x_2 \geq 0 \end{cases}$$

Então, $A = \begin{pmatrix} -1 & -1 \\ -1 & 1 \end{pmatrix}$ e $b = \begin{pmatrix} -4 \\ -1 \end{pmatrix}$.

O Lagrangiano é dado por:

$$\mathfrak{L}(x,\lambda) = 3x_1 + 2x_2 + \lambda_1(-x_1 - x_2 + 4) + \lambda_2(-x_1 + x_2 + 1)$$

Simplificando o Lagrangiano

Expandindo e agrupando termos:

$$\mathfrak{L}(x,\lambda) = (3 - \lambda_1 - \lambda_2)x_1 + (2 - \lambda_1 + \lambda_2)x_2 + 4\lambda_1 + \lambda_2$$

Condições de Otimalidade : Derivando em relação a x_1 e x_2 e igualando a zero:

$$\frac{\partial \mathfrak{L}}{\partial x_1} = 3 - \lambda_1 - \lambda_2 = 0 \implies \lambda_1 + \lambda_2 = 3$$

$$\frac{\partial \mathfrak{L}}{\partial x_2} = 2 - \lambda_1 + \lambda_2 = 0 \implies \lambda_1 - \lambda_2 = 2$$

Resolvendo este sistema de equações:

$$\begin{cases} \lambda_1 + \lambda_2 = 3 \\ \lambda_1 - \lambda_2 = 2 \end{cases}$$

Somando as equações:

$$2\lambda_1 = 5 \implies \lambda_1 = \frac{5}{2}$$

Substituindo λ_1 na primeira equação:

$$\frac{5}{2} + \lambda_2 = 3 \implies \lambda_2 = 3 - \frac{5}{2} = \frac{1}{2}$$

Problema Dual

A função dual é:

$$\mathfrak{D}(\lambda) = -\lambda_1 b_1 - \lambda_2 b_2 = -\lambda_1(-4) - \lambda_2(-1) = 4\lambda_1 + \lambda_2$$

Substituindo λ_1 e λ_2:

$$\mathfrak{D}\left(\frac{5}{2},\frac{1}{2}\right) = 4\cdot\frac{5}{2}+\frac{1}{2} = 10+\frac{1}{2} = 10.5$$

O problema dual se torna:

$$\max_{\lambda_1,\lambda_2\geq 0} \quad 4\lambda_1+\lambda_2$$
$$\text{subject to} \quad \begin{cases} \lambda_1+\lambda_2=3 \\ \lambda_1-\lambda_2=2 \end{cases}$$

$$\min_{x\in\mathbb{R}^2} \quad c^T x = 3x_1+2x_2$$
$$\text{subject to} \quad \begin{cases} x_1+x_2\geq 4 \\ x_1-x_2\geq 1 \\ x_1,x_2\geq 0 \end{cases}$$

$$\max_{\lambda_1,\lambda_2\geq 0} \quad 4\lambda_1+\lambda_2$$
$$\text{subject to} \quad \begin{cases} \lambda_1+\lambda_2=3 \\ \lambda_1-\lambda_2=2 \end{cases}$$

7.3.2 Programação Quadrática

Considere o caso de uma função objetivo quadrática convexa:

$$\min_{x \in \mathbb{R}^d} \quad \frac{1}{2} x^T Q x + c^T x$$
$$\text{subject to} \quad Ax \leq b$$

O Lagrangiano é dado por:

$$\begin{aligned}\mathfrak{L}(x, \lambda) &= \frac{1}{2} x^T Q x + c^T x + \lambda^T (Ax - b) \\ &= \frac{1}{2} x^T Q x + (c + A^T \lambda)^T x - \lambda^T b\end{aligned}$$

Rearranjando os termos, derivando em relação a x e igualando a zero:

$$Qx + (c + A^T \lambda) = 0 \longrightarrow x = -Q^{-1} + (c + A^T \lambda)$$

Substituindo o valor de x no Lagrangiano primal, obtemos o Lagrangiano dual:

$$\mathfrak{D}(\lambda) = -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda)^T - \lambda^T b$$

Logo o problema de otimização dual é dado por:

$$\begin{array}{ll} \max_{x \in \mathbb{R}^m} & -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda)^T - \lambda^T b \\ \text{subject to} & \lambda \geq 0 \end{array}$$

Exemplo Programação Quadrática

$$\min_{x \in \mathbb{R}^2} \quad \frac{1}{2} x^T \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} x + \begin{pmatrix} -2 \\ -5 \end{pmatrix}^T x$$

$$\text{subject to} \quad \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix} x \leq \begin{pmatrix} 2 \\ 2 \end{pmatrix}$$

O Lagrangiano para este problema é:

$$\mathfrak{L}(x, \lambda) = \frac{1}{2} x^T \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} x + \begin{pmatrix} -2 \\ -5 \end{pmatrix}^T x + \lambda^T \left(\begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix} x - \begin{pmatrix} 2 \\ 2 \end{pmatrix} \right)$$

Rearranjando os termos, temos:

$$\mathfrak{L}(x,\lambda) = \frac{1}{2}(2x_1^2 + 2x_2^2) + (-2+\lambda_1-\lambda_2)x_1 + (-5+2\lambda_1+2\lambda_2)x_2 - 2\lambda_1 - 2\lambda_2$$

Derivando em relação a x e igualando a zero:

$$\frac{\partial \mathfrak{L}}{\partial x_1} = 2x_1 + (-2+\lambda_1-\lambda_2) = 0$$
$$\frac{\partial \mathfrak{L}}{\partial x_2} = 2x_2 + (-5+2\lambda_1+2\lambda_2) = 0$$

Isolando x:

$$x_1 = 1 - \frac{1}{2}\lambda_1 + \frac{1}{2}\lambda_2$$
$$x_2 = \frac{5}{2} - \lambda_1 - \lambda_2$$

Substituindo no Lagrangiano primal, obtemos o dual:

$$\mathfrak{D}(\lambda) = -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b$$

Portanto, o problema dual é:

$$\begin{array}{ll} \max\limits_{\lambda \in \mathbb{R}^m} & -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b \\ \text{subject to} & \lambda \geq 0 \end{array}$$

$$\begin{array}{ll} \min\limits_{x \in \mathbb{R}^2} & \frac{1}{2}x^T \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} x + \begin{pmatrix} -2 \\ -5 \end{pmatrix}^T x \\ \text{subject to} & \begin{pmatrix} 1 & 2 \\ -1 & 2 \end{pmatrix} x \leq \begin{pmatrix} 2 \\ 2 \end{pmatrix} \end{array} \longrightarrow \begin{array}{ll} \max\limits_{\lambda \in \mathbb{R}^m} & -\frac{1}{2}(c + A^T\lambda)^T Q^{-1}(c + A^T\lambda) - \lambda^T b \\ \text{subject to} & \lambda \geq 0 \end{array}$$

7.3.3 Transformada de Legendre-Fenchel e conjugado convexo

- A transformação de Legendre-Fenchel (transformação convexa) essencialmente troca as variáveis de uma função convexa, de modo que o novo domínio seja descrito em termos das variáveis duais (conjugadas);

- Em problemas de otimização convexa temos uma forte dualidade, ou seja, as soluções do problema primal e dual são as mesmas;

- O conjugado de Legendre-Fenchel revela-se bastante útil para problemas de aprendizagem de máquina.

Conclusão

- A otimização contínua é uma área de pesquisa ativa;

- **Gradiente descendente** é amplamente utilizado para otimização, especialmente em aprendizado de máquina e redes neurais, apesar de suas limitações;

- **Dualidade e otimização convexa** são conceitos fundamentais no campo da otimização contínua:
    - <u>Dualidade</u>: limites, simplificação do problema;
    - <u>Propriedades de Problemas Convexos</u>: ótimo global, eficiência computacional.

Material Disponibilizado

- Página no GitHub:
    - Relatório (pdf)
    - Videoaula (YouTube)
    - Slides (pdf)
    - Códigos (Notebooks Google Colab - Python)

Contatos

André Rodrigues Coimbra
andre_coimbra@discente.ufg.br

Rayane Araujo Lima
rayane_lima@discente.ufg.br

Renan Rodrigues de Oliveira
renanrodrigues@discente.ufg.br

Obrigado(a)!