# Programação Não-Linear com Restrições

Prof. Alexandre Salles da Cunha

Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Belo Horizonte, Brasil

acunha@dcc.ufmg.br

2022/1

UNIVERSIDADE FEDERAL DE MINAS GERAIS

UFMG - ICEx
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

1. Métodos primais:
    - Método de Direções Viáveis (Gradiente Condicional)
    - Método do Gradiente Projetado
    - Coordenadas ativas
2. Dualidade Lagrangeana
3. Condições de Slater
4. Condições de Karush-Kuhn-Tucker
5. Métodos de Penalidades
6. Método de Barreiras
7. Método do Lagrangeano Aumentado

## Problema de otimização sobre um conjunto convexo

$$\min f_0(x) : x \in X$$

- Vamos assumir que $X \neq \emptyset$ é um conjunto convexo.
- O conjunto $X$ pode ser definido por $h(x) = 0$ e $f(x) \leq 0$, onde:
  - $h : \mathbb{R}^n \to \mathbb{R}^q$
  - $f : \mathbb{R}^n \to \mathbb{R}^m$
- Sendo convexo, as restrições $h(x) = 0$ são necessariamente restrições afim.

Supondo que as restrições $h(x) = 0$ sejam o conjunto de $q$ restrições afim $Ax = b$ (posto$(A) = q$) podemos encontrar uma matriz $N \in \mathbb{R}^{n \times (n-q)}$ que forneça uma base para $\mathcal{N}(A)$. Fatoramos (SVD) $A = U\Sigma V^T$, e tomamos $N$ como as últimas $n - q$ colunas ortonormais de $V$.

- Substituimos $x = \overline{x} + Nz$, $\overline{x}$ é uma solução qualquer de $Ax = b$ e $z \in \mathbb{R}^{n-q}$ é um novo vetor de variáveis.
- Eliminamos as restrições $h(x) = Ax - b = 0$
- Re-escrevemos $f_i(x) : i = 0, \ldots, m$ como $f_i(\overline{x} + Nz) : i = 0, \ldots, m$.

## Vantagens e desvantagens

- Redução de $n$ para $n - q$ variáveis, de $m + q$ para $m$ restrições.
- Se $A$ for esparsa e $N$ for densa, a reformulação pode não ser vantajosa.

**Definição**

Um ponto $x^* \in X$ que satisfaz a condição

$$\nabla f_0(x^*)^T (x - x^*) \geq 0, \forall x \in X$$

é denominado ponto estacionário.

- Observe que se $x^*$ é um ponto interior ou se $X = \mathbb{R}^n$ as condições enunciadas equivalem à $\nabla f_0(x^*) = 0$.
- Assim sendo, estas condições também podem ser satisfeitas para pontos que não sejam de mínimo.

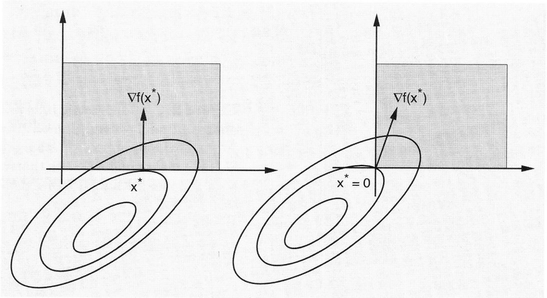
Para um ponto $\overline{x} \in X$ (viável):

- Todas as restrições $h(x) = 0$ são ativas em $x = \overline{x}$.
- Dizemos que $f_j(x) \leq 0$ é ativa em $\overline{x}$ se $f_j(\overline{x}) = 0$. Se não é ativa, é inativa.

## Exemplo - otimização no quadrante não negativo

$$\min f_0(x_1, x_2) : \quad x_1, x_2 \geq 0$$

No ponto $x^*$ de mínimo,
$\nabla f_0(x^*)^T(x - x^*) = \sum_{i=1}^{n} \frac{\partial f_0(x^*)}{\partial x_i}(x_i - x_i^*) \geq 0$. Isto implica que $\frac{\partial f_0(x^*)}{\partial x_i} \geq 0$ para todo $i$ (basta tomar $x_i = x_i^* + 1$ se $x_i^* = 0$ e $x_j = x_j^*, i \neq j$). Para $i$ tal que $x_i > 0$, $\frac{\partial f_0(x^*)}{\partial x_i} = 0$.



- As restrições ativas em $\overline{x}$ é que restringem a viabilidade nas suas vizinhanças.
- As restrições inativas não influenciam em nada a viabilidade nas vizinhanças de $\overline{x}$.
- Desempenham papel fundamental na dedução das propriedades (condições necessárias e suficientes) de um ponto de mínimo local.
- Caso soubéssemos quais são as restrições ativas no ponto de mínimo, poderíamos concentrar nosso estudo apenas à elas, tratando-as como restrições de igualdade e ignorando as restrições inativas.
- A inclusão de restrições de desigualdade confere *natureza combinatória* ao problema de PNL.

1. Método do Gradiente Condicional
2. Método do Gradiente Projetado
3. Método do Gradiente Reduzido
4. Método do Simplex Convexo

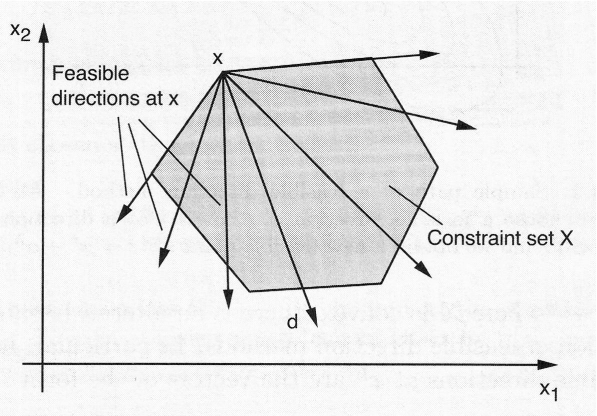
$$\begin{array}{ll} \text{minimize} & f_0(x) \\ & x \in X \end{array}$$

Vamos inicialmente estudar métodos de otimização para o problema, assumindo:

- $X \neq \emptyset$ é convexo, não vazio e fechado.
- $f_0 \in C^1$, isto é, $X$ é continuamente diferenciável em $X$.

## Definição

Dado ponto viável $x \in X$, dizemos que $d$ é uma direção viável em $x$ se $d$ é tal que $x + \alpha d \in X$ para qualquer $\alpha > 0$ suficientemente pequeno.



Um método nesta classe:

1. inicia com uma solução $x^0$ e gera uma sequência $\{x^k\}$ de pontos viáveis, através da iteração típica:

$$x^{k+1} = x^k + \alpha^k d^k$$

2. se $x^k$ não é um ponto estacionário, então a direção $d^k$ escolhida deve satisfazer:

$$\nabla f_0(x^k)^T d^k < 0$$

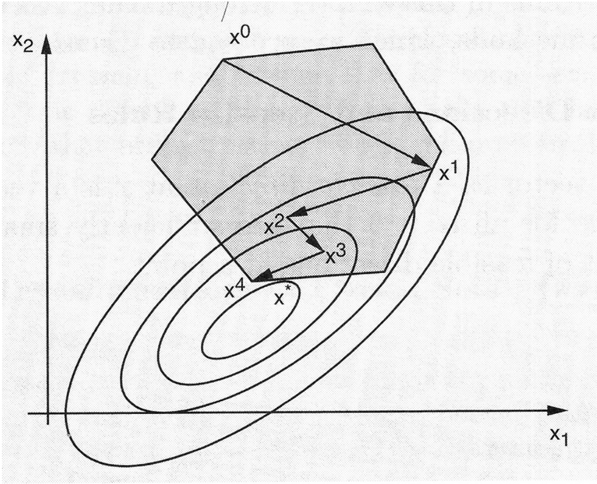
(uma direção viável de descida existe)

3. o passo $\alpha^k$ é escolhido de forma que $x^k + \alpha^k d^k \in X$.
4. se $x^k$ é estacionário, o método pára ($x^{k+1} = x^k$).

(não existe uma direção viável de descida !)

Vamos nos concentrar em métodos de direções viáveis que são métodos de descida, isto é, $f_0(x^k + \alpha^k d^k) < f_0(x^k)$.

# Um método de direções viáveis, de descida



- Quando $X$ é convexo, uma direção viável $d^k$ em $x^k$ é dada por:

$$d^k = \gamma(\overline{x}^k - x^k), \;\; \gamma > 0$$

onde $\overline{x}^k$ é qualquer outro ponto viável.

- Observe que pela convexidade de $X$, $x^k + \alpha^k(\overline{x}^k - x^k) \in X$ para qualquer $\alpha \in [0, 1]$.
- O método opera então de acordo com a seguinte iteração:

$$x^{k+1} = x^k + \alpha^k d^k, \;\; \alpha \in [0, 1],$$

onde a direção $d^k$ além de viável, deve ser de descida:

$$\nabla f_0(x^k)^T(\overline{x}^k - x^k) < 0$$

- Se $d^k = \gamma(\overline{x}^k - x^k)$ é uma direção de descida e $x^k$ é não estacionário, temos a garantia de existir um $\alpha^k$ para o qual $f_0(x^{k+1}) < f_0(x^k)$.

# Métodos de direções viáveis: idéias centrais

1. Requerem uma solução viável inicial (emprego de Fase I)
2. Definem uma direção viável de descida, ou concluem que o ponto é estacionário.
3. Implementam uma iteração típica que corresponde a uma busca unidirecional na direção escolhida.
4. Podem ser vistos como uma especialização dos métodos vistos em otimização irrestrita.

Uma vez determinada a direção $d^k$:

- Minimização exata restrita: $\alpha^k = \text{arg min}_{\alpha \in [0,1]} f_0(x^k + \alpha d^k)$

- Armijo: Fixados escalares $\beta, \sigma > 0$ tais que $\beta \in (0,1)$ e $\sigma \in (0,1)$, fazemos $\alpha^k = \beta^{m_k}$, onde $m_k$ é o primeiro inteiro não negativo $m$ tal que:

$$f_0(x^k) - f_0(x^k + \beta^m d^k) \geq -\sigma \beta^m \nabla f_0(x^k)' d^k$$

  Ou seja, tentamos sucessivamente os passos $\alpha^k = 1, \beta, \beta^2, \ldots$ até satisfazer a condição acima.

Tomando um $x^0$ qualquer, possivelmente inviável para o programa original

$$
\begin{aligned}
(P)\ \min f_0(x) & \\
h_i(x) = 0 & \qquad i = 1, \dots, q \\
g_j(x) \leq 0 & \qquad j = 1, \dots, m
\end{aligned}
$$

Resolvemos o programa auxiliar:

$$
\begin{aligned}
(PA)\ \min \sum_i z_i + \sum_j y_j & \\
h_i(x) = 0 & \qquad i : h(x^0) = 0 \\
h_i(x) + z_i = 0 & \qquad i : h_i(x^0) < 0 \\
h_i(x) - z_i = 0 & \qquad i : h_i(x^0) > 0 \\
g_j(x) \leq 0 & \qquad j : g_j(x^0) \leq 0 \\
g_j(x) - y_j \leq 0 & \qquad j : g_j(x^0) > 0 \\
z_i \geq 0, y_j \geq 0 & \qquad \forall i, j
\end{aligned}
$$

Se ao final da resolução do programa auxiliar a função objetivo ótima satisfizer $\sum_i z_i^* + \sum_j y_j^* = 0$, temos um ponto de partida para aplicar o método ao problema original (P)

$\Rightarrow$ Veja que o mesmo algoritmo é aplicado para resolver PA e então P.

Os métodos de direções viáveis se diferenciam em como a direção viável de descida é obtida. Um dos métodos mais usados na classe é o Gradiente Condicional (Método de Frank-Wolfe).

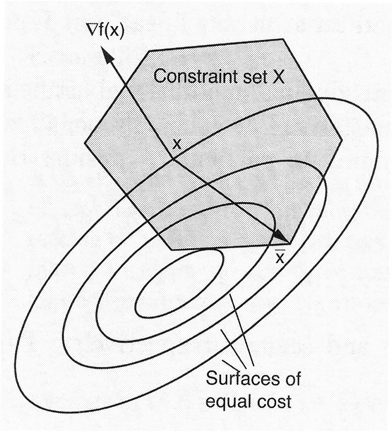
- O problema de obter uma direção viável com propriedade de descida em $x^k$ pode ser formulado como:

$$\begin{aligned} \min \quad & \nabla f_0(x^k)^T(x - x^k) \\ & x \in X \end{aligned}$$

- Se $X$ é compacto, o programa acima é limitado. Caso $X$ não seja limitado, impomos restrições lineares do tipo $\|x\|_\infty \leq 1$ e resolvemos o programa.
- Se o valor ótimo do programa auxiliar for negativo, encontramos uma direção viável, de descida. Caso contrário, $x^k$ é estacionário.

- Observe que o programa auxiliar envolve o mesmo conjunto de restrições do original, mas a função objetivo é linear.
- No caso das restrições que definem $X$ serem não lineares, o programa auxiliar é tão complicado quanto o programa original.
- Os métodos de direções viáveis, em particular o do Gradiente Condicional, são particularmente interessantes quando $X$ é um poliedro, uma vez que o subproblema é um Programa Linear, passível de solução eficiente.

- Se o método Simplex é usado para resolver o subproblema, $\overline{x}^k$ sempre é um ponto extremo de $X$.



## Problema Quadrático

$$\begin{aligned} \text{minimize} \quad f_0(x) = \frac{1}{2}(x_1^2 + x_2^2 + 0.1x_3^2) + 0.55x_3 \\ x_1 + x_2 + x_3 = 1 \\ x \geq 0 \end{aligned}$$

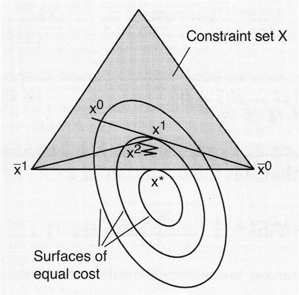
$x^* = (\frac{1}{2} \quad \frac{1}{2} \quad 0)^T, x^0 = \frac{1}{3}(1 \quad 1 \quad 1).$

## Subproblema associado - determinação da direção viável, de descida

$$\begin{aligned} \text{minimize} \quad & (x_1^k)x_1 + (x_2^k)x_2 + (0.1x_3^k + 0.55)x_3 + \\ & + \left(-(x_1^k)^2 - (x_2^k)^2 - (x_3^k)(0.1x_3^k + 0.55)\right) \\ & x_1 + x_2 + x_3 = 1 \\ & x \geq 0 \end{aligned}$$

`gradcondicional.sci`

- Se Armijo ou minimização exata forem empregadas, pode-se demonstrar que, se as direções forem *gradient-related*, todo ponto limite da sequência gerada pelo método é um ponto estacionário.
- A taxa de convergência do método é baixa: os pontos $\overline{x}^k$ gerados são tais que a direção $d^k$ é quase ortogonal à direção que levaria ao ótimo, a partir do ponto $x^k$ (zig-zag). Nos casos patológicos, a convergência é sub-linear, $\beta \to 1$.
- A taxa melhora quando o número de restrições lineares aumenta muito (*falsa* curvatura positiva), muitos pontos extremos próximos.



É uma grande classe de métodos primais que particiona o conjunto de restrições em dois grupos:

1. aquelas que serão tratadas como ativas (working set)
2. e aquelas que serão tratadas como inativas.

- Para o conjunto de restrições ativas escolhido, o método implementa um movimento na superfície de restrições ativas visando redução da função objetivo, obtendo um novo ponto. Então, corrige o conjunto de restrições ativas.
- Diferentes classes de algoritmos são definidos a partir de como é realizada a operação de obtenção do novo ponto, dado o conjunto de restrições ativas associado ao ponto atual.

Vamos ilustrar o método para um problema com restrições de desigualdades apenas. Caso restrições $h(x) = 0$ sejam impostas, o conjunto de restrições ativas sempre irá incluí-las.

$$
\begin{aligned}
\min\ & f_0(x) \\
& f_j(x) \le 0 \qquad\qquad\qquad j = 1, \ldots, m
\end{aligned}
$$

**CNPO para ponto $x$ regular**

$$
\begin{aligned}
\nabla f_0(x) + \nabla f(x)\mu &= 0 \\
f(x) &\le 0 \\
f(x)^T\mu &= 0 \\
\mu &\ge 0
\end{aligned}
$$

$$J(x) = \{j = 1, \ldots, m : f_j(x) = 0\}$$

## CNPO re-escritas em função de $J(x)$

$$\begin{aligned}
\nabla f_0(x) + \sum_{j \in J(x)} \nabla f_j(x)\mu_j &= 0 & & \\
f_j(x) &= 0 & \qquad & j \in J(x) \\
f_j(x) &< 0 & & j \notin J(x) \\
\mu_j &= 0 & & j \notin J(x) \\
\mu_j &\geq 0 & & j \in J(x)
\end{aligned}$$

# Métodos de restrições ativas

1. Se as restrições ativas no working set ótimo fossem conhecidas, resolveríamos:

$$\begin{aligned} \min \ & f_0(x) \\ & f_j(x) = 0 \qquad\qquad j \text{ ativas no working set ótimo} \end{aligned}$$

2. Como não são, escolhemos um conjunto de restrições ativas $J(x) = \{j = 1, \ldots, m : f_j(x) = 0\}$ associado a $x$ viável.
3. Este conjunto $J(x)$ é tratado como um *working set*.
4. Resolvemos o problema abaixo, obtendo $\overline{x}$:

$$\begin{aligned} \min \ & f_0(x) \\ & f_j(x) = 0 \qquad\qquad\qquad j \in J(x) \end{aligned}$$

   O método é primal: A satisfação de $f_j(x) \leq 0 : j \notin J(x)$ é monitorada durante a otimização sobre a working surface.
5. Se $\overline{\mu} \geq 0$, $(\overline{x}, \overline{\mu})$ satisfaz CNPO. Se $\exists\, \overline{\mu}_j < 0 : j \in J(x)$, o índice $j$ deve ser removido do working set.

## Interpretação da mudança do working set

- $\exists\, \overline{\mu}_k < 0 : k \in J(x)$, o índice $k$ deve ser removido do working set do novo ponto $\overline{x}$.
- Observe que $J(x) \subseteq J(\overline{x})$, logo $k \in J(\overline{x})$.
- Isto significa, que permitir que a restrição $f_k(x) = 0$ seja relaxada para $f_k(x) < 0$ (saindo do working set) pode reduzir a função objetivo:

Para $\overline{x}$ regular, $\exists y \in V(\overline{x})$ tal que $\nabla f_k(\overline{x})^T y < 0$ (cc $\nabla f_k(\overline{x}) = 0$).

$$\nabla f_0(\overline{x}) + \sum_{j \in J(\overline{x})} \nabla f_j(\overline{x}) \overline{\mu}_j = 0$$

$$\nabla f_0(\overline{x})^T y + \sum_{j \in J(\overline{x})} \overline{\mu}_j \nabla f_j(\overline{x})^T y = 0$$

$$\nabla f_0(\overline{x})^T y + \overline{\mu}_k \nabla f_k(\overline{x})^T y = 0$$

$$\nabla f_0(\overline{x})^T y = -\overline{\mu}_k \nabla f_k(\overline{x})^T y < 0$$

1. Multiplicadores de Lagrange indicam quais restrições devem sair do working set.
2. Ao se movimentar pela superfície ativa (para um working set fixo), é necessário garantir que $f_j(x) \leq 0$ para $j \notin J(x)$.
3. É comum então que neste processo alguma restrição $f_j(x) \leq 0 : j \notin J(x)$ torne-se ativa, devendo ser incorporada ao working set.

**Ingrediente principal desta classe de métodos**

Método para minimização sob restrições de igualdade (com controle da viabilidade das restrições inativas).

1. Antes de determinar o working set ótimo, diversos working sets intermediários são gerados, cada um correspondendo a um PNL distinto.
2. Para que o sinal correto dos multiplicadores de Lagrange sejam determinados, é necessário que ótimos globais dos problemas intermediários sejam determinados. Caso contrário, o working set anterior pode ser obtido novamente.
3. Para contornar esta dificuldade, o working set é modificado utilizando-se outros critérios. Por exemplo, é comum utilizar ótimos locais dos problemas intermediários, antes que o ótimo global para um dado working set tenha sido obtido (e o sinal correto dos multiplicadores inferido).
4. Consequência: zig-zagging no working set, infinitas trocas nos working sets.

- Mover ao longo de $-\nabla f_0(x^k)$ não garante a viabilidade do ponto obtido, caso o ponto de origem esteja na fronteira do conjunto de viabilidade.
- O Método do Gradiente Projetado consiste em projetar $-\nabla f_0(x^k)$ no convexo $X$, para se obter a direção de busca, de forma que aprimore a função objetivo e garanta viabilidade.
- A iteração típica é:

$$x^{k+1} = x^k - \alpha^k P \nabla f_0(x^k)$$

  onde $P$ é uma matriz de projeção, de forma que $-P\nabla f_0(x^k)$ seja viável e de descida.
- Do ponto de vista computacional, a principal desvantagem do método é o elevado custo associado à projeção, em cada iteração. Logo o método é uma boa alternativa quando a operação de projeção pode ser feita com baixo custo computacional.

### Definição

Assuma que $P \in \mathbb{R}^{n \times n}$ seja uma matriz quadrada de ordem $n$. $P$ é chamada matriz de projeção (ou idempotente) se $PP = P$. Se além disto satisfizer $P = P^T$ é chamada de matriz de projeção ortogonal.

### Proposição

*Seja $P$ uma matriz quadrada de ordem $n$. Então, o seguinte é verdadeiro:*

1. *$P$ é matriz de projeção $\Rightarrow$ $P \succeq 0$.*
2. *$P$ é uma matriz de projeção $\iff$ $I - P$ é matriz de projeção.*
3. *Seja $P$ uma matriz de projeção e $Q = I - P$. Então os subespaços lineares $L = \{Px : x \in \mathbb{R}^n\}$ e $L^\perp = \{Qx : x \in \mathbb{R}^n\}$ são ortogonais. Além disto, qualquer $x \in \mathbb{R}^n$ pode ser escrito como $x = p \oplus q$ onde $p \in L$, $q \in L^\perp$, através de $p, q$ únicos.*

## Prova

1. *Seja* $x \in \mathbb{R}^n$ *um vetor arbitrário. Então, se* $P$ *é uma matriz de projeção,* $x^T P x = x^T P P x = x^T P^T P x = \|Px\|^2 \geq 0$.
2. *Claramente* $L, L^\perp$ *são subespaços lineares.*
   *Uma vez que* $P^T Q = P^T(I - P) = P - P^T P = 0$, $L, L^\perp$ *são de fato ortogonais.*
   *(Unicidade) suponha que* $x = p' + q'$ *e que* $x = p + q$. *Então subtraindo a primeira da segunda temos* $p - p' = q' - q$. *Uma vez que* $L \cap L^\perp = 0$, $p - p' = q' - q = 0$ *e a representação é única.*

- Dado $z \in \mathbb{R}^n$ temos que $\langle Pz, (I - P)z \rangle = 0$. Ou seja, a aplicação $Pz$ projeta $z$ no espaço ortogonal ao da aplicação $Qz = (I - P)z$.

- Podemos escrever que $z$ como a soma direta de $Pz$ e $Qz$: $z = Pz \oplus Qz$.

## Projeção

Consiste em encontrar um vetor $x^* \in X$, cuja distância a $z$ seja mínima:

$$\begin{array}{ll} \text{minimize} & \|z - x\|_2^2 \\ & x \in X \end{array}$$

## Recordando alguns fatos sobre a projeção em um convexo

- Para todo $z \in \mathbb{R}^n$ existe um $x^* \in X$ único que minimiza $\|z - x\|_2^2$ sobre todo $x \in X$. Este vetor é chamado projeção de $z$ em $X$. Usaremos a notação $x^* = [z]^+$.
- Dado $z \in \mathbb{R}^n$, $x^* = [z]^+ \iff (z - x^*)(x - x^*) \leq 0, \forall x \in X$.
- No caso de $X$ ser um subespaço, $x^* \in X$ é a projeção de $z$ em $X$ se e somente se $z - x^*$ for ortogonal a $X$, isto é: $(z - x^*)^T x = 0, \forall x \in X$.

- Desejamos resolver

$$\begin{array}{rrcl} \text{minimize} & f_0(x) & & \\ & Ax & \leq & b \\ & Ex & = & e \end{array}$$

onde $A \in \mathbb{R}^{m\times n}$, $E \in \mathbb{R}^{q\times n}$, $b \in \mathbb{R}^{m\times 1}$, $e \in \mathbb{R}^{q\times 1}$ e $X = \{x \in \mathbb{R}^n : Ax \leq b, Ex = e\}$.

- Dispomos de um ponto viável tal que: $A_1 x^k = b_1$ e $A_2 x^k < b_2$, onde $A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$ $b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$.

Desejamos projetar $-\nabla f_0(x^k)$ na face de $X$ ativa em $x^k$ que é o conjunto afim $\mathcal{V} = \{x \in \mathbb{R}^n : A_1 x = b_1, Ex = e\}$.

## Problema de programação quadrática

$$\begin{aligned} \text{minimize} \quad & \|x - (-\nabla f_0(x^k))\|_2^2 \\ A_1 x &= b_1 \\ Ex &= e \end{aligned}$$

cuja solução analítica é $x^*$ tal que $x^* - x^k = d^k = -P\nabla f_0(x^k)$

- $P$ é a matriz de projeção $P$ associada à $\mathcal{N}(M)$, $M = \begin{pmatrix} A_1 \\ E \end{pmatrix}$.
- Assumimos que $M$ possui posto completo igual a $q + m_1$, $m_1$ o número de linhas de $A_1$ (veja que esta é uma condição de regularidade que é assumida).

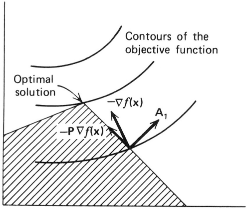
Para obter $P := I - M^T(MM^T)^{-1}M$

- $\mathbb{R}^n = \mathcal{N}(M) \oplus \mathcal{R}(M^T)$
- $d^k \in \mathcal{N}(M)$.

$$
\begin{aligned}
-\nabla f_0(x^k) &= d^k + M^T w \\
-M\nabla f_0(x^k) &= Md^k + MM^T w \\
w &= -(MM^T)^{-1}M\nabla f_0(x^k) \\
d^k &= -\nabla f_0(x^k) + M^T(MM^T)^{-1}M\nabla f_0(x^k) \\
&= -[I - M^T(MM^T)^{-1}M]\nabla f_0(x^k)
\end{aligned}
$$

Qual interpretação podemos dar ao vetor $w$ ?

- Uma vez que $MP = 0$ temos que $A_1 P = 0$, $EP = 0$ e então a matriz de projeção $P$ projeta cada linha de $A_1$ e de $E$ no vetor nulo.
- Entretanto, as linhas de $A_1$ e de $E$ são os gradientes das restrições ativas em $x^k$, a matriz $P$ é na verdade a matriz que projeta os gradientes das restrições ativas em $x^k$ no vetor zero.
- Assim sendo, $P\nabla f_0(x^k)$ corresponde à projeção de $\nabla f_0(x^k)$ no espaço nulo das restrições ativas.
  Veja: $MP\nabla f_0(x^k) = 0$, logo, $P\nabla f_0(x^k) \in \mathcal{N}(M)$



## Proposição

*Considere o poliedro* $X = \{x \in \mathbb{R}^n : Ax \leq b, Ex = e\}$ *e suponha que dispomos de* $x^k \in X$ *e que* $A_1 x^k = b_1, A_2 x^k < b_2$*, conforme definimos.*

*Então,* $d^k$ *é uma direção viável em* $x^k$ *se e somente se* $A_1 d^k \leq 0, Ed^k = 0$ *(a demonstração é óbvia).*

**Proposição**

*Considere o problema* $\min_{x \in X} f_0(x)$ *onde* $X$ *é o poliedro definido anteriormente e assuma que* $x^k$ *seja um ponto viável tal que*

$$Ex^k = e, \; A_1 x^k = b_1 \text{ e } A_2 x^k < b_2, \text{ onde } A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

*Suponha que* $f_0$ *seja diferenciável em* $x^k$.

*Então:*

1. *Se* $P$ *é uma matriz de projeção tal que* $P\nabla f_0(x^k) \neq 0$, *então* $d^k = -P\nabla f_0(x^k)$ *é uma direção de descida de* $f$ *em* $x^k$.
2. *Além disto, se* $M = \begin{pmatrix} A_1 \\ E \end{pmatrix}$ *possui posto completo e* $P := I - M^T(MM^T)^{-1}M$, *então* $d^k$ *é uma direção viável de descida.*

## Prova

1. *($-P\nabla f_0(x^k)$ é direção de descida)*
   *Observe que:* $\nabla f_0(x^k)^T d^k = -\nabla f_0(x^k)^T P \nabla f_0(x^k) < 0$, *uma vez que* $P \succeq 0$ *e, pelo enunciado* $\nabla f_0(x^k)^T P \neq 0$. *Então* $d^k$ *é de descida.*
2. *($-P\nabla f_0(x^k)$ é direção viável - já sabemos disto pois $P$ projeta em $\mathcal{N}(M)$)*
   *Se $M$ possui posto completo, $(MM^T)$ é não singular e:*

$$
\begin{aligned}
P &= I - M^T(MM^T)^{-1}M \\
MP &= M - MM^T(MM^T)^{-1}M && = 0 \\
\textit{Logo} \quad Md^k &= -MP\nabla f_0(x^k) && = 0
\end{aligned}
$$

*Se* $Md^k = 0$ *temos* $A_1 d^k = 0, Ed^k = 0$ *e* $d^k$ *é viável.*

## Tratando o caso $P\nabla f_0(x^k) = 0$

$$
\begin{aligned}
0 &= P\nabla f_0(x^k) \\
&= [I - M^T(MM^T)^{-1}M]\nabla f_0(x^k) \\
&= \nabla f_0(x^k) + M^T w \\
&= \nabla f_0(x^k) + A_1^T u + E^T v
\end{aligned}
$$

onde $w = -(MM^T)^{-1}M\nabla f_0(x^k)$.

- Se $u \geq 0$ (associados às restrições de desigualdades), temos a caracterização de que $x^k$ é um ponto regular e um ponto KKT. Veja que $x^k$ é estacionário. Paramos.
- Se $u \not\geq 0$, vamos mostrar como obter uma nova matriz de projeção e direção de descida.

### Tratando o caso $P\nabla f_0(x^k) = 0$

- Assumindo que $u_j < 0$ associado a $j-$ésima linha de $A_1 x \leq b_1$ : $a_j^T x^k = b_j$
- Excluímos a linha $a_j^T$ de $M$ e obtemos $\overline{M}$. Definimos $\overline{P} = I - \overline{M}^T(\overline{M}\overline{M}^T)^{-1}\overline{M}$ e $\overline{d}^k = -\overline{P}\nabla f_0(x^k)$.

## Vamos mostrar que $\overline{P}\nabla f_0(x^k) \neq 0$

- Uma vez que $P\nabla f_0(x^k) = 0$ temos $-\nabla f_0(x^k) = M^T w$
- Para algum $\overline{w}$ temos: $-\nabla f_0(x^k) = \overline{d}^k + \overline{M}^T \overline{w}$.

$$-\nabla f_0(x^k) = M^T w$$
$$-\nabla f_0(x^k) = \overline{M}^T \overline{w} + \overline{d}^k$$

- Se $\overline{d}^k = 0$, teríamos $u_j a_j^T = 0$. Isto porque $M$ (de onde $a_j$ foi retirada) possui posto completo. Ou seja, $-\nabla f_0(x^k)$ não pode estar no span das colunas de $M^T$ e $\overline{M}^T$ simultaneamente, sem que $u_j = 0$ (contradição). Logo $\overline{d}^k = -\overline{P}\nabla f_0(x^k) \neq 0$.
- Como $\overline{P} \succeq 0$, $\overline{d}^k$ é de descida.

$$\overline{d}^k = -\overline{P}\nabla f_0(x^k) \neq 0 \text{ também é viável.}$$

## Além disto, é viável

$$\begin{aligned} \nabla f_0(x^k) &= -M^T w \\ 0 > \nabla f_0(x^k)^T \overline{d}^k &= -w^T M \overline{d}^k \\ &= -w^T \begin{pmatrix} a_j^T \\ \overline{M} \end{pmatrix} \overline{d}^k \\ &= -u_j a_j^T \overline{d}^k && \text{já que } \overline{M}\overline{d}^k = 0 \end{aligned}$$

Como $u_j < 0$, temos que $a_j^T \overline{d}^k < 0$ e é direção viável.

## Iteração típica, dado $x^k$ viável

1. Determinamos $\mathcal{V}$, $M$ e a matriz de projeção $P := I - M^T(MM^T)^{-1}M$ e $d^k = -P\nabla f_0(x^k)$.
2. Se $d^k \neq 0$, faça a busca linear restrita $x^{k+1} = x^k + \alpha^k d^k$ e retorne para Passo 1.
   (busca linear restrita = não pode violar as desigualdades $A_2 x \leq b_2$, folgadas em $x^k$. Isto apenas limita o tamanho do passo.)
3. Se $d^k = 0$, calcule $w = (u \;\; v)^T = -(MM^T)^{-1}M\nabla f_0(x^k)$.
   1. Se $u \geq 0$, páre. $x^k$ é um ponto KKT (interpretação geométrica).
   2. Caso contrário, elimine de $M$ a linha de $A_1$ associada ao mais negativo multiplicador $u_j$ e retorne para o Passo 1.

# Método do Gradiente Projetado - Exemplo

$$\begin{aligned}
\min \;\; & x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1 - 3x_4 \\
& 2x_1 + x_2 + x_3 + 4x_4 = 7 \\
& x_1 + x_2 + 2x_3 + x_4 = 6 \\
& x_i \geq 0, \;\; i = 1, \ldots, 4
\end{aligned}$$

ponto inicial $x = (2,\ 2,\ 1,\ 0)^T$

Para o ponto inicial $x = (2,\ 2,\ 1,\ 0)^T$:

- as restrições de igualdade e $x_4 \geq 0$ são ativas.
- $M = \begin{pmatrix} 2 & 1 & 1 & 4 \\ 1 & 1 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, MM^T = \begin{pmatrix} 22 & 9 & 4 \\ 9 & 7 & 1 \\ 4 & 4 & 1 \end{pmatrix}$

  $(MM^T)^{-1} = \frac{1}{11} \begin{pmatrix} 6 & -5 & -19 \\ -5 & 6 & 14 \\ -19 & 14 & 73 \end{pmatrix}$
- $P = I - M^T(MM^T)^{-1}M = \frac{1}{11} \begin{pmatrix} 1 & -3 & 1 & 0 \\ -3 & 9 & -3 & 0 \\ 1 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

- Projetando o gradiente $\nabla f_0(x) = \begin{pmatrix} 2x_1 - 2 & 2x_2 & 2x_3 & 2x_4 - 3 \end{pmatrix}^T$ no espaço nulo das restrições ativas:

$$d = -P\nabla f_0(x) = -\frac{1}{11}\begin{pmatrix} 1 & -3 & 1 & 0 \\ -3 & 9 & -3 & 0 \\ 1 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}\begin{pmatrix} 2 \\ 4 \\ 2 \\ -3 \end{pmatrix} = \frac{1}{11}\begin{pmatrix} 8 \\ -24 \\ 8 \\ 0 \end{pmatrix}$$

- Normalizando a direção temos $d = \begin{pmatrix} 1 & -3 & 1 & 0 \end{pmatrix}^T$
- $x(\alpha) = \begin{pmatrix} 2+\alpha & 2-3\alpha & 1+\alpha & 0 \end{pmatrix}^T \geq 0 \to \alpha \leq \frac{2}{3}$.
- Busca unidirecional ao longo de $d$:
  $\alpha^* = \arg\min\{f_0(x + \alpha d) : 0 \leq \alpha \leq \frac{2}{3}\}$.
  $f_0(\alpha) = 5 - 6\alpha + 11\alpha^2$, cujo mínimo ocorre em $\alpha^* = \frac{6}{22} < \frac{2}{3}$

- $\overline{x} = \begin{pmatrix} 2 & 2 & 1 & 0 \end{pmatrix}^T + \frac{6}{22} \begin{pmatrix} 1 & -3 & 1 & 0 \end{pmatrix}^T = \begin{pmatrix} 2.0909 & 1.7272 & 1.0909 & 0 \end{pmatrix}^T$
- A função objetivo caiu de $f_0(x) = 5$ para $f_0(\overline{x}) = 4.3636$
- A matriz $M$ não foi modificada para o novo ponto, de forma que $P$ permanece a mesma.
- Basta avaliar $\nabla f_0(\overline{x})$ para obter $d = -P\nabla f_0(\overline{x})$ e implementar uma nova busca unidirecional.

- A versão original, proposta por Wolfe (1963), resolve problemas nos quais a função objetivo é não linear e o conjunto de restrições poliedral, na forma padrão. O método guarda similaridades com o Método Simplex em Programação Linear.

$$\begin{aligned} \min\ & f_0(x) \\ & Ax = b \\ & x \geq 0 \end{aligned}$$

- Posteriormente, foi generalizado (Generalized Reduced Gradient Method) por Abadie e Carpentier (1969) para resolver o problema mais geral, com restrições de não negatividade nas variáveis.

$$\begin{aligned} \min\ & f_0(x) \\ & h(x) = 0 \\ & x \geq 0 \end{aligned}$$

## Problema a ser resolvido

$$\begin{aligned} \min \ & f_0(x) \\ & Ax = b \\ & x \geq 0 \end{aligned}$$

onde $f_0 \in C^2, A \in \mathbb{R}^{m \times n}, b \in \mathbb{R}^m$.

**Hipóteses de não degeneração:**

1. Qualquer conjunto de $m$ colunas de $A$ possui posto completo.
2. Qualquer solução extrema de $P = \{x \in \mathbb{R}^n : Ax = b, x \geq 0\}$ é não degenerada. Isto significa que qualquer solução extrema de $P$ possui $n - m$ entradas nulas e exatamente $m$ entradas positivas.

Observação: Dizemos que $P$ é degenerado se alguma solução básica de $P$ possui mais de $n - m$ entradas nulas (e, consequentemente, menos de $m$ entradas positivas).

## Reformulando o problema

- Particionamos $A = (B \;\; N)$: $B \in \mathbb{R}^{m \times m}$ e $N \in \mathbb{R}^{m \times n-m}$
- Analogamente, particionamos $x = (y \;\; z)^T$, onde $y$ é o (vetor básico) e $z$ é o (vetor não básico). Então temos:

$$y \geq 0, z \geq 0$$

$$Ax = b \rightarrow By + Nz = b$$

- Ao longo do método teremos $y > 0$. As componentes de $z$ podem ser nulas ou positivas (diferentemente do método Simplex).
- Como qualquer submatriz $B \in \mathbb{R}^{m \times m}$ de $A$ possui posto completo, escrevemos as variáveis $y$ (dependentes) em função das $z$ (independentes)

$$y = B^{-1}b - B^{-1}Nz$$

**Problema reformulado**

$$\begin{aligned} \min\ & f_0(B^{-1}b - B^{-1}Nz, z) \\ & y = B^{-1}b - B^{-1}Nz \\ & y \geq 0, z \geq 0 \end{aligned} \tag{1}$$

- Em função de (1) a ideia do método (assim como no Simplex) consiste em considerar o problema apenas em função de $z$.
- Trata-se de uma variante do método do gradiente puro, para lidar com a restrição (1).

- $\nabla f_0(x) = \begin{pmatrix} \nabla_y f_0(x)^T & \nabla_z f_0(x)^T \end{pmatrix}$
  - $\nabla_y f_0(x) \in \mathbb{R}^m$: gradiente de $f_0(x)$ em relação a $y$
  - $\nabla_z f_0(x) \in \mathbb{R}^{n-m}$: gradiente de $f_0(x)$ em relação a $z$.
- Uma direção $d^T = (d_y^T \ \ d_z^T)$ viável de descida deve satisfazer:

$$\nabla f_0(x)^T d < 0$$
$$0 = Ad = Bd_y + Nd_z$$
$$d_i \geq 0 \text{ se } x_i = 0$$

Fazendo $d_y = -B^{-1}Nd_z$ satisfazemos $Bd_y + Nd_z = 0$.

- Definimos o *gradiente reduzido (análogo ao custo reduzido de PL):*

$$\begin{aligned} r^T &= (r_y^T \;\; r_z^T) \\ &:= \nabla f_0(x)^T - \nabla_y f_0(x)^T B^{-1} A \\ &= \left( \nabla_y f_0(x)^T \quad \nabla_z f_0(x)^T \right) - \nabla_y f_0(x)^T B^{-1} (B \;\; N) \\ &= \left( \nabla_y f_0(x)^T \quad \nabla_z f_0(x)^T \right) - \nabla_y f_0(x)^T (I \;\; B^{-1} N) \\ &= \left( 0 \quad (\nabla_z f_0(x)^T - \nabla_y f_0(x)^T B^{-1} N) \right) \end{aligned}$$

- Examinando o termo $\nabla f_0(x)^T d$:

$$\begin{aligned} \nabla f_0(x)^T d &= \nabla_y f_0(x)^T d_y + \nabla_z f_0(x)^T d_z \\ &= (-\nabla_y f_0(x)^T B^{-1} N + \nabla_z f_0(x)^T) d_z \\ &= r_z^T d_z \end{aligned}$$

**Devemos escolher $d_z$ de forma que $r_z^T d_z < 0$ e $d_j \geq 0$ caso $x_j = 0$.**

**Desejamos definir $d$: $r_z^T d_z < 0$ e $d_j \geq 0$ caso $x_j = 0$.**

1. Avaliando $d_z$ (componentes não básicas)
2. Naturalmente, $d_y = -B^{-1}Nd_z$ onde $d_z$ foi definido como indicado acima. Garantimos assim que

$$d = (-B^{-1}Nd_z, d_z) \in \mathcal{N}(A)$$

   e portanto $A(x + d) = b$.
3. O ponto será viável desde que $(y + d_y, z + d_z) \geq 0$ seja garantido na determinação do passo.

## Abordagem de restrições ativas

1. Definimos $J(z) = \{j : z_j = 0\}$
2. A direção abaixo é restrita à superfície ativa, isto é, todas as coordenadas $z_i : i \in J(z)$ permanecem no nível de atividade 0 ainda que seu custo reduzido seja negativo.
   Avaliando $d_z$ (componentes não básicas)
   $$d_j = \begin{cases} -r_j & \text{se } j \notin J(z) \\ 0 & j \in J(z) \end{cases}$$
3. Adia a atualização do working set: $j \in J(z)$ só sai do working set se todas as componentes $z_j : j \notin J(z)$ satisfizerem $r_j = 0$. Neste caso, caso exista $j : j \in J(z), r_j < 0$, $j$ sai do working set. Caso contrário, as condições necessárias de primeira ordem são satisfeitas.
4. O working set é atualizado quando uma nova variável dependente $y_j$ torna-se zero.

## Sem adiar a atualização do working set

1. Avaliando $d_z$ (componentes não básicas)

$$d_j = \begin{cases} -r_j & \text{se } r_j < 0 \text{ ou } z_j > 0 \\ 0 & \text{caso contrário} \end{cases}$$

   - Esta regra garante $r_z^T d_z \leq 0$ e $d_j \geq 0$ quando $x_j = 0$.
   - Se $d_z = 0$, a solução encontrada é um ponto KKT.
   - Se $d_z \neq 0$, $\nabla f_0(x)^T d = r_z^T d_z < 0$.

2. O working set também é atualizado quando uma nova variável dependente $y_j$ torna-se zero.

## Calculando o passo

Determinamos $\alpha_1, \alpha_2, \alpha_3$

- $\alpha_1 = \arg\max\{\alpha : y + \alpha d_y \geq 0\}$
- $\alpha_2 = \arg\max\{\alpha : z + \alpha d_z \geq 0\}$
- $\alpha_3 = \arg\max\{f_0(x + \alpha d) : 0 \leq \alpha \leq \min\{\alpha_1, \alpha_2\}\}$.

Fazemos $\overline{x} = x + \alpha_3 d$

- Se $\alpha_3 < \alpha_1$, recalculamos $r_z$ e repetimos o processo sem mudança do particionamento das colunas de $A$, entre básicas e não básicas.
- Se $\alpha_3 = \alpha_1$, precisamos realizar uma operação de pivoteamento. Uma variável básica $x_i > 0$ tornou-se nula e deve ser substituída por alguma outra variável $x_k$ não básica tal que $\overline{x}_k > 0$. Pela hipótese de não degeneração, esta variável $x_k$ existe. Alteramos $B$ e $N$, calculamos $r_z$ diante do novo particionamento e repetimos.

$$\begin{aligned} \min \ & x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1 - 3x_4 \\ & 2x_1 + x_2 + x_3 + 4x_4 = 7 \\ & x_1 + x_2 + 2x_3 + x_4 = 6 \\ & x_i \geq 0, \ \ i = 1, \ldots, 4 \end{aligned}$$

ponto inicial $x = (2,\ 2,\ 1,\ 0)^T$

- Dispomos de 3 variáveis não nulas. Arbitrariamente, escolhemos $y = (x_1,\ \ x_2)^T = (2,\ \ 2)^T$ como vetor básico e $z = (x_3,\ \ x_4)^T = (1,\ \ 0)^T$ como vetor não básico.
- Na forma padrão, temos o quadro:

$$\begin{aligned} x_1 &= 1 + x_3 - 3x_4 \\ x_2 &= 5 - 3x_3 + 2x_4 \\ x_i &\geq 0, \ \ i = 1, \ldots, 4 \end{aligned}$$

$$B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} B^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} N = \begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix}, \nabla f_0(x) = \begin{pmatrix} 2x_1 - 2 \\ 2x_2 \\ 2x_3 \\ 2x_4 - 3 \end{pmatrix}$$

- Cálculo de $r_z$ e $d$

$$\begin{aligned} r_z^T =& \nabla_z f_0(x)^T - \nabla_y f_0(x)^T B^{-1} N \\ =& (2 \quad -3) - (2 \quad 4) \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix} = (-8 \quad -1) \\ d_z^T =& (8 \quad 1) \\ d_y^T =& -(B^{-1} N d_z)^T = -\left( \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} 8 \\ 1 \end{pmatrix} \right)^T \\ =& (5 \quad -22) \\ d^T =& (5 \quad -22 \quad 8 \quad 1) \end{aligned}$$

- Determinação de $\alpha_1, \alpha_2$:

$$\overline{x} = \begin{pmatrix} 2 \\ 2 \\ 1 \\ 0 \end{pmatrix} + \alpha \begin{pmatrix} 5 \\ -22 \\ 8 \\ 1 \end{pmatrix} \geq 0 \rightarrow \alpha_1 = \frac{1}{11}, \alpha_2 = \infty$$

- Determinação de $\alpha_3$ (busca unidirecional, exata):

$$\begin{aligned} \alpha_3 &= \arg\min\{f_0(x + \alpha d) : 0 \leq \alpha \leq \alpha_1\} \\ &= \arg\min \left\{5 - 65\alpha + 574\alpha^2 : 0 \leq \alpha \leq \frac{1}{11}\right\} \\ &= \frac{65}{1148} \end{aligned}$$

- Determinação do novo ponto

$$\begin{aligned} \overline{x}^T &= (2 \quad 2 \quad 1 \quad 0) + \frac{65}{1148}(5 \quad -22 \quad 8 \quad 1) \\ &= (2.2831 \quad 0.7544 \quad 1.4530 \quad 0.0566) \end{aligned}$$

$$\overline{x} = (2.2831 \quad 0.7544 \quad 1.4530 \quad 0.0566)^T$$

A função objetivo foi reduzida de 5 para 3.1599

Como $\alpha_3 < \alpha_1$ nenhuma variável básica foi anulada.

Mantemos o particionamento $(B \quad N)$ de $A$ e repetimos o processo:

1. Calculando o novo gradiente reduzido
2. Calculando a direção de busca e verificando a condição de otimalidade ($d_z = 0$?)
3. Determinando $\alpha_1, \alpha_2$
4. Determinando $\alpha_3$ e implementando o passo.
5. Verificando a necessidade de redefinir o particionamento.

Trata-se de uma especialização do Método do Gradiente Reduzido em que, apenas uma variável independente é modificada por vez.

- $d_j = 1$ para alguma variável $z_j$ independente
- $d_j = 0$ para todas as demais variáveis $z_j$.
- $d_y = -B^{-1} N d_z$, de forma que $d = (d_y, d_z) \in \mathcal{N}(A)$.

Seja $z_j$ a variável independente que muda:

1. A variável $z_j$ cresce até que um mínimo local na direção $d_z$ seja atingido ou que a fronteira do conjunto de viabilidade seja atingido e uma nova restrição de não negatividade tenha se tornado justa.

## Escolhendo a direção

Calculamos o gradiente reduzido $r_z$ associado às variáveis independentes e avaliamos:

- $\hat{r}_1 = \min_{z_j}\{r_j\}$, $\hat{z}_1 = \arg\min_{z_j}\{r_j\}$
- $\hat{r}_2\hat{z}_2 = \max_j\{r_j z_j\}$, $\hat{z}_2 = \arg\max_j\{r_j z_j\}$
  - Se $\hat{r}_1 = \hat{r}_2\hat{z}_2 = 0$, páre. Caso contrário:
  - Se $\hat{r}_1 \leq -|\hat{r}_2\hat{z}_2|$, incremente a variável $\hat{z}_1$.
  - Se $\hat{r}_1 \geq -|\hat{r}_2\hat{z}_2|$, reduza a variável $\hat{z}_2$

1. A variável independente que muda é escolhida segundo o potencial de redução na função objetivo, pesando-se os custos reduzidos pela sua distância a zero.
2. Isto garante convergência global pois em toda iteração há redução da função objetivo (hipótese de não degeneração).

1. O método do Simplex Convexo pode ser visto como um método de coordenadas descentes (ex. Gauss-Southwell) no espaço das $n - m$ variáveis independentes.
2. É razoável admitir que o método requeira em torno de $n - m$ passos para realizar o progresso de uma única iteração do Gradiente Reduzido.
3. Para ser competitivo, o custo por iteração do Simplex Convexo deve ser $n - m$ vezes menor que o custo do Gradiente Reduzido (busca unidirecional + avaliação das restrições).
4. O Simplex Convexo é competitivo para programas lineares (neste caso é o próprio Método Simplex) e para programas quadráticos, pois para funções objetivo mais gerais requer, assim como o Gradiente Reduzido, o uso da busca unidirecional ao longo da direção escolhida.

## Caso com restrições de igualdade

$$\begin{aligned} &\min f_0(x) \\ &\quad h_j(x) = 0 \qquad\qquad j = 1, \ldots, q \end{aligned}$$

A caracterização das condições necessárias e suficientes de otimalidade para um candidato a mínimo local $x^*$ do PNL depende do conceito de regularidade do ponto $x^*$.

### Definição - regularidade

- Um ponto $x^*$ viável é dito regular se os vetores gradientes de todas as restrições ativas em $x^*$, no caso em questão $\nabla h_1(x^*), \ldots, \nabla h_q(x^*)$, são linearmente independentes.
- Na presença de restrições de igualdade e desigualdade, os gradientes de todas as restrições ativas em $x^*$ (desigualdades justas e igualdades) devem ser linearmente independentes, para que $x^*$ seja regular.

## Restrições de igualdade

O principal resultado é o seguinte:

se $x^*$ é um ponto de mínimo local regular, então existem escalares $\lambda_i : i = 1, \ldots, q$, denominados Multiplicadores de Lagrange, tais que vale a seguinte equação vetorial:

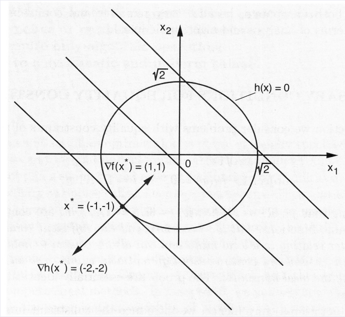
$$\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i \nabla h_i(x^*) = 0$$

- O gradiente da função objetivo $\nabla f_0$ em $x^*$ pertence ao subespaço gerado pelos gradientes $\nabla h_i : i = 1, \ldots, q$ das restrições ativas em $x^*$.

$$\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i \nabla h_i(x^*) = 0$$

## Exemplo

$$\min \quad x_1 + x_2$$
$$x_1^2 + x_2^2 = 2$$



## Definição

Definimos o subespaço de variações nulas de primeira ordem, isto é, o subespaço para o qual o vetor $x = x^* + \triangle x$ satisfaz as restrições $h(x) = 0$ até a primeira ordem, como:

$$V(x^*) = \left\{ y : \nabla h_i(x^*)^T y = 0, \;\; i = 1, \ldots, q \right\}$$

$$V(x^*) = \left\{ y : \nabla h_i(x^*)^T y = 0, \quad i = 1, \ldots, q \right\}$$

- Tome $y \in V(x^*)$;
- Uma vez que

$$\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i \nabla h_i(x^*) = 0,$$

temos que

$$\nabla f_0(x^*)^T y + \sum_{i=1}^{q} \lambda_i \nabla h_i(x^*)^T y = 0$$

e logo $\nabla f_0(x^*)^T y = 0$.

- Ou seja, $\nabla f_0(x^*) \perp V(x^*)$.
- Este resultado é análago à condição $\nabla f_0(x^*) = 0$ para um ponto estacionário de um problema de otimização irrestrita.

$$V(x^*) = \left\{ y : \nabla h_i(x^*)^T y = 0, \;\; i = 1, \ldots, q \right\}$$

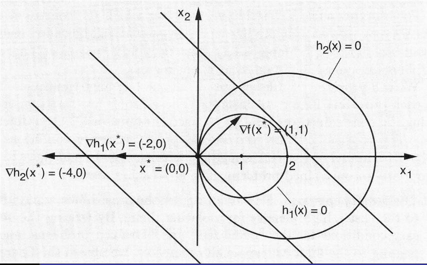
- Inferimos o comportamento de $f_0(x)$ a partir de direções suficientemente viáveis, isto é, direções viáveis em primeira ordem em torno de $\overline{x} + \alpha d$: $d \in V(\overline{x})$.
- Para $\alpha$ pequeno, o comportamento de $f_0$ nas vizinhanças de $\overline{x}$ perturbado por $\alpha d$ deve ser similar ao comportamento de $f_0(x)$ em torno de $\overline{x}$ na superfície $h(\overline{x} + \tilde{d}) = 0$ para $\tilde{d} \approx \alpha d$.

$$\min \ x_1 + x_2$$
$$x_1^2 + x_2^2 = 2$$

- Tomando o ponto $\overline{x} = (\sqrt{2} \ \ 0)^T$, $\nabla f_0(\overline{x}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$,
  $\nabla h(\overline{x}) = \begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix}$ e $V(\overline{x}) = \{y \in \mathbb{R}^2 : y_1 = 0\}$.
- Observe que existem $d^1 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ e $d^2 = -d^1$ satisfazendo $d^1, d^2 \in V(\overline{x})$, $\nabla f_0(\overline{x})^T d^1 > 0$ e $\nabla f_0(\overline{x})^T d^2 < 0$, de forma que $\overline{x}$ não é de mínimo ou de máximo local.
- Isto ocorre já que $\nabla f_0(\overline{x}) \not\perp V(\overline{x})$.

$$
\begin{array}{lrcll}
\text{minimize} & f_0(x) & = & x_1 + x_2 & \\
& h_1(x) & = & (x_1-1)^2 + x_2^2 - 1 & = \quad 0 \\
& h_2(x) & = & (x_1-2)^2 + x_2^2 - 4 & = \quad 0
\end{array}
$$

Observe que o único ponto viável é o ponto não regular $(0,0)$, onde $\nabla h_1$ e $\nabla h_2$ são linearmente dependentes, não permitindo sintetizar $\nabla f_0$ em $\mathbb{R}^2$.



Dada uma superfície $X$ do $\mathbb{R}^n$, a forma como a descrevemos por meio de restrições $h(x) = 0$, afeta a regularidade de pontos em $X$, mediante a representação escolhida para a mesma superfície.

## Exemplo

Considere as duas funções que representam o mesmo subconjunto do $\mathbb{R}^2$: $h_1(x_1, x_2) := x_1 = 0$ e $h_1(x_1, x_2) := x_1^2 = 0$. O ponto $(0 \quad 0)^T$ é regular diante da primeira representação, mas não é regular diante da segunda representação. Isto porque no primeiro caso,

$\nabla h_1(0,0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, enquanto que no segundo caso,

$\nabla h_1(0,0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Na verdade, diante da segunda representação, nenhum ponto satisfazendo $x_1^2 = 0$ é regular, pois todos possuem o vetor gradiente igual a zero.

1. A regularidade é uma forma de qualificação das restrições que permita escrever condições para a caracterização de otimalidade, por meio da caracterização de direções viáveis e de descida, fazendo aproximações (série de Taylor truncada) de primeira ordem para função objetivo e restrições.
2. Estas condições de qualificação são hipóteses que garantem a similaridade entre $X = \{x \in \mathbb{R}^n : h(x) = 0, f(x) \leq 0\}$ e sua aproximação linear, nas vizinhanças de um candidato $x^*$ a ótimo local.
3. Quando a aproximação linear captura os aspectos essenciais da geometria do problema, esta abordagem gera conclusões importantes.
4. Quando a linearização é substancialmente diferente da região de viabilidade, por exemplo, a linearização é todo um plano enquanto a região de viabilidade é um ponto apenas, a aproximação linear não permite estabelecer sobre a otimalidade.

Dado $x \in X$ onde $X = \{x \in \mathbb{R}^n : h(x) = 0, f(x) \leq 0\}$ é o conjunto de viabilidade de P.

### Sequência viável aproximando $x$

Chamamos $\{z^k\}$ de uma sequência viável aproximando $x$ se $z^k \in X$ para todo $k$ suficientemente grande (isto é, existe $\overline{k}$ tal que $z^k \in X$ para $k \geq \overline{k}$) e $z^k \to x$.

### Caracterização de uma solução ótima local

Um ponto $x^*$ é um ponto ótimo local do PNL

$$(P) \qquad \begin{aligned} &\min f_0(x) \\ &x \in X \end{aligned}$$

se todas as sequências viáveis $\{z^k\}$ aproximando $x^*$ possuem a propriedade $f_0(z^k) \geq f_0(x^*)$ para todo $k$ suficientemente grande.

### Tangente

Uma direção $d$ é uma tangente (ou vetor tangente) de $X$ em um ponto $x$ se existem:

(1) uma sequência viável $\{z^k\}$ aproximando $x$ e

(2) uma sequência de escalares positivos $\{t^k\} : t^k \to 0$ tal que

$$\lim_{k\to\infty} \frac{z^k - x}{t^k} = d$$

### Cone tangente

O conjunto de todas as tangentes de $X$ em $x$ é designado cone tangente $T_X(x)$, ou equivalentemente, $T_X(x)$ é o conjunto de todas as tangentes de curvas suaves sobre $X$ no ponto $x$.

## Direções linearizadas em $x$ ($\mathcal{F}(x)$ análogo a $V(x)$)

Dado $x \in X$, o conjunto de direções linearizadas em $x$, $\mathcal{F}(x)$, é o cone definido como a seguir:

$$\mathcal{F} := \left\{ d \;\middle|\; \begin{array}{ll} \nabla h_i(x)^T d = 0 & i = 1, \dots, q \\ \nabla f_j(x)^T d \leq 0 & j = 1, \dots, m : f_j(x) = 0 \end{array} \right.$$

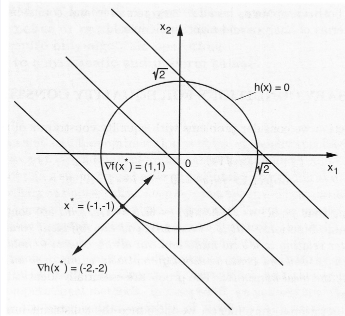
A definição de $\mathcal{F}(x)$ depende da representação escolhida para $X$ enquanto que a definição de $T_X(x)$ não depende:

- $T_X(x) \subseteq \mathcal{F}(x)$
- A regularidade do ponto $x$ garante $T_X(x) = \mathcal{F}(x)$.

Demonstração: Luenberger p. 325 ou Nocedal & Wright, p. 323

## Exemplo

$$\begin{aligned} \min \quad & x_1 + x_2 \\ & x_1^2 + x_2^2 = 2 \end{aligned}$$



## Sequências aproximando $x$ pelo terceiro quadrante

$$x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}, z^k = \begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} \\ -\frac{1}{k} \end{pmatrix}, t^k = \|z^k - x\|$$

- Observe que $z^k \in X$ para qualquer $k$.
- Diante destas escolhas
$$d = \lim_{k\to\infty} \frac{\begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} + \sqrt{2} \\ -\frac{1}{k} \end{pmatrix}}{\|z^k - x\|} = (0 \quad -1)^T$$
é uma tangente em $x$.

## Sequências aproximando $x$ pelo segundo quadrante

$$x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}, z^k = \begin{pmatrix} -\sqrt{2 - \frac{1}{k^2}} \\ \frac{1}{k} \end{pmatrix}, t^k = \|z^k - x\|$$

- Observe que $z^k \in X$ para qualquer $k$.
- Diante destas escolhas $d = (0 \;\; 1)^T$ é uma tangente em $x$.

$$T_X(x) = \{(d_1 \;\; d_2)^T \in \mathbb{R}^2 : d_1 = 0\}.$$

**$h_1(x) = x_1^2 + x_2^2 - 2 = 0$**

Para $x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$ e a representação $h_1(x) = x_1^2 + x_2^2 - 2 = 0$, temos que $\mathcal{F}(x) = \{d : (2x_1 \quad 2x_2) \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = 0\}$. Logo $\mathcal{F}(x) = \{d \in \mathbb{R}^2 : -2\sqrt{2}d_1 = 0\}$ e neste caso $\mathcal{F}(x) = T_X(x)$.

**$h_1(x) = (x_1^2 + x_2^2 - 2)^2 = 0$**

Para $x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$ e a diante da nova representação para $h_1(x) = 0$, temos que $\mathcal{F}(x) = \{d : \begin{pmatrix} 4(x_1^2 + x_2^2 - 2)x_1 \\ 4(x_1^2 + x_2^2 - 2)x_2 \end{pmatrix}^T \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} = 0\}$. Logo $\mathcal{F}(x) = \mathbb{R}^2$ e neste caso $\mathcal{F}(x) \neq T_X(x)$.

Considerando o mesmo ponto $x = \begin{pmatrix} -\sqrt{2} \\ 0 \end{pmatrix}$ para a relaxação do problema anterior:

$$\begin{array}{rc} \min & x_1 + x_2 \\ & x_1^2 + x_2^2 \leq 2 \end{array}$$

Podemos ter sequências viáveis que aproximam $x$ de dois tipos:

1. Ao longo da fronteira de $X$: as sequências viáveis que aproximam $x$ no caso de $x_1^2 + x_2^2 = 2$ continuam válidas.
2. Ao longo de retas pelo interior de $X$.
3. Ao longo de curvas pelo interior de $X$.

1. Sequências que aproximam $x = \begin{pmatrix} -\sqrt{2} & 0 \end{pmatrix}^T$ ao longo de retas pelo interior de $X$:
   - Devem ser do tipo $(z^k)^T = (-\sqrt{2} \quad 0) + \frac{1}{k} w^T$ onde $w^T = (w_1 \quad w_2)$ satisfaz $w_1 \geq 0$.
   - O ponto $z^k$ é viável se $\|z^k\| \leq \sqrt{2}$ o que ocorre se $(-\sqrt{2} + \frac{w_1}{k})^2 + \left(\frac{w_2}{k}\right)^2 \leq 2$ e portanto $k \geq \frac{w_1^2+w_2^2}{2\sqrt{2}w_1}$
2. Sequências que aproximam $x$ ao longo de curvas, pelo interior de $X$: idem, $w_1 \geq 0$ para $k \geq \bar{k}$.
3. Sequências que aproximam $x$ pela fronteira devem ter $w_1 = 0$ e $w_2 \in \mathbb{R}$.

Em resumo $T_X(x) = \{w \in \mathbb{R}^2 : w_1 \geq 0\}$.

Investigação de $\mathcal{F}(x)$ em $x = \begin{pmatrix} -\sqrt{2} & 0 \end{pmatrix}^T$:

$$\begin{array}{ll} \min & x_1 + x_2 \\ & x_1^2 + x_2^2 \leq 2 \end{array}$$

- $\mathcal{F}(x) = \{d \in \mathbb{R}^2 : \nabla f_1(x)^T d \leq 0\} = \{d \in \mathbb{R}^2 : 2(x_1 \quad x_2) \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \leq 0\}$.
- Para o ponto em questão, temos $-2\sqrt{2}d_1 \leq 0 \rightarrow d_1 \geq 0$.

Para este caso $T_X(x) = \mathcal{F}(x)$.

As condições de qualificação de restrições (por exemplo, a regularidade do ponto) visam estabelecer condições que garantam que o conjunto $\mathcal{F}(x)$ ($V(x)$), obtido linearizando a descrição algébrica de $X$, capture os aspectos geométricos essenciais de $X$ nas vizinhanças de $x$, representados por $T_X(x)$ no ponto $x$, candidato a ótimo.

**Teorema - Luenberger p. 325**

Em um ponto regular $x^*$ da superfície $S$ gerada pelo conjunto de restrições $h_i(x) = 0, \; i = 1, \ldots, m$ o plano tangente $T_X(x^*)$ e $V(x^*)$ ou $\mathcal{F}(x^*)$ são idênticos.

## Proposição

*Seja $x^*$ um ponto de mínimo local de $f_0$ sujeito às restrições $h(x) = 0$. Assuma que os gradientes $\nabla h_1(x), \ldots, \nabla h_q(x)$ sejam li em $x^*$.*

1. *Então, existe $\lambda^* = (\lambda_1^*, \ldots, \lambda_q^*)^T$ único, tal que:*

$$\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) = 0$$

2. *Se $f_0$ e $h$ forem duas vezes continuamente diferenciáveis, temos que*

$$y^T \left( \nabla^2 f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla^2 h_i(x^*) \right) y \geq 0, \quad \forall y \in V(x^*) \text{ onde}$$

$$V(x^*) = \{y : \nabla h_i(x^*)^T y = 0, \ \ i = 1, \ldots, q\}$$

*denota o subespaço de variações de primeira ordem de $h$ nulas*

$$\min \; x_1 + x_2$$
$$x_1^2 + x_2^2 = 2$$

- $\nabla^2 f_0(x) = 0_{2\times 2}$ e $\nabla^2 h(x) = Diag(2)$ para todo $x \in \mathbb{R}^n$.
- Para $\overline{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, $\overline{\lambda} = -\frac{1}{2}$ e $d \in V(\overline{x}) = \{d \in \mathbb{R}^2 : d_1 = -d_2\}$.
  $d^T \left(\nabla^2 f_0(\overline{x}) + \overline{\lambda}\nabla^2 h(\overline{x})\right) d = -2d_1^2 \leq 0$.
- A condição de 2a. ordem ($\geq 0$) só se verifica se $d = 0$ e $\overline{x}$ não pode ser ponto de mínimo (pode ser ponto de máximo local).

$$\min \; x_1 + x_2$$
$$x_1^2 + x_2^2 = 2$$

- $\nabla^2 f_0(x) = 0_{2\times 2}$ e $\nabla^2 h(x) = Diag(2)$ para todo $x \in \mathbb{R}^n$.
- Para $\hat{x} = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$, $\hat{\lambda} = \frac{1}{2}$ e $d \in V(\hat{x}) = \{d \in \mathbb{R}^2 : d_1 = -d_2\}$.
  $d^T \left(\nabla^2 f_0(\hat{x}) + \overline{\lambda} \nabla^2 h(\hat{x})\right) d = 2d_1^2 \geq 0$
- $\hat{x}$ satisfaz as condições necessárias de 1a. e 2a. ordem.

Considere a função Lagrangeana $L : \mathbb{R}^{n+q} \to \mathbb{R}$ definida como:

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{q} \lambda_i h_i(x).$$

As condições necessárias de primeira ordem, de viabilidade (primal) e as condições necessárias de segunda ordem para otimalidade de um ponto regular $x^*$ podem ser escritas em termos de $L(x, \lambda)$:

1. $\nabla_x L(x^*, \lambda^*) = 0 \Leftrightarrow \nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) = 0$
2. $\nabla_\lambda L(x^*, \lambda^*) = 0 \Leftrightarrow h_i(x) = 0, \forall i = 1, \ldots, q$
3. $y^T \nabla_{xx}^2 L(x^*, \lambda^*) y \geq 0, \qquad \forall y \in V(x^*) \quad \Leftrightarrow$
$$y^T \left( \nabla^2 f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla^2 h_i(x^*) \right) y \geq 0, \quad \forall y \in V(x^*)$$

A regularidade de $x^*$ implica na existência de multiplicadores de Lagrange

# A função Lagrangeana

$$L(x, \lambda) := f_0(x) + \sum_{i=1}^{q} \lambda_i h_i(x).$$

Todo mínimo local regular deve então satisfazer o sistema de equações em $n + q$ variáveis dado por:

- $\nabla_x L(x^*, \lambda^*) = 0$
- $\nabla_\lambda L(x^*, \lambda^*) = 0$

1. Cabe salientar que nem todo ponto que satisfizer o sistema acima é um ponto de mínimo (recorde nossa experiência com Otimização Irrestrita). Um ponto de máximo pode também satisfazer o sistema acima.
2. Mesmo introduzindo as condições de segunda ordem anteriores, ainda não temos condições suficientes para caracterizar um mínimo local regular.

$$\text{minimize} \quad \frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right)$$
$$x_1 + x_2 + x_3 = 3$$

- Consideremos o vetor $(x^*, \lambda^*) = (1, 1, 1, -1)^T$
- Observe que $\nabla h(x^*) = (1, 1, 1)^T$ (é li), logo o ponto é regular.
- Observe que $\nabla f_0(x^*)^T + \lambda^* h(x^*)^T = (1, 1, 1) - 1(1, 1, 1) = 0$, logo as condições necessárias de primeira ordem são satisfeitas.
- Uma vez que $\nabla^2 f_0(x^*) = \text{Diag}(1)$ e $\nabla^2 h(x) = 0_{3\times 3}$, temos que $y^T \nabla^2 L(x^*, \lambda^*) y > 0, \forall y \in \mathbb{R}^n$, em particular para $y \in V(x^*)$.
- Portanto $x^*$ é um mínimo global para o problema.

$$\text{minimize} \quad -\frac{1}{2}\left(x_1^2 + x_2^2 + x_3^2\right)$$
$$x_1 + x_2 + x_3 = 3$$

- Este problema admite mínimo ?
- Consideremos agora o vetor $(x^*, \lambda^*) = (1, 1, 1, 1)^T$
- $\nabla h(x^*) = (1, 1, 1)^T$ e o ponto é regular.
- Observe que $\nabla f_0(x^*)^T + \lambda^* h(x^*)^T = (-1, -1, -1) + (1, 1, 1) = 0$, logo as condições necessárias de primeira ordem são satisfeitas.
- Uma vez que $\nabla^2 f_0(x^*) = \text{Diag}(-1)$ e $\nabla^2 h(x) = 0_{3\times 3}$, $d \in V(x)$ para qualquer $x$ deve satisfazer $d_1 + d_2 + d_3 = 0$. Fixando $d_3 = -(d_1 + d_2)$, temos que $d^T \nabla^2 L(x^*, \lambda^*) d = -d_1^2 - d_2^2 - (d_1 + d_2)^2 \leq 0$ e as condições necessárias de segunda ordem não são satisfeitas.

## Proposição

**(Condições Suficientes de Otimalidade)**

Assuma que $f_0$ e $h$ sejam funções duas vezes diferenciáveis e seja $x^* \in \mathbb{R}^n$ e $\lambda^* \in \mathbb{R}^q$ satisfazendo:

1. $\nabla_x L(x^*, \lambda^*) = 0, \;\; \nabla_\lambda L(x^*, \lambda^*) = 0$
2. $y^T \nabla^2_{xx} L(x^*, \lambda^*) y > 0$ para todo $y \neq 0, y \in V(x^*)$.

Então:

- $x^*$ é um mínimo local estrito de $f_0$ sujeito a $h(x) = 0$.
- De fato, existem $\gamma > 0$ e $\epsilon > 0$ tais que: $f_0(x) \geq f_0(x^*) + \frac{\gamma}{2}||x - x^*||^2$ para qualquer $x : h(x) = 0$ e $||x - x^*|| \leq \epsilon$.

## Exemplo

$$\begin{aligned} \min \quad & -(x_1x_2 + x_1x_3 + x_2x_3) \\ & x_1 + x_2 + x_3 = 3 \end{aligned}$$

**Função Lagrangeana**

$$L(x, \lambda) = -(x_1x_2 + x_1x_3 + x_2x_3) + \lambda(x_1 + x_2 + x_3 - 3)$$

CNPO: $\nabla_x L(x^*, \lambda^*) = 0, \nabla_\lambda L(x^*, \lambda^*) = 0$ que implica em:

$$\begin{aligned} -x_2^* - x_3^* + \lambda^* &= 0 \\ -x_1^* - x_3^* + \lambda^* &= 0 \\ -x_1^* - x_2^* + \lambda^* &= 0 \\ x_1^* + x_2^* + x_3^* &= 3 \end{aligned}$$

cuja solução única é $(x^*, \lambda^*) = (1, 1, 1, -2)$

$$\nabla^2_{xx} L(x^*, \lambda^*) = \begin{pmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{pmatrix}$$

- $\nabla^2_{xx} L(x^*, \lambda^*)$ (seus autovalores são: -2,1,1) é indefinida, mas....
- Para os vetores $y \in V(x^*), y \neq 0$ temos:
  - $y \in V(x^*)$ temos $y_1 + y_2 + y_3 = 0$.
  - $y^T \nabla^2_{xx} L(x^*, \lambda^*) y = -y_1(y_2 + y_3) - y_2(y_1 + y_3) - y_3(y_1 + y_2) = y_1^2 + y_2^2 + y_3^2 > 0$.
- Para todo $y \in V(x^*) \setminus \{0\}$, $\nabla^2_{xx} L(x^*, \lambda^*) \succ 0$ e as condições suficientes de segunda ordem são verificadas.

1. Teorema de Multiplicadores de Lagrange - Condições necessárias e suficientes para otimalidade de um ponto mínimo regular sob restrições de igualdade.
2. Condições Suficientes de Otimalidade para um ponto de mínimo regular sob restrições de igualdade.

Baseia-se na aplicação do Método de Penalidades, cujas ideias centrais são:

1. Relaxamos as restrições do problema
2. Penalizamos a violação das restrições na função objetivo: adicionamos o produto de cada restrição por uma quantidade positiva (penalidade) e somamos à função objetivo.
3. Aplicamos as condições necessárias de primeira e segunda ordem para um mínimo local regular do problema relaxado/penalizado e
4. Tomamos o limite destas condições, na medida em que as penalidades crescem.

## Função de custo penalizada

$$F^k(x) = f_0(x) + \frac{k}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x - x^*\|^2 \qquad k = 1, 2, \ldots$$

onde:

- $x^*$: é um mínimo local de $f_0(x) : h(x) = 0$.
- O termo $\frac{k}{2}\|h(x)\|^2$ penaliza a violação das restrições
- $\alpha$ é uma quantidade positiva tão pequena quanto se queira.
- O termo $\frac{\alpha}{2}\|x - x^*\|^2$ é inserido para facilitar a prova, garantindo que $x^*$ é um mínimo local estrito de $f_0(x) + \frac{\alpha}{2}\|x - x^*\|^2$ sujeito a $h(x) = 0$.

# Método de Penalidades

- Dado que $x^*$ é um mínimo local do problema, é possível escolher uma vizinhança $\epsilon > 0$ de forma que para todo $x$ viável $(h(x) = 0)$ satisfazendo $x \in S := \{z \in \mathbb{R}^n : \|z - x^*\| \leq \epsilon\}$ tenhamos $f_0(x^*) \leq f_0(x)$.
- Então vamos assumir que $x^k$ resolve o problema

$$\begin{array}{ll} \text{minimize} & F^k(x) \\ & x \in S \end{array}$$

Observe que $x^k$ existe uma vez que $S$ é compacto.

## Primeira parte da prova

Vamos mostrar que $\{x^k\}$ converge para $x^*$, isto é, quando a penalidade cresce a solução do problema penalizado irrestrito aproxima-se de uma solução local regular do problema com restrições.

## Mostrando que $\lim_{k\to\infty} x^k = x^*$

- Para todo $k$ temos:

$$\begin{array}{rcll} F^k(x^k) & = & f_0(x^k) + \frac{k}{2}\|h(x)\|^2 + \frac{\alpha}{2}\|x^k - x^*\| & \\ & \leq & F^k(x^*) & \text{otimal. de } x^k \\ & = & f_0(x^*) & \text{já que } h(x^*) = 0 \end{array}$$

- Como $f_0(x)$ é limitada no compacto $S$, o lado esquerdo da desigualdade precisa ficar limitado quando $k \to \infty$
- Para que isto ocorra, o termo $\frac{k}{2}\|h(x)\|^2$ precisa assumir valor finito quando $k \to \infty$ e então temos $\lim_{k\to\infty} \|h(x)\| = 0$.
- Então todo ponto limite $\overline{x}$ de $\{x^k\}$ satisfaz $h(\overline{x}) = 0$ e

$$f_0(\overline{x}) + \frac{\alpha}{2}\|\overline{x} - x^*\|^2 \leq f_0(x^*)$$

### Mostrando que $\lim_{k\to\infty} x^k = x^*$ (continuação)

- Sabemos que $\overline{x} \in S$ uma vez que $x^k \in S$ para todo $k$, e portanto $f_0(x^*) \leq f_0(\overline{x})$, já que $x^*$ é ótimo local em $S$.
- A desigualdade anterior estabelece $f_0(\overline{x}) + \frac{\alpha}{2}\|\overline{x} - x^*\|^2 \leq f_0(x^*)$ e uma vez que $\frac{\alpha}{2}\|\overline{x} - x^*\|^2 \geq 0$, conclui-se $f_0(\overline{x}) \leq f_0(x^*)$.
- Combinando com $f_0(x^*) \leq f_0(\overline{x})$, temos que ter $\frac{\alpha}{2}\|\overline{x} - x^*\|^2 = 0$. Logo, $\overline{x} = x^*$

# Observações

1. O método de penalidades sugere um algoritmo para obter $\overline{x}$, desde que a penalidade inicial garanta que o mínimo de $F^k$, removido o termo $\frac{\alpha}{2}\|x - x^*\|^2$, seja ilimitado.

2. Note que $x^k$ é um ponto interior a $S$ para $k$ suficientemente grande. Ou seja, para uma penalidade suficientemente grande, o mínimo irrestrito de $f_0(x) + \frac{k}{2}\|h(x)\|^2$ é limitado.

3. Por este motivo, a partir deste valor suficientemente grande de $k$, $x^k$ é um mínimo local do problema irrestrito que consiste em minimizar $F^k(x)$. Ou seja, não precisamos considerar a possibilidade de $x^k$ estar na fronteira de $S$ e aplicar condições para mínimo na fronteira.

Vamos provar o Teorema dos Multiplicadores de Lagrange, aplicando as condições necessárias de primeira e segunda ordem para mínimo local irrestrito de $F^k(x)$, para $k$ suficientemente grande.

# Prova do Teorema dos Multiplicadores de Lagrange (cond. 1a e 2a ordem)

Fazendo $k \to \infty$ e usando as conclusões que deduzimos com o Método das Penalidades, vamos:

1. Obter $\lambda^*$.
2. Provar que para $x^*$ ponto de mínimo regular

$$\nabla f_0(x) + \nabla h(x^*)\lambda^* = 0.$$

3. Provar a condição necessária de 2a ordem, impondo que $\nabla^2 L(x^*, \lambda^*) \succeq 0$ em $V(x^*)$.

- Aplicando as CNPO a $F^k(x^k)$, temos:

$$0 = \nabla F^k(x^k) = \nabla f_0(x^k) + k\nabla h(x^k)h(x^k) + \alpha(x^k - x^*) \quad (2)$$

- Pelo enunciado, temos que $\nabla h(x^*)$ possui posto completo, igual a $q < n$. Por continuidade, existe $k$ suficientemente grande para o qual o posto de $\nabla h(x^k)$ também é $q$. Logo $[\nabla h(x^k)^T \nabla h(x^k)]^{-1}$ existe.
- Pre multiplique (2) por $[\nabla h(x^k)^T \nabla h(x^k)]^{-1} \nabla h(x^k)^T$ e obtenha:

$$kh(x^k) = -[\nabla h(x^k)^T \nabla h(x^k)]^{-1} \nabla h(x^k)^T \left( \nabla f_0(x^k) + \alpha(x^k - x^*) \right)$$

- Sabemos que $k \to \infty$, $x^k \to x^*$ e $kh(x^k)$ converge para um valor finito. Denomine então por

$$\lambda^* = -[\nabla h(x^*)^T \nabla h(x^*)]^{-1} \nabla h(x^*)^T \nabla f_0(x^*)$$

- Tome o limite em (2) quando $k \to \infty$ e obtenha:

$$\nabla f_0(x^*) + \nabla h(x^*)\lambda^* = 0$$

- Aplicando as CNSO a $F^k(x^k)$:

$$\nabla^2 F^k(x^k) = \nabla^2 f_0(x^k) + k\nabla h(x^k)\nabla h(x^k)^T + k\sum_{i=1}^{m} h_i(x^k)\nabla^2 h_i(x^k) + \alpha I \quad (3)$$

é semipositiva definida, para todo $k$ suficientemente grande e $\alpha > 0$.

- Tome qualquer $y \in V(x^*)$ isto é $y : \nabla h(x^*)^T y = 0$ e considere que $y^k$ seja a projeção de $y$ no espaço nulo de $\nabla h(x^k)^T$, isto é:

$$(I - P)y = y^k = y - \nabla h(x^k)[\nabla h(x^k)^T \nabla h(x^k)]^{-1} \nabla h(x^k)^T y$$

- Como $\nabla h(x^k)^T y^k = 0$ e como $\nabla^2 F^k(x^k) \succeq 0$ para $k$ sufic. grande, temos:

$$\begin{array}{lll} 0 & \leq & (y^k)^T \nabla^2 F^k(x^k) y^k \\ 0 & \leq & (y^k)^T \left(\nabla^2 f_0(x^k) + k\sum_{i=1}^m h_i(x^k)\nabla^2 h_i(x^k)\right) y^k + \alpha\|y^k\|^2 \end{array}$$

- Uma vez que:
  - $kh_i(x^k) \to \lambda_i^*$
  - $x^k \to x^*$
  - $\nabla h(x^k)^T y = 0 \to \nabla h(x^*)^T y = 0$

  temos que $y^k \to y$ e então temos:

$$0 \leq y^T \left( \nabla^2 f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y + \alpha \|y\|^2$$

- Uma vez que $\alpha > 0$ pode ser arbitrariamente pequeno, temos:

$$0 \leq y^T \left( \nabla^2 f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y, \ \forall y \in V(x^*)$$

  e a prova está completa.

- O entendimento da demonstração da suficiência das condições auxilia o entendimento da razão pela qual os métodos que iremos estudar funcionam.
- Vamos usar uma função denominada Função Lagrangeana Aumentada que nada mais é que a Função Lagrangena de um Problema de Otimização similar ao problema original, modificada pela introdução de uma penalidade quadrática na violação das restrições $h(x) = 0$.

Antes de demonstrar as condições suficientes, precisamos de alguns resultados auxiliares.

**Teorema de Bolzano-Weierstrass**

Toda sequência limitada no $\mathbb{R}^n$ possui pelo menos um ponto limite (ou equivalentemente, pelo menos uma subsequência convergente).

**Lema Auxiliar**

Sejam $P$ e $Q$ duas matrizes simétricas tais que $Q \succeq 0$ e que $P \succ 0$ no espaço nulo de $Q$. Isto é, $x^T Px > 0$ para qualquer $x : Qx = 0$.
Então existe um escalar $\overline{c}$ tal que $P + cQ \succ 0$ para qualquer $c > \overline{c}$.

Por contradição.

- Assuma então que o resultado enunciado não vale. Então, para qualquer inteiro $k$, existe um vetor $x^k$, $\|x^k\| = 1$ tal que:

$$(x^k)^T P x^k + k(x^k)^T Q x^k \leq 0$$

- Uma vez que $\|x^k\| = 1$, a sequência $\{x^k\}$ é limitada dado que é contida em $\mathcal{B} = \{x \in \mathbb{R}^n : \|x\| \leq 1\}$, e pelo Teorema de Bolzano-Weierstrass, existe uma subsequência $\{x^k\}_{k \in K}$ que converge para algum vetor $\overline{x}$.
- Tomando o limite superior da desigualdade acima temos:

$$\lim_{k \to \infty, k \in K} \sup \left\{ (x^k)^T P x^k + k(x^k)^T Q x^k \right\} \leq 0 \quad \to$$

$$\overline{x}^T P \overline{x} + \lim_{k \to \infty, k \in K} \sup \left\{ k(x^k)^T Q x^k \right\} \leq 0$$

$$\overline{x}^T P \overline{x} + \lim_{k \to \infty, k \in K} \sup \left\{ k (x^k)^T Q x^k \right\} \leq 0$$

- Para que o lado esquerdo seja limitado superiormente ($\leq 0$), temos que ter $(x^k)^T Q x^k \to 0$ (observe que $\|x\| = 1$, e logo $(x^k)^T Q x^k \in [0, \lambda_{max}(Q)]$ e $k \to \infty$).
- Logo $\lim_{k \to \infty, k \in K} (x^k)^T Q x^k = \overline{x}^T Q \overline{x} = 0$ e consequentemente $\overline{x}^T P \overline{x} \leq 0$ para $\overline{x} : Q\overline{x} = 0$
- Pelo enunciado do teorema temos que para $\overline{x} \in \mathcal{N}(Q), \overline{x}^T P \overline{x} > 0$. Portanto, temos uma contradição. $\square$

**Função Lagrangeana Aumentada**

$$L_c(x, \lambda) = f_0(x) + \lambda^T h(x) + \frac{c}{2}\|h(x)\|^2$$

onde $c$ é um escalar positivo.

Observe que:

- A função $L_c(x, \lambda)$ corresponde à função Lagrangeana para o seguinte problema:

$$\begin{aligned} \min \quad & f_0(x) + \frac{c}{2}||h(x)||^2 \\ & h(x) = 0 \end{aligned}$$

- Minimizar este problema equivale a minimizar $f_0(x) : h(x) = 0$.

$$L_c(x, \lambda) = f_0(x) + \lambda^T h(x) + \frac{c}{2} ||h(x)||^2$$

- Vamos considerar que $x^*, \lambda^*$ satisfaçam as condições suficientes para otimalidade local ($\min f_0(x) : h(x) = 0$) enunciadas pelo Teorema
- E vamos investigar $\nabla_x L_c(x^*, \lambda^*)$ e $\nabla^2_{xx} L_c(x^*, \lambda^*)$ diante disto.

$$L_c(x,\lambda) = f_0(x) + \lambda^T h(x) + \frac{c}{2}||h(x)||^2$$

$$\begin{aligned}
\nabla_x L_c(x^*,\lambda^*) &= \\
\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) + c\sum_{i=1}^{q} h_i(x^*)\nabla h_i(x^*) &= \\
\nabla f_0(x^*) + \nabla h(x^*)\lambda^* + c\nabla h(x^*)h(x^*) &= \\
\nabla_x L(x^*,\lambda^*) &= 0
\end{aligned}$$

A última expressão é verdade pois assumimos que $x^*, \lambda^*$ satisfazem as condições de primeira ordem enunciadas no Teorema.

$$L_c(x,\lambda) = f_0(x) + \lambda^T h(x) + \frac{c}{2}||h(x)||^2$$

$$\nabla_x L_c(x^*,\lambda^*) = \nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) + c \sum_{i=1}^{q} h_i(x^*) \nabla h_i(x^*)$$

Então avaliando $\nabla^2_{xx} L_c(x^*,\lambda^*)$:

$$\nabla^2_{xx} L_c(x^*,\lambda^*) =$$

$$\nabla^2 f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla^2 h_i(x^*) + c \sum_{i=1}^{q} h_i(x^*) \nabla^2 h_i(x^*) + c \sum_{i=1}^{q} \nabla h_i(x^*) \nabla h_i(x^*)^T =$$

$$\nabla^2 f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla^2 h_i(x^*) + c \nabla h(x^*) \nabla h(x^*)^T =$$

$$\nabla^2_{xx} L(x^*,\lambda^*) + c \nabla h(x^*) \nabla h(x^*)^T$$

$$\nabla^2_{xx} L_c(x^*, \lambda^*) = \nabla^2_{xx} L(x^*, \lambda^*) + c \nabla h(x^*) \nabla h(x^*)^T$$

Vamos assumir que (condições suficientes descritas no Teorema) para qualquer $y \in \mathbb{R}^n : \nabla h(x^*)^T y = 0$ (ou $y^T \nabla h(x^*) \nabla h(x^*)^T y = 0$) tenhamos:

$$y^T \left( \nabla^2_{xx} L(x^*, \lambda^*) \right) y > 0$$

Observe que:

- Aplicando o Lema Auxiliar, fazendo $P = \nabla^2_{xx} L(x^*, \lambda^*)$ e $Q = \nabla h(x^*) \nabla h(x^*)^T \succeq 0$, existe $\overline{c}$ tal que $\nabla^2_{xx} L_c(x^*, \lambda^*) \succ 0$.
- Usando as condições suficientes de otimalidade para otimização irrestrita, concluímos que para $c > \overline{c}$, $x^*$ é um ótimo local irrestrito de $L_c(\cdot, \lambda^*)$.

- Então $x^*$ é um ótimo local estrito para $L_c(\cdot, \lambda^*)$
- Em particular, existem $\gamma > 0$ e $\epsilon > 0$ tais que

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}||x - x^*||^2, \ \ ||x - x^*|| < \epsilon$$

- Observe que para todo $x$ viável ($h(x) = 0$) temos $L_c(x, \lambda^*) = f_0(x)$ e $\nabla_\lambda L(x^*, \lambda^*) = h(x^*) = 0$
- Finalmente, temos que:

$$f_0(x) = L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}||x - x^*||^2 = f_0(x^*) + \frac{\gamma}{2}||x - x^*||^2 \geq f_0(x^*)$$

para qualquer $x : h(x) = 0$ e $||x - x^*|| < \epsilon$ e a prova está completa.

## Revisando as condições necessárias de otimalidade

$$\begin{aligned} (\mathrm{P})\ \min \quad & f_0(x) \\ & h_i(x) = 0 \quad i = 1, \ldots, q \\ & f_j(x) \leq 0 \quad j = 1, \ldots, m \end{aligned}$$

- Dado $x$ viável, definimos $A(x) = \{j : f_j(x) = 0\}$, o conjunto dos índices de restrições de desigualdades que são justas para $x$.
- Vamos assumir que $x^*$ seja um mínimo local para (P). Neste caso, $x^*$ também deve ser um mínimo local para:

$$\begin{aligned} (\mathrm{PE})\ \min \quad & f_0(x) \\ & h_i(x) = 0 \quad i = 1, \ldots, q \\ & f_j(x) = 0 \quad j \in A(x^*) \end{aligned}$$

as restrições folgadas em $x^*$ ($j \notin A(x^*)$) foram eliminadas.

- Esta observação nos leva a crer que se $x^*$ for um ponto regular para o programa (PE), devem haver multiplicadores $\lambda_i^* : i = 1, \ldots, m$ e $\mu_j^* : j \in A(x^*)$ tais que as CNPO sejam atendidas:

$$\nabla f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j^* \nabla f_j(x^*) = 0$$

- Atribuindo multiplicadores $\mu_j^* = 0$ para as restrições inativas em $x^*$, i.e., $j \notin A(x^*)$ temos:

$$\begin{gathered} \nabla f_0(x^*) + \textstyle\sum_{i=1}^{q} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{m} \mu_j^* \nabla f_j(x^*) = 0 \\ \mu_j^* = 0, j \notin A(x^*) \end{gathered}$$

- Os multiplicadores associados a $f_j(x) \leq 0$: devem ser não negativos.

  Observe que se a restrição $f_j(x) \leq u_j = 0$ for relaxada para $f_j(x) \leq u_j$ para $u_j > 0$, o custo ótimo de P deve diminuir.

  (o multiplicador $\mu_j$ representa o simétrico da taxa de variação de $f_0$ com o aumento de uma unidade de $u_j$).

$$\mu_j \approx -\frac{\Delta f_0 \text{ decorrente de } \delta u_j}{\delta u_j}$$

## Definição

Um vetor viável $x$ é dito regular se:

- $\nabla h_i(x) : i = 1, \ldots, q$ e $\nabla f_j(x) : j \in A(x)$ são linearmente independentes.
- Também dizemos que $x$ é regular no caso em que não existam restrições de igualdade e todas as restrições de desigualdade sejam inativas em $x$.

$$\begin{aligned}
\min \quad & f_0(x) \\
& h_i(x) = 0 \;\; i = 1, \ldots, q \\
& f_j(x) \leq 0 \;\; j = 1, \ldots, m
\end{aligned}$$

$$L(x, \lambda, \mu) = f_0(x) + \sum_{i=1}^{q} \lambda_i h_i(x) + \sum_{j=1}^{m} \mu_j f_j(x)$$

**Teorema - Condições de Karush-Kuhn-Tucker (KKT)**

Seja $x^*$ **um mínimo local regular** para

$$
\begin{aligned}
(P)\ \min \quad & f_0(x) \\
& h_i(x) = 0 \quad i = 1, \ldots, q \\
& f_j(x) \leq 0 \quad j = 1, \ldots, m
\end{aligned}
$$

onde $f_0, h_i, f_j : \mathbb{R}^n \to \mathbb{R}, \forall i, j$ são funções continuamente diferenciáveis. **Então** existem multiplicadores de Lagrange únicos $\lambda^*$ e $\mu^*$ tais que:

$$
\begin{aligned}
\nabla_x L(x^*, \lambda^*, \mu^*) &= 0 & \\
\mu_j^* &\geq 0 & j = 1, \ldots, m \\
\mu_j^* &= 0 & j \notin A(x^*)
\end{aligned}
$$

**Se em adição a isto**, tivermos $f, h_i, f_j \in C^2, \forall i, j$ **então**:

$$
y^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) y \geq 0 \quad \forall y \in \mathbb{R}^n : \quad \begin{aligned} \nabla h_i(x^*)^T y &= 0 \\ \nabla f_j(x^*)^T y &= 0, j \in A(x^*). \end{aligned}
$$

$$\min \quad \frac{1}{2}(x_1^2 + x_2^2 + x_3^2)$$
$$x_1 + x_2 + x_3 \leq -3$$

$$L(x, \mu) = \frac{1}{2}(x_1^2 + x_2^2 + x_3^2) + \mu(x_1 + x_2 + x_3 + 3)$$

- Das CNPO, temos $\nabla L_x(x^*, \lambda^*) = 0 \rightarrow x_i^* + \mu^* = 0, i = 1, 2, 3$
- Precisamos avaliar dois casos:
  - A restrição é folgada e então $\mu^* = 0$. Neste caso $x_1^* = x_2^* = x_3^* = 0$, violando a restrição do problema. Logo a restrição não é folgada.
  - A restrição é justa. Observe então que $x_1^* + x_2^* + x_3^* = -3$ equivale a impor também que (CNPO) $\nabla_\mu L(x^*, \mu^*) = 0$. Assim sendo, temos $x_1^* = x_2^* = x_3^* = -1$ e $\mu^* = 1$

- Analisando as condições de KKT de segunda ordem: $y^T \nabla^2_{xx} L(x^*, \mu^*) y \geq 0$ para qualquer $y : \nabla h(x^*)^T y = 0, \nabla f_j(x^*)^T y = 0, j \in A(x^*)$.
- Como não há restrições de igualdade, temos que a condição torna-se $y^T \nabla^2_{xx} L(x^*, \mu^*) y \geq 0$ para qualquer $y \in \mathbb{R}^3 : y_1 + y_2 + y_3 = 0$.
- Entretanto, como $\nabla^2_{xx} L(x^*, \mu^*) = \mathrm{Diag}(1) \succ 0$ para qualquer $y$ (em particular satisfazendo $y_1 + y_2 + y_3 = 0$) a condição necessária de segunda ordem é atendida.

# Exemplo

$$
\begin{aligned}
\min \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\
& x_1^2 + x_2^2 \leq 5 \\
& 3x_1 + x_2 \leq 6
\end{aligned}
$$

Para encontrar uma solução, devemos:

1. considerar as combinações de restrições ativas e inativas: zero, uma ou todas as duas ativas.
2. Para a escolha considerada, impomos as condições de primeira ordem pertinentes e verificamos se a solução primal-dual satisfaz:
   1. viabilidade (a restrição que foi considerada inativa é satisfeita)
   2. os multiplicadores das restrições ativas são não negativos
   3. as condições de segunda ordem, no plano tangente pertinente.

$$
\begin{aligned}
\min \quad & 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2 \\
& x_1^2 + x_2^2 \leq 5 \\
& 3x_1 + x_2 \leq 6
\end{aligned}
$$

### Zero restrições ativas

- Impondo $\nabla f_0(x) = 0 \rightarrow x = \begin{pmatrix} 0 \\ 5 \end{pmatrix}$, logo a primeira restrição é violada e a segunda é folgada.
- Não precisamos observar a condição de segunda ordem para mínimo irrestrito.

$$\min \quad 2x_1^2 + 2x_1x_2 + x_2^2 - 10x_1 - 10x_2$$
$$x_1^2 + x_2^2 \leq 5$$
$$3x_1 + x_2 \leq 6$$

## Primeira restrição ativa, segunda folgada

- $\mu_2^* = 0$.
- CNPO: $\nabla_x L(x, \mu^*) = 0 \to (x^*, \mu_1)^T = \begin{pmatrix} 1 & 2 & 1 \end{pmatrix}$, o ponto é regular.
- $f_2(x^*) < 0$ (se fosse justa com $\mu_2^* = 0$ seria uma restrição degenerada).
- $V(x^*) = \{y \in \mathbb{R}^2 : \nabla f_1(x^*)^T y = 0\} = \{y \in \mathbb{R}^2 : y_1 + 2y_2 = 0\}$
- $\nabla^2_{xx} L(x^*, \mu^*) = \begin{pmatrix} 4 + 2\mu_1 & 2 \\ 2 & 2 \end{pmatrix} \succ 0 \to$ CNSO são satisfeitas.

**Proposição: Condições Suficientes de Segunda Ordem**

Assuma que $f_0, f, h$ sejam funções duas vezes diferenciáveis e seja $x^* \in \mathbb{R}^n, \lambda^* \in \mathbb{R}^q, \mu^* \in \mathbb{R}^m$ satisfazendo:

$$\begin{array}{ccc} \nabla_x L(x^*, \lambda^*, \mu^*) = 0 & h(x^*) = 0 & f(x^*) \leq 0 \\ \mu_j^* \geq 0 & j = 1, \ldots, m & \\ \mu_j^* = 0 & j \notin A(x^*) & \end{array}$$

$y^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y > 0$ para todo $y$ satisfazendo:

$$\nabla h_i(x^*)^T y = 0, \ \forall i = 1, \ldots, q \quad \text{e} \quad \nabla f_j(x^*)^T y = 0, \forall j \in A(x^*).$$

Assuma também que $\mu_j^* > 0$ para qualquer $j \in A(x^*)$.
Então $x^*$ é um mínino local de $f_0(x)$ restrito a $h(x) = 0, f(x) \leq 0$.

## Re-escrevendo as condições Suficientes de Segunda Ordem

### Proposição: Condições Suficientes de Segunda Ordem

Assuma que $f_0, f, h$ sejam funções duas vezes diferenciáveis e seja $x^* \in \mathbb{R}^n, \lambda^* \in \mathbb{R}^q, \mu^* \in \mathbb{R}^m$ satisfazendo:

$$\begin{array}{ccc} \nabla_x L(x^*, \lambda^*, \mu^*) = 0 & h(x^*) = 0 & f(x^*) \leq 0 \\ \mu_j^* \geq 0 & j = 1, \ldots, m & \\ \mu_j^* = 0 & j \notin A(x^*) & \end{array}$$

$$y^T \nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) y > 0 \qquad \text{para todo } y \text{ satisfazendo:}$$

$$\nabla h_i(x^*)^T y = 0, \ \forall i = 1, \ldots, q \quad \text{e}$$

$$\nabla f_j(x^*)^T y = 0, \forall j \in A(x^*) \text{ tal que } \mu_j > 0$$

Então $x^*$ é um mínino local de $f_0(x)$ restrito a $h(x) = 0, f(x) \leq 0$.

Observe que $\nabla_{xx}^2 L(x^*, \lambda^*, \mu^*) \succ 0$ em um subespaço maior, que contém o plano tangente às restrições ativas em $x^*$.

1. Função Dual Lagrangeana
2. Função perturbação ou função primal
3. Dualidade Fraca
4. Interpretação geométrica da dualidade
5. Dualidade forte, condições de qualificação, condições de Slater
6. Dualidade em Programação Convexa
7. Dualidade *local*

## Par primal-dual de problemas de otimização

Associado a qualquer problema de programação não linear, existe um outro, denominado Problema Dual Lagrangeano, muito proximamente relacionado ao primeiro, dito primal.

- Diante de convexidade do problema primal e de certas condições de qualificação, ambos possuem funções objetivo iguais em suas soluções ótimas (Dualidade Forte).
- Satisfeitas algumas hipóteses adicionais, é possível resolver o primal indiretamente: resolve-se o dual e, posteriormente, recupera-se a solução primal.

## Problema Primal

$$
\begin{aligned}
(P) \qquad & \min f_0(x) & \\
& h_i(x) = 0 & i = 1, \ldots, q \\
& f_j(x) \leq 0 & j = 1, \ldots, m \\
& x \in X &
\end{aligned}
$$

onde $X$ é um subconjunto qualquer do $\mathbb{R}^n$.

## Função Dual Lagrangeana

Associando $v \in \mathbb{R}^q$ às restrições $h(x) = 0$ $u \in \mathbb{R}^m_+$ às restrições $f(x) \leq 0$, a Função Dual Lagrangena é:

$$
\theta(u, v) = \inf \left\{ f_0(x) + \sum_{i=1}^{q} v_i h_i(x) + \sum_{i=1}^{m} u_i f_i(x) : x \in X \right\} \tag{4}
$$

$$(D) \quad \sup_{u \geq 0} \theta(u, v) \tag{5}$$

$$(D) \quad \sup_{u \geq 0} \inf \left\{ f_0(x) + \sum_{i=1}^{q} v_i h_i(x) + \sum_{i=1}^{m} u_i f_i(x) : x \in X \right\}$$

- O ato de incorporar os termos $u^T f(x)$ e $v^T h(x)$ à função objetivo $f_0(x)$ para definir $\theta(u, v)$ é chamado de dualização.
- Associado ao mesmo problema (P), diversos problemas duais Lagrangeanos podem ser construídos, dependendo de quais restrições forem relaxadas e dualizadas.

## Teorema - Dualidade Fraca

Seja $\overline{x}$ uma solução viável de (P), $x^*$ uma solução ótima de (P) e $(u, v)$ um vetor dual viável, isto é: $u \geq 0$.
Então $\theta(u, v) \leq f_0(x^*)$.

## Prova

$$\begin{aligned}
\theta(u, v) &= \inf\{f_0(x) + v^T h(x) + u^T f(x) : x \in X\} \\
&\leq \inf\{f_0(x) + v^T h(x) + u^T f(x) : f(x) \leq 0, h(x) = 0, x \in X\} \\
&\leq \inf\{f_0(x) : f(x) \leq 0, h(x) = 0, x \in X\} \\
&= f_0(x^*) \\
&\leq f_0(\overline{x})
\end{aligned}$$

Recordando: ($\theta(u, v)$ é) é o ínfimo ponto a ponto de uma função afim de $u$ e $v$, parametrizada em $x \in X$. Portanto é côncava.
Supondo $u_1, u_2$ tais que $\theta(u_1), \theta(u_2) > -\infty$

$$\begin{aligned}
\theta(\alpha u_1 + (1-\alpha)u_2, \alpha v_1 + (1-\alpha)v_2) &= \inf \Big\{ f_0(x) + (\alpha u_1 + (1-\alpha)u_2)^T f(x \\
&\qquad + (\alpha v_1 + (1-\alpha)v_2)^T h(x) : x \in X \\
&\geq \inf \Big\{ \alpha f_0(x) + \alpha u_1^T f(x) + \alpha v_1^T h(x) : x \in X \\
&\quad + \inf \Big\{ (1-\alpha) f_0(x) + (1-\alpha) u_2^T f(x) + (1-\alpha) v_2^T h(x) : x \in X \\
&= \alpha\theta(u_1, v_1) + (1-\alpha)\theta(u_2, v_2
\end{aligned}$$

Logo o Problema Dual Lagrangeano é um problema de otimização convexa, mesmo que o problema primal não seja.

**Vamos considerar apenas uma restrição explícita de desigualdade no primal**

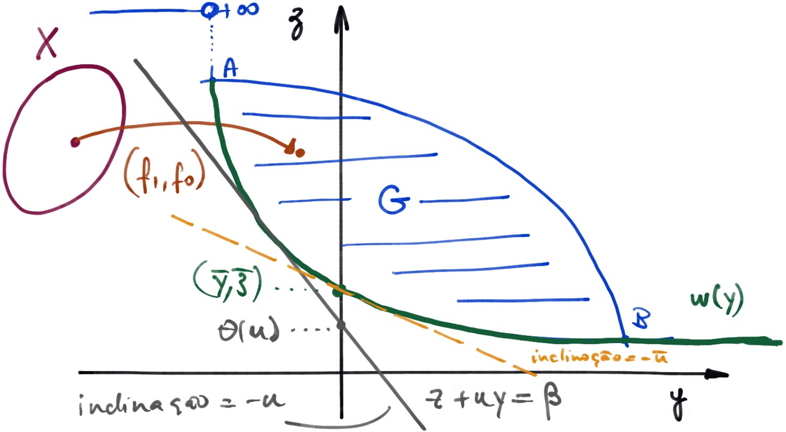
$$\begin{aligned} \min\ & f_0(x) \\ & f_1(x) \leq 0 \\ & x \in X \end{aligned}$$

**Função primal ou função perturbação**

$$w(y) = \inf\left\{f_0(x) : f_1(x) \leq y, x \in X\right\}$$

- Vamos considerar a imagem de $X$ diante do mapeamento $f_1, f_0$: $G = \{(y, z) : y = f_1(x), z = f_0(x), \text{ para algum } x \in X\}$.
- $w(y)$ é o envelope inferior de $G$.
- O problema primal consiste em encontrar, dentre os pontos em $G : y \leq 0$, o de menor ordenada, indicado por $(\bar{y}, \bar{z})$.
- Suponha que um determinado $u \geq 0$ seja dado. Encontrar $\theta(u)$ requer a minimização de $f_0(x) + uf_1(x)$ sobre $x \in X$. Fazendo $z = f_0(x)$ e $y = f_1(x)$, desejamos minimizar $z + uy$ sobre os pontos de $X$. Todos os pontos $(y, z)$ ao longo da reta $z + uy$ possuem o mesmo objetivo. Assim, para encontrar $\theta(u)$, devemos descer a reta de inclinação $-u$ até que suporte $G$.
- O valor de $\theta(u)$ é dado pelo ponto que este hiperplano suporte $G$ com inclinação $-u$ intercepta o eixo $z$ $(y = 0)$.

Seja $f^* = \overline{z}$ o valor ótimo de P e $\theta(\overline{u})$ o valor ótimo do Problema Dual Lagrangeano. A diferença $f^* - \theta(\overline{u})$ é chamada gap de dualidade.



+∞
z
X
A
$(f_1, f_0)$
G
$(\bar{y}, \bar{z})$
$\theta(u)$
w(y)
B
inclinação = $-\bar{u}$
inclinação = $-u$
$z + uy = \beta$
y

## Problema primal

$$\begin{aligned} \min x_1^2 + x_2^2 \\ -x_1 - x_2 + 4 \leq 0 \\ x_1, x_2 \geq 0 \end{aligned}$$

Solução ótima: $x^* = (2 \quad 2)^T$ e $f_0(x^*) = 8$.

## Função Dual Lagrangeana

Dualizando $f_1(x) = -x_1 - x_2 + 4 \leq 0$ com multiplicador $u \in \mathbb{R}_+$ e fazendo $X = \mathbb{R}^2_+$, a função Dual Lagrangeana é:

$$\begin{aligned} \theta(u) &= \inf\{x_1^2 + x_2^2 + u(-x_1 - x_2 + 4) : x \in X\} \\ &= 4u + \inf\{x_1^2 - ux_1 : x_1 \geq 0\} + \inf\{x_2^2 - ux_2 : x_2 \geq 0\} \end{aligned}$$

- Para um $u$ fixo, temos que o ínfimo ocorre em: $x_1 = x_2 = \frac{u}{2}$ se $u \geq 0$ e $x_1 = x_2 = 0$ caso $u < 0$ (dual inviável).
- Então $\theta(u) = \begin{cases} -\frac{1}{2}u^2 + 4u & u \geq 0 \\ 4u & u < 0 \end{cases}$
- Observe que $\theta(u)$ é uma função côncava.
- O máximo $\theta(u)$ ocorre em $\overline{u} = 4$ e $\theta(\overline{u}) = 8$, não havendo gap de dualidade.

A construção de $G$ sob o mapeamento $f_1(x) = -x_1 - x_2 + 4 = y$ e $z = x_1^2 + x_2^2$ para $x \in X = \mathbb{R}^2_+$

Os envelopes inferior $w(y)$ e superior $s(y)$ de $G$ podem ser obtidos resolvendo-se os problemas de otimização:

**$w(y)$**

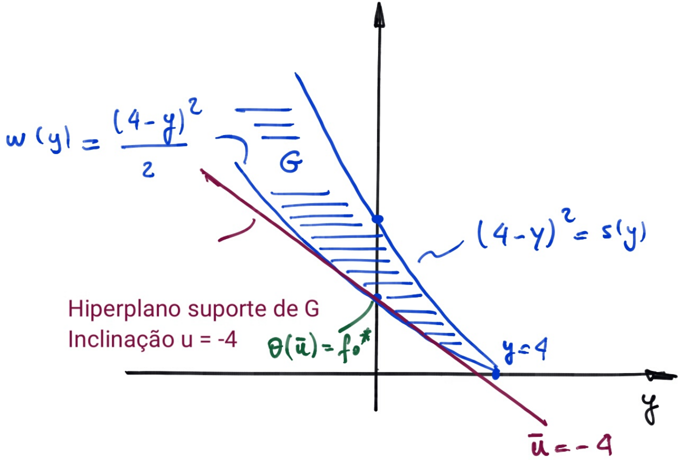
$$\begin{aligned} \min\ & x_1^2 + x_2^2 \\ & -x_1 - x_2 + 4 = y \\ & x_1, x_2 \geq 0 \end{aligned}$$

Solução ótima $w(y) = \frac{(4-y)^2}{2}$ para $y \leq 4$.

**$s(y)$**

$$\begin{aligned} \max\ & x_1^2 + x_2^2 \\ & -x_1 - x_2 + 4 = y \\ & x_1, x_2 \geq 0 \end{aligned}$$

Solução ótima $s(y) = (4-y)^2$ para $y \leq 4$.



w (y) = (4-y)^2 / 2
G
(4-y)^2 = s(y)
Hiperplano suporte de G
Inclinação u = -4
θ(ū) = f₀*
y=4
y
ū = -4

## Problema primal

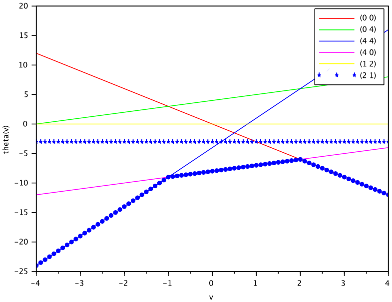
$$
\begin{aligned}
\min f_0(x) &= -2x_1 + x_2 \\
h_1(x) &= x_1 + x_2 - 3 = 0 \\
(x_1, x_2) &\in X
\end{aligned}
$$

onde $X = \{(0,0), (0,4), (4,4), (4,0), (1,2), (2,1)\}$ é um conjunto discreto de pontos (não convexo).
Solução ótima: $x^* = (2 \quad 1)^T$ e $f_0(x^*) = -3$.

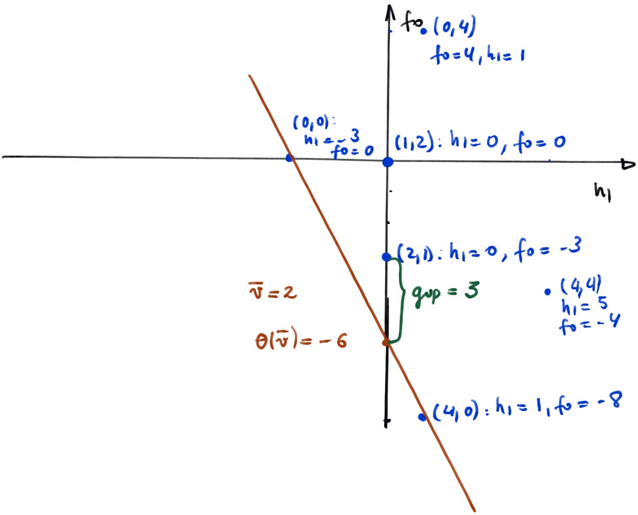
Dualizando $h_1(x) = x_1 + x_2 - 3 = 0$ com multiplicador $v \in \mathbb{R}$ temos:

$$\theta(v) = \inf\{-2x_1 + x_2 + v(x_1 + x_2 - 3) : x \in X\}$$

| $i$ | $x^i \in X$ | $L_i(v)$ | $h_1(x^i)$ |
|---|---|---|---|
| 1 | $(0 \quad 0)^T$ | $L_1(v) = -3v$ | $-3$ |
| 2 | $(0 \quad 4)^T$ | $L_2(v) = 4 + v$ | 1 |
| 3 | $(4 \quad 4)^T$ | $L_3(v) = -4 + 5v$ | 5 |
| 4 | $(4 \quad 0)^T$ | $L_4(v) = -8 + v$ | 1 |
| 5 | $(1 \quad 2)^T$ | $L_5(v) = 0$ | 0 |
| 6 | $(2 \quad 1)^T$ | $L_5(v) = -3$ | 0 |




Temos então que $\theta(v) = \begin{cases} -4+5v & v \leq -1 \\ -8+v & -1 \leq v \leq 2 \\ -3v & v \geq 2 \end{cases}$ cujo máximo ocorre em $\overline{v} = 2$, $\theta(\overline{v}) = -6$. Gap de dualidade $= -3+6 = 3$.



f0
(0,4)
f0=4, h1=1
(0,0):
h1=-3
f0=0
(1,2): h1=0, f0=0
h1
(2,1): h1=0, f0=-3
gap = 3
(4,4)
h1=5
f0=-4
v̄=2
θ(v̄)=-6
(4,0): h1=1, f0=-8

**PPL - $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m, c \in \mathbb{R}^n$**

$$\begin{aligned} \text{(Primal)}\ \ \min\ & c^T x \\ & Ax \leq b \end{aligned}$$

$$L(x,\lambda) = c^T x + \lambda^T(Ax - b) = (c^T + \lambda^T A)x - \lambda^T b \text{ para } \lambda \geq 0$$

$$\theta(\lambda) = \min_{x\in\mathbb{R}^n} L(x,\lambda) = \begin{cases} -\infty & \text{se } c + A^T\lambda \neq 0 \\ -\lambda^T b & \text{se } c + A^T\lambda = 0 \end{cases}$$

$$\begin{aligned} \text{(Dual)}\ \ \max\ & \theta(\lambda) = -\lambda^T b \\ & c + A^T\lambda = 0 \\ & \lambda \geq 0 \end{aligned}$$

$$(P) \quad \min \; x^T x$$
$$Ax = b$$

$$L(x, v) = x^T x + v^T (Ax - b)$$

- Como $L(x, v)$ é convexa, impondo $\nabla_x L(x, v) = 0$, obtemos a expressão $x(v) = -\frac{1}{2} A^T v$ para seu minimizador.
- Resolvendo o dual

$$\max \theta(v) =$$
$$\max \min L(x, v) =$$
$$\max \left(-\frac{1}{2} A^T v\right)^T \left(-\frac{1}{2} A^T v\right) + v^T A \left(-\frac{1}{2} A^T x\right) - v^T b =$$
$$\max \; -\frac{1}{4} v^T A A^T v - v^T b$$

$$\max -\frac{1}{4}v^T AA^T v - v^T b$$

- Resolvendo o problema convexo acima para $v$ e recuperando a solução primal:

$$\begin{aligned}
-\frac{1}{2}AA^T v - b = 0 & \\
v^* &= -2(AA^T)^\dagger b \\
x^* &= A^T(AA^T)^\dagger b
\end{aligned}$$

## Problema não convexo (formulação não linear para $x_i \in \{-1, 1\}$)

$$
\begin{aligned}
\min \ & x^T W x \\
& x_i^2 - 1 = 0 \qquad\qquad i = 1, \ldots, n
\end{aligned}
$$

onde $W \in \mathcal{S}^n$.

- $C^1 = \{i : x_i = 1\}$
- $C^2 = \{i : x_i = -1\}$
- $\{1, \ldots, n\} = C^1 \cup C^2$, $C^1 \cap C^2 = \emptyset$.
- Quando $x_i x_j = -1$, $i$ e $j$ não estão no mesmo conjunto.
- $w_{ij}$ representa o custo de manter o par de items $i$ e $j$ na mesma partição.

$$
\begin{aligned}
\min\ & x^T W x \\
& x_i^2 - 1 = 0 \qquad\qquad i = 1, \ldots, n
\end{aligned}
$$

$$L(x, v) = x^T W x + \sum_{i=1}^{n} v_i (x_i^2 - 1)$$

$$\theta(v) = \inf_{x \in \mathbb{R}^n} x^T (W + \mathrm{Diag}(v)) x - \mathbf{1}^T v = \begin{cases} -\mathbf{1}^T v & \text{se } W + \mathrm{Diag}(v) \succeq 0 \\ -\infty & c.c. \end{cases}$$

## Dual do problema - Problema de Programação Semidefinida

$$
\begin{aligned}
\max\ & -\mathbf{1}^T v \\
& W + \mathrm{Diag}(v) \succeq 0
\end{aligned}
$$

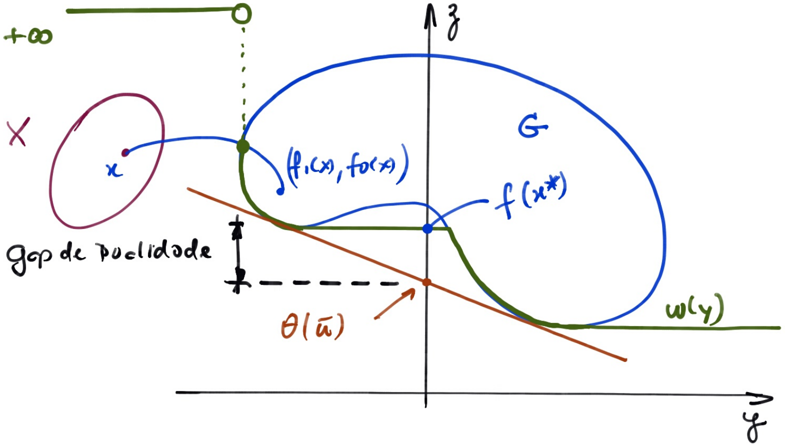
**Dual do problema - Problema de Programação Semidefinida**

$$\begin{aligned} \max \; & -\mathbf{1}^T v \\ & W + \mathrm{Diag}(v) \succeq 0 \end{aligned}$$

- Uma solução viável para o dual é $v = -\lambda_{min}\mathbf{1}$, onde $\lambda_{min}$ é o menor autovalor de $W$.
- Assim sendo, um limite inferior para o valor ótimo $p^*$ do problema primal é

$$p^* \geq -\mathbf{1}^T(-\lambda_{min}\mathbf{1}) = n\lambda_{min}$$

Observação: Resolver um problema de Programação Semidefinida a uma precisão constante é um problema polinomial. Considerando que o problema de particionamento de conjuntos é NP-Completo, a Dualidade Forte não deve valer para este par primal-dual.


+∞
z
X
x
G
(f1(x), f0(x))
f(x*)
gap de Dualidade
θ(ū)
w(y)
y

$$
\begin{aligned}
(P) \quad p^* = \min f_0(x) & \\
h_i(x) = 0 & \qquad i = 1, \ldots, q \\
f_j(x) \leq 0 & \qquad j = 1, \ldots, m \\
(D) \quad d^* = \max_{u \geq 0, v} \theta(u, v) &
\end{aligned}
$$

Dizemos que a dualidade forte se verifica quando $\theta(u^*, v^*) = f_0(x^*)$ para um vetor primal-dual viável $(x^*, u^*, v^*)$.

Condições suficientes para a dualidade forte ser verificada:

- Convexidade do problema de otimização
- Qualificação das restrições: condições de Slater.

**Teorema - Condições de Slater para Problemas Convexos**

Assuma que as funções $f_i(x) : i = 0, \ldots, m$ sejam convexas e que $h(x)$ sejam afim. Assuma ainda as primeiras $k : 0 \leq k \leq m$ restrições $f_i(x) : i = 1, \ldots, m$ sejam afim.

Se existe um ponto $x \in$ rel int$(\mathcal{D})$ tal que:

- $f_i(x) \leq 0 : i = 1, \ldots, k$
- $f_i(x) < 0 : i = k + 1, \ldots, m$
- $h(x) = 0$

então a dualidade forte entre P e D é verificada, isto é $p^* = d^*$. Além disto, se $p^* > -\infty$ (P não é ilimitado), então o valor ótimo é atingido pelo dual, isto é, existem $u^*, v^*$ tais que $d^* = \theta(u^*, v^*) = p^*$.

$\mathcal{D}$ = interseção do domínio de definição das funções $f_i(x) : i = 0, \ldots, m, h_i(x) : i = 1, \ldots, q$.

$$
\begin{aligned}
(P) \;\; p^* &= \min c^T x \\
& Ax \leq b \\
(D) \;\; d^* &= \max \; -\lambda^T b \\
& c + A^T \lambda = 0 \\
& \lambda \geq 0
\end{aligned}
$$

1. Se $X = \{x \in \mathbb{R}^n : Ax \leq b\} \neq \emptyset$ e $p^* > -\infty$, existe $\lambda^* : -b^T \lambda^* = c^T x^*$.
2. Se P é ilimitado ($p^* = -\infty$) então $D$ é inviável, $d^* = -\infty$.
3. Se D é ilimitado ($d^* = \infty$) então $P$ é inviável, $p^* = \infty$.
4. A única quebra da dualidade forte ocorre caso ambos, P e D, sejam inviáveis.

- Na ausência de convexidade, a Teoria de Dualidade é empregada localmente. Os algoritmos
  - Buscam encontrar soluções que anulam o gradiente da função objetivo.
  - Investigam direções de descida $d : \nabla f_0(x^k)^T d < 0$ ou promovem deslocamentos na direção do gradiente.

- A teoria de Dualidade global, mais rica e pontente, na ausência de dualidade, é substituída por uma teoria local de Dualidade, mais fraca, mas bastante útil e que nos ajuda a compreender o funcionamento de algoritmos duais.

- Os algoritmos ditos duais tentam encontrar um multiplicador de Lagrange $\lambda^*$ que nos permita encontrar uma solução primal $x^*$ satisfazendo:

$$\nabla f_0(x^k) + \nabla h(x^*)\lambda^* = 0$$

## Problema Não Convexo

$$\min f_0(x)$$
$$h(x) = 0$$

Nos concentramos em uma solução ótima local $x^*$, regular para o problema acima. As condições suficientes de otimalidade impõem que existe $\lambda^* \in \mathbb{R}^q$ satisfazendo:

1. $\nabla_x L(x^*, \lambda^*) = \nabla f_0(x^*) + \nabla h(x^*)\lambda^* = 0$
2. $\nabla_\lambda L(x^*, \lambda^*) = h(x^*) = 0$
3. $\nabla^2_{xx} L(x^*, \lambda^*) = \nabla^2 f_0(x^k) + \sum_{i=1}^q \lambda_i^* \nabla^2 h_i(x^k) \succ 0$ para todo $y \in V(x^*)$
   (positividade da Hessiana da Lagrangeana no cone tangente de $x^*$).

Vamos assumir a seguinte hipótese, fundamental para o desenvolvimento da Teoria de Dualidade Local:

### Hipótese de convexidade local

Os pontos $x^*, \lambda^*$ satisfazem as condições suficientes de otimalidade local e ainda, que a Hessiana $\nabla^2_{xx} L(x^*, \lambda^*)$ é positiva definida em $\mathbb{R}^n$:

$$L(x^*, \lambda^*) = \nabla^2 f_0(x^*) + \sum_{i=1}^{q} \lambda_i^* \nabla^2 h_i(x^*) \succ 0$$

para todo $y \in \mathbb{R}^n$ e não apenas para $y \in V(x^*)$.

**Consequências da hipótese de convexidade local:**

1. $x^*$ é um mínimo local do problema restrito, mas também é um minimizador local de $L(x, \lambda^*) = f_0(x) + (\lambda^*)^T h(x)$, uma vez que satisfaz as condições de primeira ordem, em conjunto com $\lambda^*$.

2. Para um $\lambda$ suficientemente próximo a $\lambda^*$, o mínimo local $x(\lambda)$ de $f_0(x) + \lambda^T h(x)$ deve ser próximo a $x^*$.
   O Teorema da Função Implícita garante que a solução do sistema não linear em $x$:
   $$\nabla f_0(x) + \nabla h(x)\lambda = 0$$
   é próxima de $x^*$ quando $\lambda$ é próxima de $\lambda^*$, uma vez que $\nabla^2_{xx} L(x^*, \lambda^*)$ é não singular ($\nabla^2_{xx} L(x^*, \lambda^*) \succ 0$ por hipótese)

Localmente, existe uma correspondência entre $\lambda, x$ por meio da resolução do problema irrestrito $\min f_0(x) + \lambda^T h(x)$ e esta corresondência é continuamente diferenciável, pelo Teorema da Função Implícita.

Vamos definir a função dual $\theta(\lambda)$, próximo a $\lambda^*$, como

$$\theta(\lambda) = \min_x \{f_0(x) + \lambda^T h(x)\} \tag{6}$$

onde **a minimização acima se dá nas vizinhanças de $x^*$.**

Vamos mostrar que localmente, a otimização de (6) é equivalente à maximização irrestrita da função dual $\theta$ com relação a $\lambda$.

Vamos assumir que $x(\lambda)$ é a solução única de

$$\min_x \{ f_0(x) + \lambda^T h(x) \}$$

**nas vizinhanças de $x^*$.**

## Lema 1

A função dual $\theta(\lambda)$ possui gradiente (em relação a $\lambda$ !)

$$\nabla_\lambda \theta(\lambda) = h(x(\lambda))$$

## Prova

Como $\theta(\lambda) = \min_x\{f_0(x) + \lambda^T h(x)\}$, em $x(\lambda)$, temos que:

$$\begin{aligned}
\theta(\lambda) &= f_0(x(\lambda)) + \lambda^T h(x(\lambda)) \\
\nabla_\lambda \theta(\lambda) &= \nabla_\lambda x(\lambda) \nabla_x f_0(x(\lambda)) + \nabla_\lambda x(\lambda) \nabla_x h(x(\lambda))\lambda + h(x(\lambda)) \\
&= \nabla_\lambda x(\lambda)(\nabla_x f_0(x(\lambda)) + \nabla_x h(x(\lambda))\lambda) + h(x(\lambda)) \\
&= h(x(\lambda))
\end{aligned}$$

uma vez que $x(\lambda), \lambda$ satisfazem $\nabla_x f_0(x(\lambda)) + \nabla_x h(x(\lambda))\lambda = 0$.

O gradiente da função dual $\theta$ em $\lambda$ é fácil de ser calculado.

## Lema 2

A Hessiana da função dual Lagrangeana é dada por:

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = -\nabla_x h(x(\lambda))^T \left(\nabla^2_{xx} L(x(\lambda))\right)^{-1} \nabla_x h(x(\lambda))$$

## Prova

A Hessiana é a derivada do gradiente. Então, pelo Lema anterior:

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = \nabla_\lambda x(\lambda) \nabla_x h(x(\lambda))$$

Diferenciando $\nabla_x f_0(x(\lambda)) + \nabla_x h(x(\lambda))\lambda = 0$ em relação a $\lambda$:

$$\nabla_\lambda x(\lambda) \left( \nabla^2_{xx} f_0(x(\lambda)) + \sum_{i=1}^{q} \lambda_i \nabla^2_{xx} h_i(x(\lambda)) \right) + \nabla_x h(x(\lambda))^T = 0$$

$$\nabla_\lambda x(\lambda) = -\nabla h(x(\lambda))^T \left( \nabla^2_{xx} f_0(x(\lambda)) + \sum_{i=1}^{q} \nabla^2_{xx} h_i(x(\lambda)) \right)^{-1}$$

## Lema 2

A Hessiana da função dual Lagrangeana é dada por:

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = -\nabla_x h(x(\lambda))^T \left(\nabla^2_{xx} L(x(\lambda))\right)^{-1} \nabla_x h(x(\lambda))$$

## Prova - continua

Substituindo:

$$\nabla_\lambda x(\lambda) = -\nabla h(x(\lambda))^T (\nabla^2_{xx} L(x(\lambda)))^{-1}$$

em

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = \nabla_\lambda x(\lambda) \nabla_x h(x(\lambda))$$

temos

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = -\nabla_x h(x(\lambda))^T \left(\nabla^2_{xx} L(x(\lambda))\right)^{-1} \nabla_x h(x(\lambda))$$

$$\nabla^2_{\lambda\lambda}\theta(\lambda) = -\nabla_x h(x(\lambda))^T \left(\nabla^2_{xx} L(x(\lambda))\right)^{-1} \nabla_x h(x(\lambda))$$

Observações sobre os lemas anteriores:

1. Nas vizinhanças de $x^*$ regular satisfazendo a hipótese de convexidade local, $(\nabla_{xx} L(x(\lambda)))^{-1} \succ 0$.
2. Pela continuidade de $\nabla_x h(x^*)$, $\nabla_x h(x(\lambda))$ possui posto completo $q$ para um $x(\lambda)$ nas vizinhanças de $x^*$.
3. Consequentemente a Hessiana da função dual satisfaz $\nabla^2_{\lambda\lambda}\theta(\lambda) \prec 0$ (é negativa definida em $\mathbb{R}^n$).
4. Logo, localmente, maximizar $\theta(\lambda)$ equivale a minimizar $L(x, \lambda)$. O valor $\theta(\lambda)$ pode não fornecer um limite inferior globalmente válido.

## Teorema

Suponha que o problema $\min\{f_0(x) : h(x) = 0\}$ possua uma solução ótima local regular $x^*$, com correspondente multiplicador de Lagrange $\lambda^*$ e que $f_0(x^*) = r^*$. Suponha que além de $x^*$ ser regular para o conjunto de restrições $h(x) = 0$, a Hessiana da Lagrangeana satisfaça a hipótese de convexidade local, isto é, suponha que $\nabla^2_{xx} L(x^*, \lambda^*)$ seja positiva definida em $\mathbb{R}^n$. Então o problema dual $\max \theta(\lambda)$ possui uma solução ótima local em $\lambda^*$ com correspondente valor ótimo $\theta(\lambda^*) = r^*$.

## Prova

Para $\lambda = \lambda^*$, $\nabla\theta(\lambda^*) = h(x^*) = 0$, logo $\lambda^*$ satisfaz as condições necessárias de primeira ordem para um ponto de máximo local irrestrito de $\theta(\lambda)$. Como $\nabla^2_{\lambda\lambda}\theta(\lambda^*) \prec 0$, as condições suficientes de otimalidade também são satisfeitas.

$$\min f_0(x,y) = -\,xy$$
$$(x-3)^2 + y^2 = 5$$

$$L(x,y,\lambda) = -xy + \lambda((x-3)^2 + y^2 - 5)$$

- Condições necessárias de primeira ordem

$$\begin{aligned} y &= 2(x-3)\lambda \\ x &= 2y\lambda \\ 5 &= (x-3)^2 + y^2 \end{aligned}$$

  Uma solução do sistema é o ponto regular: $x = 4, y = 2, \lambda = 1$, $f_0(4,2) = -8$.

- $\nabla^2_{xx} L(4,2,1) = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \succ 0$. Logo, a hipótese de convexidade local é verificada.

$$L(x, y, \lambda) = -xy + \lambda((x-3)^2 + y^2 - 5)$$

O sistema $\begin{cases} y = & 2(x-3)\lambda \\ x = & 2y\lambda \end{cases}$ permite escrever:

$x(\lambda) = \frac{12\lambda^2}{(4\lambda^2-1)^2}$

$y(\lambda) = \frac{6\lambda}{(4\lambda^2-1)^2}$.

$$\max_{\lambda} \theta(\lambda) = \max_{\lambda} \frac{4\lambda + 4\lambda^3 - 80\lambda^5}{(4\lambda^2-1)^2}$$

que admite uma **solução máxima local** (para $\lambda > \frac{1}{2}$) em $\lambda = 1$ e objetivo $\theta(\lambda) = -8$.

- São métodos que aproximam um problema de otimização com restrições por uma sequência de problemas de otimização irrestritos.
- No caso dos Métodos de Penalidades, a aproximação se dá introduzindo na função objetivo uma penalização pela violação das restrições.
- No caso dos Métodos de Barreira, se dá pela introdução na função objetivo de um termo que favorece os pontos no interior do domínio, em detrimento daqueles na fronteira.
- Associados a estes métodos estão parâmetros $c$ ou $\mu$ que controlam a severidade da penalidade ou da barreira e consequentemente o quanto o problema irrestrito aproxima o problema original.
- Para um problema com $n$ variáveis e $m$ restrições, estes métodos operam diretamente no espaço das $n$ variáveis.

Questões essenciais a investigar:

1. O quão bem o problema irrestrito aproxima o original: o papel de $c \to \infty$ na função de fazer o problema irrestrito aproximar bem o original e fazer convergir a solução do irrestrito para a solução do original.

2. Como resolver o problema irrestrito quando sua função objetivo incorpora um termo de barreira ou penalidade severa. Na medida em que a severidade da penlização cresce, o problema irrestrito torna-se mal condicionado.

3. Encontrar maneiras de contornar as dificuldades de convergência tipicamente observadas quando $c \to \infty$.

## Problema original

$$\min f_0(x)$$
$$x \in X$$

## Problema irrestrito (penalizado)

$$\min f_0(x) + cP(x)$$

onde a função $P(x)$, chamada função penalidade deve:

1. Ser contínua em $\mathbb{R}^n$
2. $P(x) \geq 0, \forall x \in \mathbb{R}^n$
3. $P(x) = 0$ se e somente se $x \in X$

Exemplos de penalidades:

- Se $X = \{x \in \mathbb{R}^n : f_i(x) \leq 0, i = 1, \ldots, m\}$, uma função penalidade muito importante é:

$$P(x) = \frac{1}{2} \sum_{i=1}^{m} (\max\{0, f_i(x)\})^2$$

(A restrição $f_i(x) \leq 0$ pode ser reformulada como $f_i(x) + z_i^2 = 0$)

Embora a função $\max\{0, f_i(x)\}$ seja não diferenciável para os pontos $x : f_i(x) = 0$, a função $(\max\{0, f_i(x)\})^2 \in C^1$, i.e., possui primeiras derivadas contínuas.
Como exemplo a função $f_j(x) = x - a \leq 0$ e a função $P(x) = (\max\{0, x - a\})^2$:

- Se $x < a$, $P(x) = 0$, $\frac{dP}{dx} = 0$.
- Se $x \to a_+$, $\frac{dP}{dx} = (x - a)$, $\lim_{x \to a_+}(x - a) = 0_+$, de forma que $\frac{dP}{dx}$ é contínua em $x = 0$.

O procedimento para resolver $\min f_0(x) : x \in X$ onde $P(x)$ é uma função penalidade para $X$ é o seguinte.

Seja $\{c_k\} : k = 1, 2, \ldots,$ uma sequência escalar que tenda para $\infty$, $c_k \geq 0$, $c_{k+1} > c_k$ para qualquer $k$.

1. Defina a função
$$q(c, x) = f_0(x) + cP(x)$$
2. Obtenha $x^k$ que resolve
$$\begin{aligned} \min\ & q(c_k, x) \\ & x \in \mathbb{R}^n \end{aligned} \tag{7}$$
3. Assumimos que para todo $k$ o problema (7) admite uma solução $x^k$.

**Lema 1**

Pela definição de $x^k$ e considerando que $c_{k+1} > c_k$ temos:

1. $q(c_{k+1}, x^{k+1}) \geq q(c_k, x^k)$
2. $P(x^k) \geq P(x^{k+1})$
3. $f_0(x^{k+1}) \geq f_0(x^k)$

Prova (Lema 1):

1.
$$\begin{aligned} q(c_{k+1}, x^{k+1}) &= f_0(x^{k+1}) + c_{k+1} P(x^{k+1}) \\ &\geq f_0(x^{k+1}) + c_k P(x^{k+1}) \\ &\geq f_0(x^k) + c_k P(x^k) \\ &= q(c_k, x^k) \end{aligned}$$

2.
$$\begin{aligned} f_0(x^k) + c_k P(x^k) &\leq f_0(x^{k+1}) + c_k P(x^{k+1}) \\ f_0(x^{k+1}) + c_{k+1} P(x^{k+1}) &\leq f_0(x^k) + c_{k+1} P(x^k) \\ P(x^{k+1})(c_{k+1} - c_k) &\leq P(x^k)(c_{k+1} - c_k) \end{aligned}$$

3. $f_0(x^{k+1}) + c_k P(x^{k+1}) \geq f_0(x^k) + c_k P(x^k)$. Uma vez que $P(x^{k+1}) \leq P(x^k)$, temos que $f_0(x^{k+1}) \geq f_0(x^k)$.

**Lema 2**

Seja $x^*$ a solução de min $f_0(x) : x \in X$. Então para todo $k$:

$$f_0(x^*) \geq q(c_k, x^k) \geq f_0(x^k)$$

**Prova - Lema 2**

$$\begin{aligned} f_0(x^*) &= f_0(x^*) + c_k P(x^*) \\ &\geq f_0(x^k) + c_k P(x^k) \quad = q(c_k, x^k) \\ &\geq f_0(x^k) \end{aligned}$$

**Teorema - Convergência global do método**

Seja $\{x^k\}$ a sequência gerada pelo Método de Penalidades.
Então, qualquer ponto limite $\overline{x}$ da sequência é uma solução para
$\min\ f_0(x) : x \in X$.

Para a prova, considere que a subsequência $\{x^k\} : k \in \mathcal{K}$ seja uma subsequência convergente de $\{x^k\}$ com ponto limite $\overline{x}$ e seja $f_0^* = f_0(x^*)$ o valor da função objetivo ótima.

- Pela continuidade de $f_0$: $\lim_{k \in \mathcal{K}} f_0(x^k) = f_0(\overline{x})$
- Pelo Lemas 1 e 2 $\{q(c_k, x^k)\}$ é uma sequência não decrescente e limitada superiormente por $f_0^*$:

$$\lim_{k \in \mathcal{K}} q(c_k, x^k) = \overline{q} \leq f_0^*$$

- Subtraindo a penúltima da última expressão:

$$\lim_{k \in \mathcal{K}} c_k P(x^k) = \overline{q} - f_0(\overline{x})$$

- Como $P(x^k) \geq 0$ e $c_k \to \infty$, isto implica em $\lim_{k \in \mathcal{K}} P(x^k) = 0$
- Pela continuidade de $P(x)$, $P(\overline{x}) = 0$ e logo $\overline{x} \in X$ (é viável).
- A otimalidade de $\overline{x}$ decorre da continuidade de $f_0$ e de que, pelo Lema 2, $f_0(\overline{x}) = \lim_{k \in \mathcal{K}} f_0(x^k) \leq f_0^*$ (se $f_0(\overline{x}) \neq f_0^*$ temos uma contradição à otimalidade de $x^*$).

**Problema a resolver**

$$\min f_0(x) = x_1^2 + x_2^2$$
$$x_1 + x_2 - 1 = 0$$

**Problema perturbado**

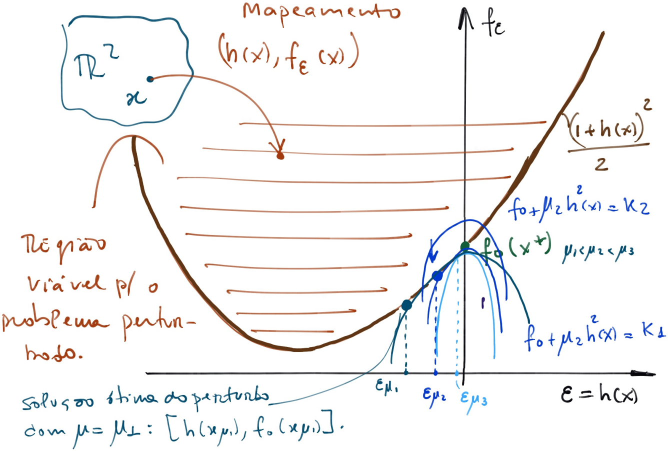
$$\min f_\epsilon(x) = x_1^2 + x_2^2$$
$$x_1 + x_2 - 1 = \epsilon$$

Substituindo $x_2 = \epsilon + 1 - x_1$ em $f_\epsilon(x)$, $f_\epsilon(x_1) = x_1^2 + (\epsilon + 1 - x_1)^2$.
Logo:

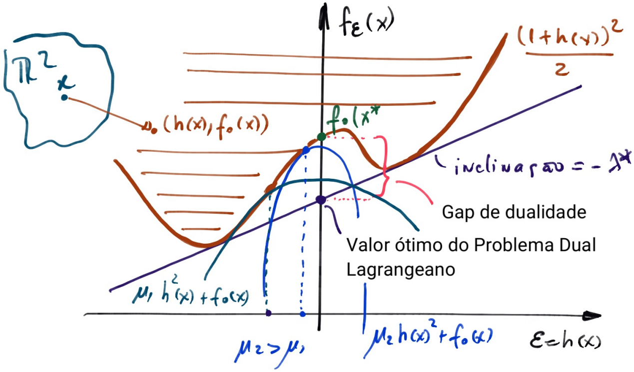
1. O mínimo do problema perturbado ocorre em em $x_1 = x_2 = \frac{1+\epsilon}{2}$ e é dado por $\frac{(1+\epsilon)^2}{2} = \frac{(1+h(x))^2}{2}$.
2. Para qualquer $\epsilon \in \mathbb{R}$, $\sup f_\epsilon(x) = +\infty$.

Vamos investigar o mapeamento $(\epsilon, f_\epsilon)$ que transforma um $x \in \mathbb{R}^2$ em um ponto em $\mathbb{R}^2$.

Ilustração do caso onde o mapeamento $(h(x), f_\epsilon(x))$ é convexo.



Diferentemente da função dual Lagrangeana que possui suporte linear, a interseção de $f_0(x) + \mu\|h(x)\|^2$ com a curva $\frac{(1+h(x))^2}{2}$ consegue se aproximar de $f_0^*$ tanto quanto se deseje, fazendo-se $\mu \to \infty$, uma vez que a penalidade quadrática possui suporte não linear.



**Problema penalizado irrestrito**

Sem perda de generalidade, vamos assumir que desejamos resolver:

$$\begin{aligned} &\min f_0(x) \\ &\quad h_i(x) = 0 \qquad\qquad i = 1, \ldots, q \end{aligned}$$

por uma sequência de problemas $\min q(c_k, x) = f_0(x) + \frac{c_k}{2}\|h(x)\|^2$, parametrizados por $c_k \to \infty$.

A Hessiana $Q(x^k)$ de $q(c, x)$ avaliada em $x^k$ é

$$Q(x^k) = \nabla^2_{xx} L(x, \lambda) + c_k \nabla h(x^k) \nabla h(x^k)^T$$

onde $\nabla^2_{xx} L(x, \lambda)$ é a Hessiana da Lagrangeana $L(x, \lambda) = f_0(x) + \lambda^T h(x)$.

$$Q(x^k) = \nabla^2_{xx} L(x, \lambda) + c_k \nabla h(x^k) \nabla h(x^k)^T$$

$$\nabla^2_{xx} L(x, \lambda) = \nabla^2 f_0(x^k) + \sum_{i=1}^{q} \lambda_i \nabla^2 h_i(x^k)$$

- A estrutura de autovalores de $Q(x^k)$ indica que a otimização de $q(c_k, x)$ torna-se mais difícil quando $c_k$ aumenta.
- Assumindo que $\nabla h(x^k)$ possua posto completo $q$ no minimizador $x^*$, pela continuidade da Hessiana, para um valor $k \geq \overline{k}$, $Q(x^k)$ terá $q$ autovalores tendendo a $+\infty$ e $n - q$ autovalores que embora dependam de $c_k$ são bem comportados.
- É de se esperar que um Método do Gradiente Puro tenha convergência sublinear para valores $c_k$ muito elevados.

Alternativas para a resolução dos subproblemas:

1. Método de Newton (ou alguma de suas variantes)
   A ordem de convergência quadrática, nas vizinhanças de $x^k$ minimizador de $q(c_k, x)$, é invariante quanto à estrutura de autovalores. Porém, o malcondicionamento pode dificultar a obtenção da direção de Newton. Desde que a direção possa ser determinada com precisão, tomando-se cuidados adequados na resolução do sistema linear associado, o método apresenta boas propriedades de convergência.
2. Método do Gradiente Conjugado.
   Sendo $q$ o posto de $\nabla h(x^k)\nabla h(x^k)^T$ (nas vizinhanças de $x^*$), e quando $q << m$, o método do Gradiente Conjugado é uma excelente alternativa, desde que implementada re-inicialização, a cada $q+1$ passos. (Ver Seção 9.5 - Partial Conjugate Gradient Methods, em Luenberger, 3a. Edição). A taxa de convergência independerá de $c_k$.

## Problema original

$$\min f_0(x)$$
$$x \in X$$

$X$ deve ter interior não vazio.

## Problema irrestrito com termo Barreira

$$\min f_0(x) + \frac{1}{c} B(x) \quad \text{ou} \quad \min f_0(x) + \mu B(x)$$

onde $c > 0$ ($\mu > 0$) e a função $B(x)$, chamada função Barreira, deve:

1. Ser contínua no interior de $X$
2. $B(x) \geq 0, \forall x \in X$
3. $B(x) \to \infty$ quando $x \to bd(X)$.

- Também denominados de Métodos Interiores ou Métodos de Pontos Interiores.
- O conjunto de viabilidade $X$ deve possuir interior não vazio.
- Qualquer ponto na fronteira de $X$ deve ser alcançável por algum caminho que percorre o interior de $X$, isto é, $X$ deve ser robusto.
- Este tipo de condição é normalmente satisfeita quando $X = \{f_i(x) \leq 0 : i = 1, \ldots, m\}$ é um conjunto de restrições de desigualdade.
- As restrições de igualdade, caso existentes na definição de $X$, devem ser tratadas por uma função de penalidade ou explicitamente nos subproblemas, via SQP - Sequential Quadratic Programming, ou via Método de Newton, quando o conjunto de restrições de igualdade, $h(x) = 0$ é um conjunto afim.
- Não se associa barreira à restrição de igualdade.

Assumimos que $X = \{x \in \mathbb{R}^n : f_i(x) \leq 0 : i = 1, \ldots, m\}$

## Exemplo de barreiras

1. $B(x) = -\sum_{i=1}^{m} \frac{1}{f_i(x)}$
2. Função Barreira Logarítmica:

$$B(x) = -\sum_{i=1}^{m} \ln(-f_i(x))$$

O procedimento para resolver $\min f_0(x) : x \in X$ onde $B(x)$ é uma função barreira para $X$ é o seguinte. Seja $\{c_k\} : k = 1, 2, \ldots,$ uma sequência escalar que tenda a $\infty$ tal que $c_k > 0$, $c_{k+1} > c_k$ para qualquer $k$.

1. Defina a função
$$r(c, x) = f_0(x) + \frac{1}{c_k} B(x)$$
2. Obtenha $x^k$ que resolve
$$\begin{aligned} \min\ & r(c_k, x) \\ x \in\ & \text{interior de } X \end{aligned} \tag{8}$$
3. Assumimos que para todo $k$ o problema (8) admite uma solução $x^k$.
4. Se o algoritmo de otimização irrestrita for implementado cuidadosamente, a restrição $x \in$ interior de $X$ não precisa ser explicitamente considerada, uma vez que a barreira afasta a solução $x^k$ da fronteira.

Comportamento das funções envolvidas
(resultados "análogos" aos obtidos para o Método de Penalidades)

**Lema 1**

$$r(c_{k+1}, x^{k+1}) \leq r(c_k, x^k)$$
$$B(x^{k+1}) \geq B(x^k)$$
$$f_0(x^{k+1}) \leq f_0(x^k)$$

Demonstração - Parte (1)

**$r(c_{k+1}, x^{k+1}) \leq r(c_k, x^k)$**

$$\begin{aligned} r(c_{k+1}, x^{k+1}) &= f_0(x^{k+1}) + \frac{1}{c_{k+1}} B(x^{k+1}) \\ &\leq f_0(x^k) + \frac{1}{c_{k+1}} B(x^k) \\ &\leq f_0(x^k) + \frac{1}{c_k} B(x^k) \\ &= r(c_k, x^k) \end{aligned}$$

Demonstração - Parte (2)

**$B(x^k) \leq B(x^{k+1})$**

$$f_0(x^k) + \frac{1}{c_k} B(x^k) \leq f_0(x^{k+1}) + \frac{1}{c_k} B(x^{k+1})$$

$$f_0(x^{k+1}) + \frac{1}{c_{k+1}} B(x^{k+1}) \leq f_0(x^k) + \frac{1}{c_{k+1}} B(x^k)$$

$$B(x^{k+1}) \left( \frac{1}{c_{k+1}} - \frac{1}{c_k} \right) \leq B(x^k) \left( \frac{1}{c_{k+1}} - \frac{1}{c_k} \right)$$

$$B(x^{k+1}) \left( \frac{1}{c_k} - \frac{1}{c_{k+1}} \right) \geq B(x^k) \left( \frac{1}{c_k} - \frac{1}{c_{k+1}} \right)$$

Observe que $\left( \frac{1}{c_k} - \frac{1}{c_{k+1}} \right) \geq 0$ e então $B(x^{k+1}) \geq B(x^k)$

Demonstração - Parte (3)

**$f_0(x^{k+1}) \leq f_0(x^k)$**

$$f_0(x^k) + \frac{1}{c_{k+1}} B(x^k) \geq f_0(x^{k+1}) + \frac{1}{c_{k+1}} B(x^{k+1})$$

$$f_0(x^k) - f_0(x^{k+1}) \geq \frac{1}{c_{k+1}} (B(x^{k+1}) - B(x^k)) \geq 0$$

$$f_0(x^k) \geq f_0(x^{k+1})$$

Comportamento das funções envolvidas
(resultados análogos ao do Método de Penalidades)

**Lema 2**

Seja $x^*$ a solução ótima de $\min f_0(x) : x \in X$. Então, para qualquer $k$:

$$r(c_k, x^k) \geq f_0(x^k) \geq f_0(x^*)$$

Como $x^k \in$ interior de $X$, $x^k$ é viável. Então temos:
$r(c_k, x^k) = f_0(x^k) + \frac{1}{c_k} B(x^k) \geq f_0(x^k) \geq f_0(x^*)$

**Teorema - convergência de $\{x^k\}$ para $x^*$**

Qualquer ponto limite $\overline{x}$ da sequência gerada pelo Método das Barreiras é uma solução de $\min\ f_0(x) : x \in X$.

Para a prova, considere que a subsequência $\{x^k\} : k \in \mathcal{K}$ seja uma subsequência convergente da sequência $\{x^k\}$ gerada pelo Método das Barreiras, com ponto limite $\overline{x}$ e seja $f_0^* = f(x^*)$ o valor da função objetivo ótima.

- Pela continuidade de $f_0$: $\lim_{k \in \mathcal{K}} f_0(x^k) = f_0(\overline{x})$
- Pelo Lemas 1 e 2 $\{r(c_k, x^k)\}$ é uma sequência decrescente e limitada inferiormente por $f_0^*$:

$$\lim_{k \in \mathcal{K}} r(c_k, x^k) = \overline{r} \geq f_0^*$$

- Subtraindo a penúltima da última expressão:

$$\lim_{k \in \mathcal{K}} \frac{1}{c_k} B(x^k) = \overline{r} - f_0(\overline{x}) \leq f_0^* - f_0(\overline{x})$$

- Como $c_k \geq 0, B(x^k) \geq 0$ temos, por um lado, $0 \leq f_0^* - f(\overline{x})$. Por outro, pela otimalidade de $x^*$, temos que $f_0^* - f(\overline{x}) \leq 0$. Logo, $f_0^* = f(\overline{x})$ e o resultado segue pela viabilidade de $\overline{x}$.

- Não se associa uma barreira à uma restrição na forma de igualdade.
  As restrições de igualdade devem ser tratadas por um termo de penalidade ou, no Método de Barreiras, devem ser explicitamente consideradas no subproblema do método (isto é, as soluções do subproblema serão restritas ao interior relativo de $X$.) Assim sendo, são necessários métodos específicos para a resolução destes subproblemas.
- Vamos investigar como tratar restrições de desigualdades para o método de Penalidades. Assumimos a partir de agora que qualquer restrição na forma de igualdade foi substituída por duas desigualdades e que o problema a ser resolvido é:

$$\begin{aligned} &\min f_0(x) \\ &\quad f_i(x) \leq 0 \qquad\qquad i = 1, \ldots, m \end{aligned}$$

$$\min f_0(x)$$
$$f_i(x) \leq 0 \qquad\qquad i = 1, \ldots, m$$

- A penalidade só deve ser incorrida se houver violação da restrição. Assim, é comum definir, para todo $i = 1, \ldots, m$ a função

$$f_i^+(x) := \max\{0, f_i(x)\}$$

de forma que $f^+(x) := \begin{pmatrix} f_1^+(x) & f_2^+(x) & \ldots & f_m^+(x) \end{pmatrix}^T$ representa o vetor $m-$dimensional destas entradas.
- Definimos então $P(x) := \gamma(f^+(x))$, onde $\gamma : \mathbb{R}^n \to \mathbb{R}$ é uma função contínua, escolhida de forma que $P(x)$ tenha as propriedades necessárias de uma função de penalidades.

## Exemplos de funções $\gamma$

1. $P(x) = \frac{1}{2}\|f^+(x)\|^2$, isto é, $\gamma(y) = \frac{1}{2}\|y\|^2$.
2. Se $B \in \mathcal{S}^m_{++}$, $P(x) = f^+(x)^T B f^+(x)$, ou seja $\gamma(y) = y^T B y$

## Propriedade desejável de $P(x) := \gamma(f^+(x))$

Já assumimos desde o início que $\{f_i : i = 0, 1, \ldots, m\}$ são funções diferenciáveis. Assim sendo, idealmente, $P(x)$ deve ser uma função diferenciável, uma vez que o problema $\min f_0(x) + c_k P(x)$ será resolvido várias vezes e, para tanto, é conveniente assumir que $P \in C^1$ e que podemos empregar algoritmos baseados na disponibilidade de

$$\nabla f_0(x) + c_k \nabla P(x).$$

Restrições em $\gamma$ para garantir que $P(x) = \gamma(f^+(x)) \in C^1$ (diferenciabilidade de $P(x)$)

Definimos $\nabla f_i^+(x) = \begin{cases} \nabla f_i(x) & x : f_i(x) \geq 0 \\ 0 & x : f_i(x) < 0 \end{cases}$

- $\nabla f_i^+(x)$ é normalmente uma função descontínua em $x : f_i(x) = 0$.

- Para $y \in \mathbb{R}^m, \nabla\gamma = (\nabla\gamma_1, \ldots, \nabla\gamma_m)^T \in \mathbb{R}^m$, vamos impor que $\gamma(y)$ satisfaça a seguinte propriedade

$$y_i = 0 \rightarrow \nabla\gamma_i = 0 \tag{9}$$

- Diante desta hipótese sobre $\gamma$, o vetor de derivadas de $\gamma$ em relação a $x$, $\nabla_x\gamma(f^+(x))$, é uma função contínua, sendo dado por:

$$\nabla_x\gamma(f^+(x)) = \nabla f(x)\nabla\gamma(f^+(x))$$

$$\nabla_x \gamma(f^+(x)) = \nabla f(x) \nabla \gamma(f^+(x))$$

- $\nabla f_i(x)$ pode substituir $\nabla f_i^+(x)$ porque é multiplicado por $\nabla \gamma(f_i^+(x))$ satisfazendo a propriedade (9) ($f_i^+(x) = 0 \rightarrow \nabla \gamma(f^+(x))_i = 0$). Embora haja a descontinuidade em $x : f_i(x) = 0$, ela é regularizada pela propriedade exigida em $\gamma$.
- Esta propriedade é satisfeita, por exemplo, por $\gamma(y) = \frac{1}{2}\|y\|^2$, ou $P(x) = \frac{1}{2}\|f^+(x)\|^2$.

Uma vez que

$$\nabla_x \gamma(f^+(x)) = \nabla f(x) \nabla \gamma(f^+(x))$$

- A solução $x^k$ de min $f_0(x) + c_k \gamma(f^+(x))$ satisfaz as condições necessárias de primeira ordem para um problema irrestrito:

$$0 = \nabla f_0(x^k) + c_k \nabla_x \gamma(f^+(x)) = \nabla f_0(x^k) + c_k \nabla f(x^k) \nabla \gamma(f^+(x^k))$$

- que pode ser re-escrita como

$$0 = \nabla f_0(x^k) + \nabla f(x^k) \lambda^k$$

$$\text{onde } \lambda^k := c_k \nabla \gamma(f^+(x^k))$$

Associado a cada $x^k$, existe um multiplicador de Lagrange $\lambda^k$, obtido após a resolução do problema penalizado, irrestrito.

É um dos métodos de uso geral em PNL mais efetivos.

$$
\begin{aligned}
\min \qquad & f_0(x) \\
& h(x) = 0 \\
& x \in X
\end{aligned}
$$

Vamos considerar o caso onde $X = \mathbb{R}^n$.

- Função Lagrangena Aumentada associada, para $c > 0$:

$$L_c(x, \lambda) = f_0(x) + \lambda^T h(x) + \frac{c}{2}||h(x)||^2$$

- Como já vimos, o parâmetro $c$ controla a severidade da penalização da violação das restrições.

Dois mecanismos são normalmente usados para, através da minimização irrestrita de $L_c(\cdot, \lambda)$ fornecer pontos próximos a $x^*$, um minimizador (local) de $f_0(x) : h(x) = 0$.

1. Fazemos $\lambda = \lambda^*$. Conforme mostramos, se $c$ é maior que um valor $\overline{c}$, então existem $\gamma > 0$, $\epsilon > 0$ tais que:

$$L_c(x, \lambda^*) \geq L_c(x^*, \lambda^*) + \frac{\gamma}{2}||x - x^*||^2, \forall x : ||x - x^*|| < \epsilon$$

   Este resultado sugere que se $\lambda \approx \lambda^*$, uma razoável aproximação de $x^*$ deve ser obtida via minimização irrestrita de $L_c(\cdots, \lambda)$.

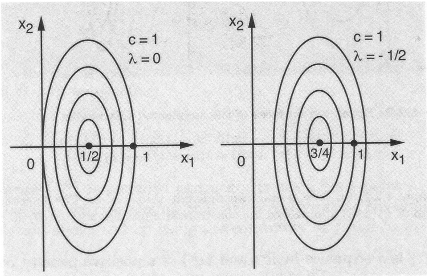
2. Tomamos um valor bastante alto de $c$. Nestes casos, o custo da inviabilidade é elevado e então o mínimo irrestrito de $L_c(\dot{,}\lambda)$ deve ser quase viável. Uma vez que $L_c(x, \lambda) = f_0(x)$ para todo $x$ viável, devemos esperar $L_c(x, \lambda) \approx f_0(x)$ para $x$ próximo da viabilidade.

# Exemplo - abordagem 1

$$\min \quad \frac{1}{2}\left(x_1^2 + x_2^2\right)$$
$$x_1 = 1$$

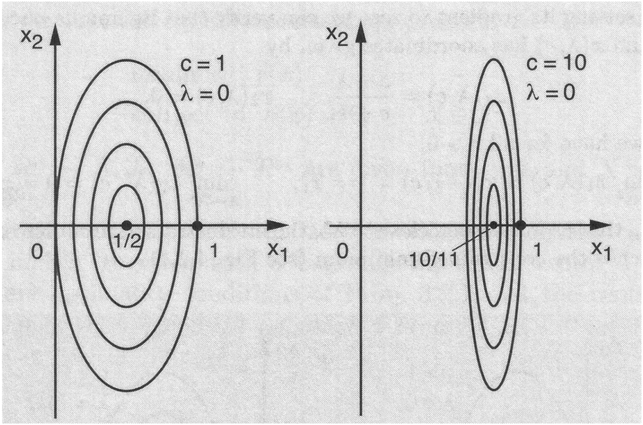
onde o vetor primal-dual ótimo é dado por $(x^*, \lambda^*) = (1, 0, -1)$.

- $L_c(x, \lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2$, para $c > 0$.
- Impondo $\nabla_x L_c(x, \lambda) = 0$, o mínimo irrestrito possui coordenadas:
  - $x_1(c, \lambda) = \frac{c-\lambda}{c+1}$. Logo para $c > 0$, temos $\lim_{\lambda \to \lambda^*} x_1(\lambda, c) = x_1(-1, c) = 1$. Logo $\lim_{\lambda \to \lambda^*} x_2(\lambda, c) = 0$
  - $x_2(\lambda, c) = 0$.



$L_c(x,\lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2.$

- $x_1(c,\lambda) = \frac{c-\lambda}{c+1}$. Logo $\lim_{c\to\infty} \frac{c-\lambda}{c+1} = 1 = x_1^*$
- $x_2(c,\lambda) = 0$. Logo $\lim_{c\to\infty} x_2(c,\lambda) = 0 = x_2^*$.



$$\min \quad \frac{1}{2}\left(x_1^2 + x_2^2\right)$$
$$x_1 = 1$$

- $$L_c(x,\lambda) = \frac{1}{2}\left(x_1^2 + x_2^2\right) + \lambda(x_1 - 1) + \frac{c}{2}(x_1 - 1)^2$$
- $$\nabla^2_{xx} L_c(x,\lambda) = I + c\begin{pmatrix} 1 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & 0 \end{pmatrix} = \begin{pmatrix} 1+c & 0 \\ 0 & 1 \end{pmatrix}$$

Autovalores de $\nabla^2_{xx} L_c(x,\lambda)$ : $1 + c$ e $1$. Quando $c \to \infty$, o problema torna-se malcondicionado.

- A cada iteração $k+1$, fazemos $\lambda^{k+1} \leftarrow \lambda^k + c^k h(x^k)$, como uma melhor aproximação para o vetor de multiplicadores ótimo. Então minimizamos $L_c(\cdot, \lambda^{k+1})$. O método baseado nesta aproximação de multiplicadores é chamado de *Método dos Multiplicadores*.
- Para evitar o malcondicionamento, recomenda-se empregar o Método de Newton para minimizar $L_c(\cdot, \lambda^k)$.
- Além disto, é usual empregar $x^k$ como ponto de partida para a minimização irrestrita de $L_c(\cdot, \lambda^{k+1})$.
- A vantagem do uso combinado destas idéias é poder empregar, em virtude de uma melhor aproximação de $\lambda^*$, uma taxa menor de crescimento da penalidade $c^k$. Assim sendo, há uma tendência do problema ser menos malcondicionado.
- Se $c^k$ aumenta muito rapidamente, a sequência $\{x^k\}$ tende a convergir mais rapidamente, mas o problema de malcondicionamento deve ser mais evidente.

Consiste em resolver uma sequência de Problemas:

$$\min \quad L_{c^k}(x, \lambda^k) = f_0(x) + \sum_{i=1}^{q} \lambda_i^k h_i(x) + \frac{c^k}{2} \left( \sum_{i=1}^{q} h_i(x)^2 \right)$$
$$x \in X$$

onde $\{\lambda^k\}$ é uma sequência em $\mathbb{R}^q$ e $\{c^k\}$ é uma sequência de penalidades positivas.

- Na versão original do método (década de 1960 - Método de Penalidades Puros que vimos), os multiplicadores não eram aproximados; os multiplicadores eram fixados em zero. A idéia de aproximá-los veio mais tarde.
- Para sua validade, o método depende de incrementar $c^k$ para $\infty$.

## Proposição

Assuma que $f_0, h$ são funções contínuas, que $X$ é um conjunto fechado e que o conjunto $\{x \in X : h(x) = 0\}$ é não vazio. Para $k = 0, 1, \ldots,$, seja $x^k$ um mínimo global do problema

$$\begin{array}{ll} \min & L_{c^k}(x, \lambda^k) \\ & x \in X, \end{array}$$

onde $\{\lambda^k\}$ é limitada, $0 < c^k < c^{k+1}, \ \forall k$ e $c^k \to \infty$.

Então todo ponto limite da sequência $\{x^k\}$ é um mínimo global do problema original.

- O resultado assume que o mínimo global irrestrito é obtido de forma exata. Entretanto, os métodos de otimização irrestrita são terminados quando $\|\nabla_x L_c(x^k, \lambda^k)\| \leq \epsilon^k$, onde $\epsilon^k \to 0$.

## Proposição

Assuma que $f_0, h$ sejam funções diferenciáveis e que $X = \mathbb{R}^n$. Para $k = 0, 1, \ldots,$ assuma que $x^k$ satisfaça $\|\nabla_x L_c(x^k, \lambda^k)\| \leq \epsilon^k$, onde $\{\lambda^k\}$ é limitada e $\{\epsilon^k\}, \{c^k\}$ satisfazem:

- $0 < c^k < c^{k+1}, \ \ \forall k, \ \ c^k \to \infty$
- $0 \leq \epsilon^k, \ \ \forall k, \ \ \epsilon^k \to 0.$

Assuma que a subsequência $\{x^k\}_K$ convirja para o vetor $x^*$ tal que $\nabla h(x^*)$ possua posto $q$. Então: $\{\lambda^k + c^k h(x^k)\}_K \to \lambda^*$, onde $\lambda^*$ em conjunto com $x^*$ satisfazem as condições necessárias de primeira ordem:

$$\nabla f_0(x^*) + \nabla h(x^*)^T \lambda^* = 0, \ \ h(x^*) = 0.$$

# Duas visões sobre o Método do Lagrangeano Aumentado

1. Como um método que usa uma função de penalidade *exata*
2. Como um método dual

- Para um valor fixo de $\lambda \in \mathbb{R}^q$, a função Lagrangeana Aumentada $L_c(x, \lambda)$ é simplesmente a função de Penalidade Quadrática para:

$$(Paux) \ \min f_0(x) + \lambda^T h(x)$$
$$h(x) = 0$$

  Este programa é equivalente ao original, uma vez que combinações das restrições de igualdade anexadas à função objetivo não alteram a solução ótima.

- O que aconteceria com a solução irrestrita de

$$\min f_0(x) + \lambda^T h(x) + \frac{c}{2}\|h(x)\|^2$$

  caso, ao invés de um valor $\lambda$ qualquer, o vetor de multiplicador $\lambda^*$ associado a uma solução ótima local regular do programa original fosse utilizado na Lagrangeana ?

- O gradiente da Lagrangeana aumentada seria nulo em $x^*, \lambda^*$:

$$\nabla L_c(x^*, \lambda^*) = \nabla f_0(x^*) + \nabla h(x^*)\lambda^* + c\nabla h(x^*)h(x^*) = 0$$

pois

$$\nabla f_0(x^*) + \nabla h(x^*)\lambda^* = 0, \text{ com } h(x^*) = 0$$

para $x^*, \lambda^*$ que satisfazem as CNPO para o problema original. Ou seja, a Função Lagrangeana Aumentada pode ser vista como uma função penalidade exata para *Paux*.

**Definição**

Uma função de penalidade é dita exata se a solução do problema penalizado fornece a solução do problema original para um valor finito da penalidade.

## Esquema do algoritmo

1. Alguma regra para incrementar (ou manter) $c$ ao longo das interações do método.
2. Estimativa inicial $\lambda^k$ para o multiplicador ótimo $\lambda^*$ no ponto regular $x^*$.
3. Resolva
$$\min f_0(x) + (\lambda^k)^T h(x) + \frac{c}{2}\|h(x)\|^2$$
obtendo $x^k$ como solução.
4. Obtenha uma nova aproximação do multiplicador $\lambda^{k+1}$.

- O multiplicador ótimo $\lambda^*$ em conjunto com a solução regular $x^*$ do problema original satisfazem $\nabla f_0(x^*) + \nabla h(x^*)\lambda^* = 0$. Então, para $\lambda^k$ fixo, o multiplicador de Lagrange ótimo de

$$\min f_0(x) + (\lambda^k)^T h(x)$$
$$h(x) = 0$$

  é $\alpha$ tal que $\nabla f_0(x^*) + \nabla h(x)\lambda^k + \nabla h(x^*)\alpha = 0$.
- Então: $\alpha + \lambda^k = \lambda^*$.
- Como $f_0(x) + (\lambda^k)^T h(x) + \frac{c}{2}\|h(x)\|^2$ é uma função de penalidade para o problema acima, pelos resultados da seção de métodos de penalidades, uma aproximação para o multiplicador de Lagrange ótimo $\alpha$ é $\alpha \approx c\nabla\gamma(g^+(x^k))$ que no caso de restrições de igualdade é $\alpha \approx ch(x^k)$. Logo, o método opera segundo:

$$\lambda^{k+1} = \lambda^k + ch(x^k)$$

## A grande vantagem do método

- Uma vez que se $\lambda^*$ for empregado, a função Lagrangeana aumentada funciona como uma função de penalidade exata, temos expectativa de que com o uso de uma boa aproximação $\lambda^k$ para $\lambda^*$, construída como descrito anteriormente, a penalidade $c$ possa permanecer finita.
- A penalidade $c$ é ajustada ao longo do método, mas não há a necessidade de que cresça tanto quanto no método de penalidade pura.

O problema do malcondiconamento numérico é bastante contornado com o Método do Lagrangeano Aumentado.

Analogamente, $L_c(x, \lambda)$ é a função Lagrangeana para

$$\min f_0(x) + \frac{c}{2}\|h(x)\|^2$$
$$h(x) = 0$$

que é equivalente ao problema original, uma vez que a introdução da penalidade quadrática à função objetivo não altera a função objetivo ótima, a solução ótima ou os multiplicadores de Lagrange ótimos.

Entretanto, enquanto a função Lagrangeana $L(x, \lambda) = f_0(x) + \lambda^T h(x)$ pode não ser convexa, um valor suficientemente grande de $c$ torna a Lagrangeana aumentada localmente convexa, caso as condições suficientes de otimalidade sejam satisfeitas por $\lambda^*, x^*$.
Recordando, o Lema auxiliar aplicado à:

$$\nabla^2_{xx} L_c(x^*, \lambda^*) = \nabla^2_{xx} L(x^*, \lambda^*) + c\nabla h(x^*)\nabla h(x^*)^T$$

$$(y^T \nabla^2_{xx} L(x^*, \lambda^*) y > 0 \text{ apenas para } y \in V(x^*))$$

- Por continuidade, o argumento acima pode ser estendido para uma vizinhança de $x^*, \lambda^*$:

  Para um $\lambda$ próximo a $\lambda^*$, $L(x, \lambda)$ deve ter um único minimizador, $x(\lambda)$, próximo a $x^*$.

- Se um valor de $\lambda$ for tal que $h(x(\lambda)) = 0$, este valor de $\lambda$ deve ser necessariamente $\lambda^*$ e $x(\lambda)$ deve ser $x^*$, uma vez que $\lambda, x(\lambda)$ satisfazem as condições necessárias para o problema original:

$$\nabla f_0(x(\lambda)) + \nabla h(x(\lambda))\lambda + c\nabla h(x(\lambda))h(x(\lambda)) = 0$$

- Então, o problema de encontrar o valor correto para $\lambda$ consiste em resolver:

$$h(x(\lambda)) = 0$$

para o qual a iteração típica

$$\lambda^{k+1} = \lambda^{k} + ch(x(\lambda^{k}))$$

formaliza um processo de aproximação sucessiva.

- No método do Lagrangeano Aumentado, a iteração principal é em relação ao multiplicador de Lagrange $\lambda$, e esta visão introduz grandes melhoramentos no método.

- Como já discutimos a Lagrangeana do problema

$$\min f_0(x) + \frac{c}{2}\|h(x)\|^2$$
$$h(x) = 0$$

ou, equivalentemente, a Lagrangeana aumentada do problema original, é localmente convexa para $c$ suficientemente grande.

- Então a Teoria de Dualidade Local é aplicável neste caso.

Aplicação da Teoria de Dualidade Local: Assumindo que $L_c(x^*, \lambda^*) \succ 0$ para o valor $c$ escolhido.

Definindo

$$\theta(\lambda) = \min_x L_c(x, \lambda) = \min_x \{ f_0(x) + \lambda^T h(x) + \frac{c}{2} \|h(x)\|^2 \}$$

em uma vizinhança de $x^*, \lambda^*$.

- Se $x(\lambda)$ é a solução irrestrita da minimização de $L_c(x, \lambda)$, sabemos que $\nabla_\lambda \theta(\lambda) = h(x(\lambda))$.
- Então a iteração principal do Método do Lagrangeano Aumentado

$$\lambda^{k+1} = \lambda^k + ch(x(\lambda)) = \lambda^k + c\nabla_\lambda \theta(\lambda^k)$$

pode ser entendida como um passo do método do Gradiente puro, com passo $c$, visando maximizar $\theta(\lambda)$ nas vizinhanças de $\lambda^k$.

- O passo $c$ pode ser uma boa escolha, mas em alguns casos vale a pena introduzir uma busca unidirecional. A taxa de convergência do Método do Gradiente para a maximização de $\theta(\lambda)$ depende da raão dos autovalores de $\nabla^2_{\lambda\lambda}\theta(\lambda^*) \prec 0$.
- É possível mostrar que quando $c \to \infty$, a razão entre os autovalores de $\nabla^2_{\lambda\lambda}\theta(\lambda^*)$ tende para a unidade e o problema dual fica melhor condicionado.

**Sumário: Efeito da introdução de um termo de penalidade**

Ao introduzirmos um termo de penalidade e fazendo $c \to \infty$, o número de condição do problema primal (Hessiana da Lagrangeana Aumentada) torna-se progressivamente ruim. Enquanto isto, o número de condição do problema dual (da Hessiana de $\theta(\lambda)$) torna-se progressivamente bom, próximo à unidade. Entretanto, para a aplicação do Método Dual, é necessário resolver um problema penalizado malcondicionado ($\min_x L(x, \lambda)$) a cada passo.