Métodos Iterativos para Otimização Sem Restrições

Fernanda Costa

Departamento de Matemática
Universidade do Minho

Outline

Métodos iterativos para otimização sem restrições

Em geral, não se consegue resolver

$$\underset{w \in \mathbb{R}^d}{\text{minimizar}} \; F(w)$$

analiticamente, e recorre-se a métodos iterativos.

Métodos iterativos para Otimização

- Começam a partir de um dado ponto inicial $w^{(0)} \in \mathbb{R}^d$ (aproximação inicial à solução);
- Dado $w^{(k)}$, calculam um novo ponto (melhor) $w^{(k+1)}$, e este processo repete-se ($k = 0, 1, 2, \ldots$).
- Geram uma sucessão de pontos $w^{(1)}, w^{(2)}, \ldots, w^{(k)}, \ldots$ que converge para a solução ótima w^*.
- O método *termina* quando deixa de ser possível obter progresso significativo ou quando a solução ótima w^* é aproximada com precisão previamente fixada.

Critério de paragem

- Ao utilizar um algoritmo de otimização, procura-se um ponto $w^{(k)}$ onde

$$\nabla F(w^{(k)}) = \mathbf{0}.$$

Na prática, devido à aritmética de precisão finita, adota-se como critério de paragem

$$\|\nabla F(w^{(k)})\|_\infty < \varepsilon, \tag{1}$$

com $\varepsilon > 0$ suficientemente pequeno.

- A segunda condição de otimalidade (Hessiana semidefinida positiva) não é, em geral, utilizada.

Assim, o algoritmo apenas garante um ponto estacionário (ponto máximo, mínimo ou sela).

- Contudo, como os algoritmos que vamos estudar geram direções de descida, espera-se, na prática, convergência para um mínimo local.

- Num algoritmo prático, utilizam-se outras condições para além da condição (1). Por vezes, a função pode apresentar problemas numéricos ou ser ilimitada inferiormente, que impedem a verificação da condição (1).

- Para evitar que o algoritmo execute indefinidamente, é habitual impor um *número máximo* ao *número de iterações*, *Kmax*, ou ao *número de avaliações da função*, *NFmax*, ou ao *tempo de execução*, *Tmax*.

- Além disso, para detetar situações em que o algoritmo não está a fazer progressos significativos, podem definir-se condições adicionais baseadas na proximidade entre dois pontos consecutivos e na variação do valor da função objetivo, em termos absoluto ou relativo:

$$\left\| w^{(k)} - w^{(k-1)} \right\| \leq \varepsilon_2 \quad \text{ou} \quad \frac{\left\| w^{(k)} - w^{(k-1)} \right\|_\infty}{\left\| w^{(k)} \right\|_\infty} \leq \varepsilon_2$$

$$\left| F(w^{(k)}) - F(w^{(k-1)}) \right| \quad \text{ou} \quad \frac{\left| F(w^{(k)}) - F(w^{(k-1)}) \right|}{\left| F(w^{(k)}) \right|} \leq \varepsilon_3$$

.

Critérios de paragem em algoritmos para otimização sem restrições.

Critério (proposto por Wolfe). Parar o algoritmo se:

1. $\left\|\nabla F(w^{(k)})\right\|_\infty \leq \varepsilon_1$ e **(medida de estacionaridade)**

2. $\dfrac{\left\|w^{(k)} - w^{(k-1)}\right\|_\infty}{\|w^{(k)}\|_\infty} \leq \varepsilon_2$ e

3. $\dfrac{\left|F(w^{(k)}) - F(w^{(k-1)})\right|}{|F(w^{(k)})|} \leq \varepsilon_3$

onde $\varepsilon_1 > 0, \varepsilon_2 > 0, \varepsilon_3 > 0$ são pequenas tolerâncias próximas de zero. Na prática, $\varepsilon_1 = 10^{-6}$ e $\varepsilon_2 = \varepsilon_3 = 10^{-6}$.

Critério (proposto por Gill e Murray): Terminar o algoritmo se:

1. $\|\nabla F(w^{(k)})\|_\infty \leq \varepsilon^{\frac{1}{3}} \left(1 + |F(w^{(k)})|\right)$ e
2. $\|w^{(k)} - w^{(k-1)}\|_\infty \leq \varepsilon \left(1 + \|w^{(k)}\|_\infty\right)$ e
3. $|F(w^{(k)}) - F(w^{(k-1)})| \leq \varepsilon^2 \left(1 + |F(w^{(k)})|\right)$

onde $\varepsilon > 0$ é uma pequena tolerância próxima de zero.

Observações:

- Este critério é de aplicação mais geral. Pode ser aplicado a problemas em que a solução ótima é o vetor nulo ou o valor ótimo da função objetivo é zero.

Métodos de procura unidirecional (geral)

Os métodos de procura unidirecional tem, em cada iteração, três passos principais:

Estrutura geral

1. Escolher uma direção de procura $s^{(k)}$ a partir do ponto atual $w^{(k)}$;
2. Calcular a distância a percorrer nessa direção, η_k, realizando uma procura unidirecional;
3. Mover para o novo ponto: $w^{(k+1)} = w^{(k)} + \eta_k \, s^{(k)}$.

O bom desempenho de um método de procura unidirecional depende das escolhas da *direção de procura* $s^{(k)}$ e do *comprimento do passo* η_k.

De preferência $s^{(k)}$ é uma *direção de descida*, isto é, $s^{(k)T} \nabla F(w^{(k)}) < 0$, o que garante que a função F decresce ao longo dessa direção.

O *método de procura undirecional (geral)* é definido pelo seguinte algoritmo iterativo.

Algoritmo 1: Método de procura unidirecional (geral)

1. **Entrada**: F, $w^{(0)}$, tolerância $\varepsilon > 0$
2. Fazer $k = 0$
3. **Enquanto** $(\|\nabla F(w^{(k)})\|_\infty > \varepsilon)$ **fazer**
4. Calcular uma direção de procura $s^{(k)}$ tal que $s^{(k)\,T}\nabla F(w^{(k)}) < 0$
5. Calcular o comprimento do passo η_k tal que
$$F(w^{(k)} + \eta_k s^{(k)}) < F(w^{(k)})$$
6. Fazer $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)}$
7. Fazer $k = k + 1$
8. **Fim enquanto**
9. **Saída:** $w^* \leftarrow w^{(k)}$, $F^* \leftarrow F(w^{(k)})$ (solução ótima e valor ótimo)

Existem vários métodos para escolher a *direção de procura* $s^{(k)}$. Na aula seguinte, estudaremos os seguintes:

- Método de descida máxima (ou Método do gradiente).
- Método de Newton.
- Métodos quasi-Newton.

Como escolher o *comprimento do passo* η_k?

- A escolha de η_k deverá conduzir a um decréscimo substancial de F e ser pouco dispendiosa.

Comprimento do passo

A procura exata do η_k:

$$\underset{\eta \in \mathbb{R}_0^+}{\text{minimizar}} \; \phi(\eta) = F(w^{(k)} + \eta s^{(k)})$$

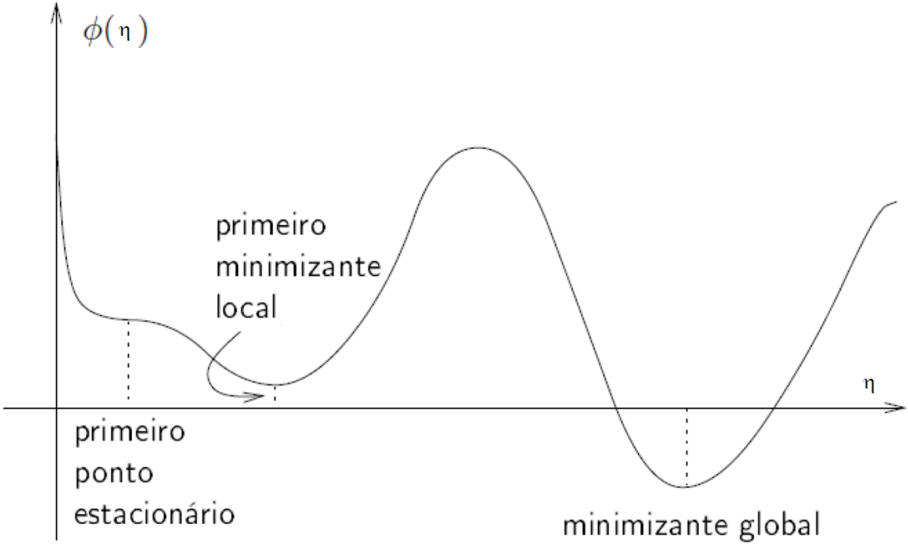
é impraticável, devido ao elevado custo computacional e, em geral, não é necessária. Na prática, recorre-se a técnicas de procura não exata.

Vamos tentar decrescer

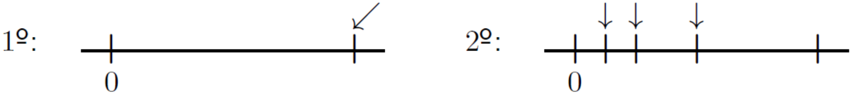
$$\phi(\eta) = F(w^{(k)} + \eta s^{(k)}), \quad \eta \geq 0$$

a partir de $\eta = 0$.

Contudo, minimizar $\phi(\eta)$ pode ser difícil.


$\phi(\eta)$
primeiro
minimizante
local
η
primeiro
ponto
estacionário
minimizante global

Pretende-se testar um número finito (pequeno) de valores para η, por ex.,



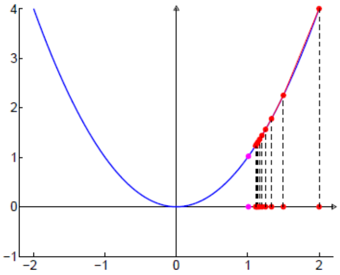
Uma condição natural a impor ao parâmetro η_k é a **condição de decréscimo simples**:

$$F\left(w^{(k)} + \eta_k s^{(k)}\right) < F\left(w^{(k)}\right).$$

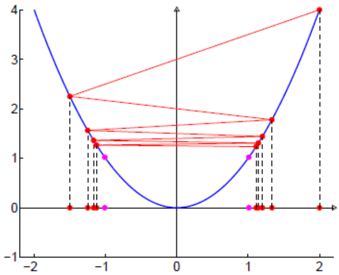
Mas, esta condição, por si só, não garante a convergência para um ponto estacionário, podendo o método convergir antes de o atingir!

Exemplo: Considere $F : \mathbb{R} \to \mathbb{R}$ dada por $F(w) = w^2$, que tem um minimizante global em $w^* = 0$. (a) Utilize o algoritmo de descida geral com $s^{(k)} = -1$, $\eta_k = \frac{1}{2^{k+1}}$ e ponto inicial dado por $w^{(0)} = 2$. Desenhe os iterados gerados pelo algoritmo, definidos por $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)}$, usando o MatLab. O que pode concluir? (b) Resolva a alínea anterior com com $s^{(k)} = (-1)^{k+1}$ e $\eta_k = 2 + \frac{3}{2^{k+1}}$.

Solução:



(a) passos muito curtos
$w^{(k)} = 1 + \frac{1}{2^k}$
$w^{(k)} \to 1$
1 não é minimizante de F



(b) passos muito longos
$w^{(k)} = 1 + \frac{1}{2^k}$ (k ímpar); $w^{(k)} = -1 - \frac{1}{2^k}$ (k par)
$w^{(k)} \to -1$ (k par)
$w^{(k)} \to 1$ (k ímpar)
-1 e 1 não são minimizantes de F

Em ambos os casos, o algoritmo satisfaz a condição de decréscimo simples, ($F(w^{(k+1)}) < F(w^{(k)})$ para $k = 0, 1, \ldots$), mas não converge para o minimizante!

Para simplificar a exposição da procura undirecional, vamos considerar a função de 1 variável

$$\phi(\eta) = F(w^{(k)} + \eta s^{(k)}) \tag{2}$$

onde $\eta = 0$ corresponde ao início da procura undirecional ($w^{(k)}$), e portanto $\phi(0) = F(w^{(k)})$.

Pela regra da cadeia, usando $w = w^{(k)} + \eta s^{(k)}$, a derivada de ϕ é

$$\phi'(\eta) = \frac{\partial(F(w))}{\partial \eta} = \sum_{i=1}^{d} \frac{\partial F}{\partial w_i} \frac{dw_i}{d\eta} \Longleftrightarrow$$

$$\phi'(\eta) = \nabla F(w^{(k)} + \eta s^{(k)})^T s^{(k)}$$

que é a derivada direcional de F ao longo da direção de procura $s^{(k)}$. O declive no início de cada procura unidirecional é

$$\phi'(0) = \nabla F(w^{(k)})^T s^{(k)}.$$

Como $s^{(k)}$ tem de ser uma direção de descida, $\phi'(0) < 0$ é sempre negativo!

Condições de Wolfe

As condições de Wolfe englobam a **condição de decréscimo suficiente** (CDS) e a **condição de curvatura** (CC).

A CDS, também designada por **condição de Armijo**, é dada pela desigualdade:

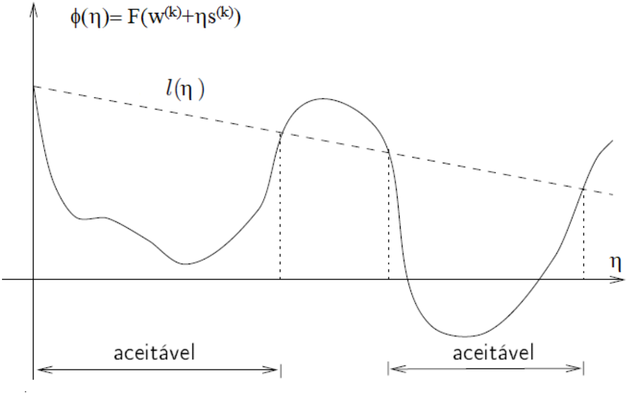
$$F(w^{(k)} + \eta s^{(k)}) - F(w^{(k)}) \leq c_1 \eta \underbrace{\nabla F(w^{(k)})^T s^{(k)}}_{\phi'(0)} \Longleftrightarrow$$

$$\underbrace{F(w^{(k)} + \eta s^{(k)})}_{\phi(\eta)} \leq \underbrace{F(w^{(k)}) + c_1 \eta \nabla F(w^{(k)})^T s^{(k)}}_{l(\eta)} \tag{3}$$

onde $0 < c_1 < 1$ é uma constante. A quantidade $\eta\phi'(0)$ representa o decréscimo esperado da função F. O multiplicador c_1 diz que (3) será satisfeita desde que se obtenha pelo menos uma pequena fração do decréscimo esperado.

Em (3), $l(\eta)$ é uma função linear de declive negativo $c_1\eta\phi'(0)$. Como $c_1 \in (0, 1)$, a reta $l(\eta)$ situa-se acima do gráfico de $\phi(\eta)$ para valores de η pequenos.

A CDS diz que η é aceitável se $\phi(\eta) \leq l(\eta)$. Os intervalos onde CDS é satisfeita são apresentados na Figura. Na prática, c_1 é escolhido bastante pequeno, $c_1 = 10^{-4}$.



Condição de decréscimo suficiente

A CDS, por si só, não é suficiente para garantir que o algoritmo faça um progresso razoável porque, como podemos ver da Figura, pode ser satisfeita por comprimentos de passo muito pequenos!

Para excluir passos demasiado curtos, introduz-se uma segunda condição, a CC, que exige que η_k satisfaça:

$$\underbrace{\nabla F(w^{(k)} + \eta_k s^{(k)})^T s^{(k)}}_{\phi'(\eta_k)} \geq c_2 \underbrace{\nabla F^T(w^{(k)}) s^{(k)}}_{\phi'(0)} \tag{4}$$

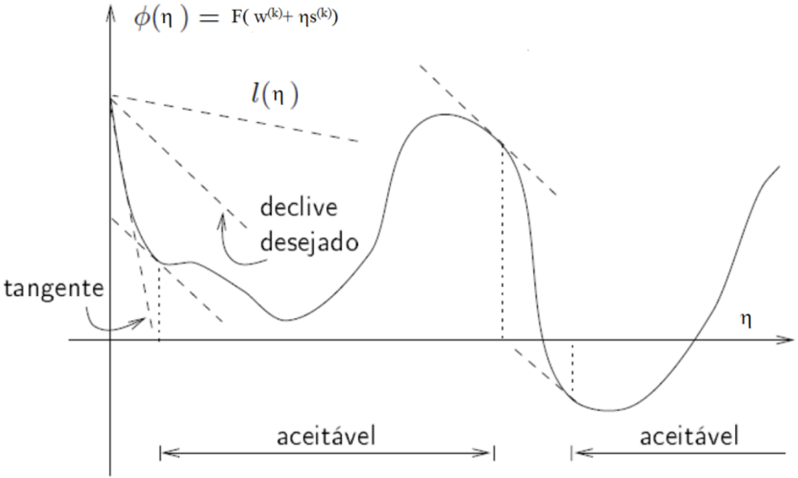
com $c_2 \in (c_1, 1)$, onde c_1 é a constante de (3).

A CC rejecta declives demasiados negativos (o declive em η_k, $\phi'(\eta_k)$, tem de ser maior c_2 vezes o declive inicial $\phi'(0)$). Isto faz sentido porque, se o declive $\phi'(\eta)$ é fortemente negativo, isso indica que ainda é possível reduzir significativamente o valor de F avançando mais na direção $s^{(k)}$.

Por outro lado, se $\phi'(\eta)$ for apenas ligeiramente negativo, ou mesmo positivo, isso indica que não é razoável esperar uma redução muito maior nessa direção. Portanto, faz sentido terminar a procura.

Na prática, $c_2 = 0.9$ quando a direção de procura $s^{(k)}$ é dada pelo método de Newton ou pelo método quasi-Newton.

Os intervalos onde CC é satisfeita são apresentados na Figura.



Condição de curvatura

Lema

Seja $F : \mathbb{R}^d \rightarrow \mathbb{R}$ continuamente diferenciável. Seja $s^{(k)}$ uma direcção de descida em $w^{(k)}$ e F limitada inferiormente ao longo da semi-recta $\{w^{(k)} + \eta s^{(k)} : \eta > 0\}$. Então, se $0 < c_1 < c_2 < 1$, existem intervalos de comprimentos de passo satisfazendo as condições de Wolfe, CDS (3) e CC (4).

Demonstração: $\phi(\eta) = F(w^{(k)} + \eta s^{(k)})$ é limitada inferiormente.
Uma vez que $0 < c_1 < 1$, $l(\eta) = F(w^{(k)}) + c_1 \eta \nabla F(w^{(k)})^T s^{(k)}$ tende para $-\infty$ quando $\eta \rightarrow +\infty$.
Por isso intersecta ϕ pelo menos uma vez.
Seja η' o menor valor de η dessa intersecção, assim

$$F(w^{(k)} + \eta' s^{(k)}) = F(w^{(k)}) + c_1 \eta' \nabla F(w^{(k)})^T s^{(k)}$$

Logo a CDS é válida $\forall \eta < \eta'$.
Pelo teorema do valor médio

$$\exists \eta'' \in (0, \eta') : F(w^{(k)} + \eta' s^{(k)}) - F(w^{(k)}) = \eta' \nabla F(w^{(k)} + \eta'' s^{(k)})^T s^{(k)}$$

Tendo em conta as duas últimas equações, vem

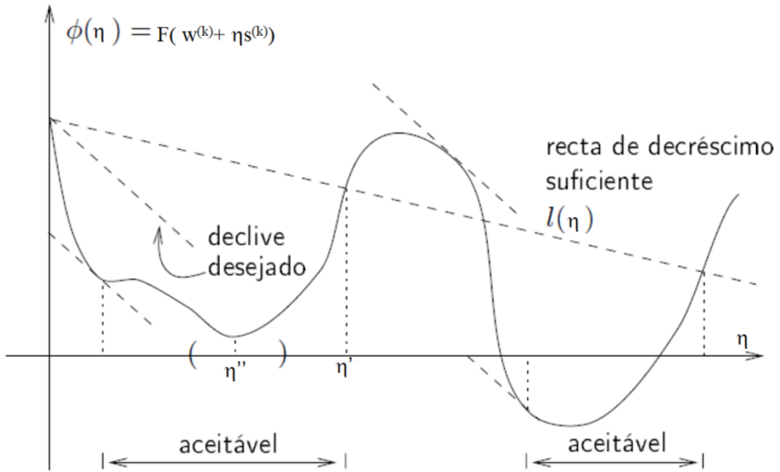
$$\nabla F(w^{(k)} + \eta'' s^{(k)})^T s^{(k)} = c_1 \nabla F(w^{(k)})^T s^{(k)}$$

Uma vez que $c_1 < c_2$ e $\nabla F(w^{(k)})^T s^{(k)} < 0$, vem

$$\nabla F(w^{(k)} + \eta'' s^{(k)})^T s^{(k)} = c_1 \nabla F(w^{(k)})^T s^{(k)} > c_2 \nabla F(w^{(k)})^T s^{(k)}$$

Assim, η'' satisfaz as condições de Wolfe e pelo facto de ∇F ser contínuo, existe um intervalo em torno de η'' onde as condições de Wolfe são satisfeitas. $\square$

Os intervalos onde as condições de Wolfe, CDS e CC, são satisfeitas são apresentados na Figura.



Intervalos de comprimentos de passo satisfazendo as condições de Wolfe

Decréscimo Suficiente e Backtracking

Como referido a CDS (3), por si só, não garante progresso adequado ao longo da direção de procura.

Contudo, se o algoritmo de procura unidirecional selecionar os comprimentos de passo candidatos de forma apropriada, recorrendo à estratégia de **backtracking**, podemos dispensar a CC (4) e usar apenas a CDS (3).

Algoritmo: Condição de Armijo com backtracking

1. **Entrada:** $\eta_0 > 0$, $\rho \in (0,1)$, $c \in (0,1)$, F, $\nabla F(w^{(k)})$, $w^{(k)}$, $s^{(k)}$
2. Fazer $\eta = \eta_0$
3. **Enquanto** $F(w^{(k)} + \eta s^{(k)}) > F(w^{(k)}) + c\eta \nabla F(w^{(k)})^T s^{(k)}$ **fazer**
4. Fazer $\eta = \rho\, \eta$
5. **Fim enquanto**
6. **Saída:** $\eta_k \leftarrow \eta$ (valor para η_k)

Notas (estratégia de *backtracking*):

- Inicia-se com um comprimento de passo relativamente grande e este é sucessivamente reduzido, multiplicando por um fator $\rho < 1$, até que a CDS seja satisfeita.
- Trata-se de uma heurística simples para garantir a verificação da CDS.
- Na prática, o comprimento de passo inicial é frequentemente escolhido como $\eta_0 = 1$ quando a direção de procura $s^{(k)}$ é obtida pelo método de Newton ou por métodos quasi-Newton. Noutros métodos, como o método do gradiente, podem ser usados outros valores iniciais.
- O procedimento termina sempre, pois a CDS é satisfeita para η_k suficientemente pequeno.
- Na prática, o fator de contração é geralmente fixo, sendo habitual a escolha $\rho = \frac{1}{2}$.

- Diz-se que um método tem *convergência de primeira ordem* se a sucessão de iterados $\{w^{(k)}\}$ gerada pelo método converge para um *ponto estacionário*.
  - Usa-se o termo *convergência global* se o método converge para um ponto estacionário qualquer que seja o ponto inicial $w^{(0)}$.
  - Usa-se o termo *convergência local* se o método converge para um ponto estacionário apenas se o ponto inicial $w^{(0)}$ está suficientemente próximo de uma solução w^*.

Convergência dos métodos de procura unidirecional

- Para garantir convergência global, é necessário escolher bem tanto os comprimentos de passo η_k quanto as direções de procura $s^{(k)}$.

- Apresentamos um teorema que estabelece condições sobre a direção de procura $s^{(k)}$ para garantir convergência global, baseado no ângulo θ_k entre $s^{(k)}$ e a direção de máxima descida $-\nabla F(w^{(k)})$:

$$\cos\theta_k = \frac{-\nabla F(w^{(k)})^\top s^{(k)}}{\|\nabla F(w^{(k)})\| \, \|s^{(k)}\|} \tag{5}$$

  e assumindo que η_k satisfaz as condições de Wolfe.

- Este Teorema é importante: por exemplo, mostra que o método da descida máxima (método do gradiente) é *globalmente convergente*; para outros métodos, descreve até que ponto $s^{(k)}$ pode desviar-se da direção de máxima $-\nabla f(w^{(k)})$ e ainda assim gerar um algoritmo *globalmente convergente*.

Teorema (Zoutendijk)

Considere-se um método iterativo $w^{(k+1)} = w^{(k)} + \eta s^{(k)}$, *em que* $s^{(k)}$ *é uma direção de descida e* η_k *satisfaz as condições de Wolfe.*

Suponhamos que F *é limitada inferiormente em* $\mathbb{R}^d$ *e que é continuamente diferenciável num conjunto aberto* $\mathcal{N}$ *contendo o conjunto de nível*

$$\mathcal{L}(w^{(0)}) = \{\, w : F(w) \leq F(w^{(0)}) \,\},$$

onde $w^{(0)}$ *é o ponto inicial da iteração.*

Suponhamos, ainda, que ∇F *é contínua à Lipschitz (com constante* $L > 0$*) em* $\mathcal{N}$*:*

$$\|\nabla F(w) - \nabla F(\tilde{w})\| \leq L\|w - \tilde{w}\|.$$

Então, tem-se

$$\sum_{k=0}^{+\infty} \cos^2\theta_k \, \|\nabla F(w^{(k)})\|^2 < +\infty \qquad \textit{(condição de Zoutendijk).}$$

Do teorema de Zoutendijk, resulta que

$$\cos^2\theta_k \,\|\nabla F(w^{(k)})\|^2 \longrightarrow 0.$$

Se existir uma constante $\varepsilon > 0$ tal que

$$\cos\theta_k \geq \varepsilon > 0 \quad \text{para todo } k,$$

então

$$\cos^2\theta_k \,\|\nabla F(w^{(k)})\|^2 \longrightarrow 0 \quad \Longleftrightarrow \quad \|\nabla F(w^{(k)})\| \longrightarrow 0.$$

Por outras palavras, podemos garantir que as normas do gradiente $\|\nabla F(w^{(k)})\|$ convergem para zero, desde que as direções de procura $s^{(k)}$ nunca se tornem quase ortogonais ao vetor gradiente $\nabla F(w^{(k)})$.

Exemplos:

- **Método de descida máxima:**

$$s^{(k)} = -\nabla f(w^{(k)}) \;\Rightarrow\; \cos\theta_k = 1 > 0 \quad \forall k \;\Rightarrow\; \text{convergência global.}$$

  Método simples e intuitivo, mas pode ser lento.

- **Métodos do tipo Newton:**

$$s^{(k)} = -B_k^{-1}\nabla f(w^{(k)}),$$

  com B_k uma matriz simétrica e definida positiva (SPD) e com número de condição limitado, $cond(B_k) = \|B_k\|\,\|B_k^{-1}\| \leq M$, então:

$$\cos\theta_k \geq \frac{1}{M} > 0 \quad \forall k \;\Rightarrow\; \text{convergência global.}$$

Se B_k são definidas positivas, com número de condição limitado, e η_k satisfaz as condições de Wolfe, então o método é *globalmente convergente*.

Isto inclui, o método de descida máxima, alguns métodos de Newton e quasi-Newton.

Taxas de convergência

FALTA FAZER!

Resolução matemática do exemplo da página 14:

(a): Matematicamente temos $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)} = w^{(k)} - \eta_k$, para $k = 0, 1, 2, \ldots$:

$$w^{(1)} = w^{(0)} - \eta_0$$

$$w^{(2)} = w^{(1)} - \eta_1 = w^{(0)} - \eta_0 - \eta_1$$

$$\ldots = \ldots$$

$$w^{(k+1)} = w^{(k)} - \eta_k = w^{(0)} - \sum_{i=0}^{k} \eta_i.$$

Portanto,

$$w^{(k+1)} = 2 - \sum_{i=0}^{k} \frac{1}{2^{i+1}} = 2 - \frac{1}{2}\sum_{i=0}^{k} \frac{1}{2^i} = 2 - \frac{1}{2}\left(\frac{1-\left(\frac{1}{2}\right)^{k+1}}{1-\frac{1}{2}}\right) = 2 - \left(1 - \left(\frac{1}{2}\right)^{k+1}\right)$$

$$= 1 + \left(\frac{1}{2}\right)^{k+1}$$

Observar que $0 < w^{(k+1)} < w^{(k)}$ e F é crescente para $w > 0$. Portanto, é verificada a condição de decréscimo simples:

$$F\left(w^{(k+1)}\right) < F\left(w^{(k)}\right), \quad \text{para } k = 0, 1, \ldots.$$

Temos, para $k \to \infty$, $w^{(k)} = 1 + \left(\frac{1}{2}\right)^k \to 1$ e $F(w^{(k)}) \to 1$. Mas o minimizante é 0, portanto o 1 não é ponto estacionário.

(b): Matematicamente temos $w^{(k+1)} = w^{(k)} + \eta_k s^{(k)} = w^{(k)} + (-1)^{k+1}\eta_k$, para $k = 0, 1, 2, \ldots$:

$$
\begin{aligned}
w^{(1)} &= w^{(0)} - \eta_0 \\
w^{(2)} &= w^{(1)} + \eta_1 = w^{(0)} - \eta_0 + \eta_1 \\
\ldots &= \ldots
\end{aligned}
$$

$$w^{(k+1)} = w^{(k)} + (-1)^{k+1}\eta_k = w^{(0)} + \sum_{i=0}^{k}(-1)^{i+1}\eta_i.$$

Portanto,

$$
\begin{aligned}
w^{(k+1)} &= 2 + \sum_{i=0}^{k}(-1)^{i+1}(2 + \frac{3}{2^{i+1}}) = 2 + 2\sum_{i=0}^{k}(-1)^{i+1} + 3\sum_{i=0}^{k}\left(\frac{-1}{2}\right)^{i+1} \\
&= 2 + 2\begin{cases} -1, & k \text{ par} \\ 0, & k \text{ ímpar} \end{cases} + 3\left(-\frac{1}{2}\,\frac{1-\left(-\frac{1}{2}\right)^{k+1}}{1+\frac{1}{2}}\right) \\
&= 2 + \begin{cases} -2, & k \text{ par} \\ 0, & k \text{ ímpar} \end{cases} + 3\left(-\frac{1}{3}\left(1-\left(-\frac{1}{2}\right)^{k+1}\right)\right) \\
&= 2 + \begin{cases} -2, & k \text{ par} \\ 0, & k \text{ ímpar} \end{cases} + \left(-1+\left(-\frac{1}{2}\right)^{k+1}\right) \\
&= 2 + \begin{cases} -3-\left(\frac{1}{2}\right)^{k+1}, & k \text{ par}, \\ -1+\left(\frac{1}{2}\right)^{k+1}, & k \text{ ímpar}. \end{cases} = \begin{cases} -1-\left(\frac{1}{2}\right)^{k+1}, & k \text{ par}, \\ 1+\left(\frac{1}{2}\right)^{k+1}, & k \text{ ímpar}. \end{cases}
\end{aligned}
$$

Portanto temos, para $k \to \infty$, $w^{(k)} \to -1$ se k par; $w^{(k)} \to 1$ se k ímpar. E $F(w^{(k)}) \to 1$, mas o minimizante é 0, portanto o 1 e -1 não são pontos estacionários.

Demonstração do Teorema (Zoutendijk):

Pela condição de decréscimo suficiente (CDS), tem-se

$$F\left(w^{(k+1)}\right) \leq F\left(w^{(k)}\right) + c_1 \eta_k \nabla F\left(w^{(k)}\right)^\top s^{(k)}.$$

Necessitamos de um limite inferior para η_k. Pela condição de curvatura (CC),

$$\nabla F\left(w^{(k+1)}\right)^\top s^{(k)} \geq c_2 \nabla F\left(w^{(k)}\right)^\top s^{(k)}.$$

Logo, subtraindo $\nabla F\left(w^{(k)}\right)^\top s^{(k)}$ a ambos os membros,

$$\left(\nabla F\left(w^{(k+1)}\right) - \nabla F\left(w^{(k)}\right)\right)^\top s^{(k)} \geq (c_2 - 1) \nabla F\left(w^{(k)}\right)^\top s^{(k)}.$$

Como ∇F é Lipschitz contínua com constante L, obtemos

$$\left\| \left(\nabla F\left(w^{(k+1)}\right) - \nabla F\left(w^{(k)}\right)\right)^\top s^{(k)} \right\| \leq L \eta_k \|s^{(k)}\|^2.$$

Assim,

$$(c_2 - 1) \nabla F\left(w^{(k)}\right)^\top s^{(k)} \leq L \eta_k \|s^{(k)}\|^2,$$

ou, equivalentemente,

$$\eta_k \geq \frac{c_2 - 1}{L} \frac{\nabla F\left(w^{(k)}\right)^\top s^{(k)}}{\|s^{(k)}\|^2}.$$

Temos, pela CDS,

$$F\left(w^{(k+1)}\right) \leq F\left(w^{(k)}\right) + c_1 \eta_k \nabla F\left(w^{(k)}\right)^\top s^{(k)},$$

e, substituindo este limite em CDS, obtemos

$$F\left(w^{(k+1)}\right) \leq F\left(w^{(k)}\right) - c_1 \frac{1-c_2}{L} \frac{\left(\nabla F\left(w^{(k)}\right)^\top s^{(k)}\right)^2}{\|s^{(k)}\|^2}.$$

Definindo

$$c = c_1 \frac{1-c_2}{L} > 0,$$

e usando a relação com o ângulo θ_k entre $s^{(k)}$ e $-\nabla F(w^{(k)})$:

$$\cos\theta_k = \frac{-\nabla F(w^{(k)})^\top s^{(k)}}{\|\nabla F(w^{(k)})\| \, \|s^{(k)}\|} \Longrightarrow -\nabla F(w^{(k)})^\top s^{(k)} = \cos\theta_k \, \|\nabla F(w^{(k)})\| \, \|s^{(k)}\|$$

obtemos

$$F\left(w^{(k+1)}\right) \leq F\left(w^{(k)}\right) - c \, \cos^2\theta_k \, \|\nabla F(w^{(k)})\|^2.$$

Esta desigualdade mostra que a função decresce pelo menos proporcional a $\cos^2\theta_k \, \|\nabla f(w^{(k)})\|^2$ em cada iteração, o que será fundamental para concluir a convergência da soma usada no teorema de Zoutendijk.

Somando a desigualdade de decrescimento obtida sobre as iterações, temos

$$F\left(w^{(k+1)}\right) \leq F(w^{(0)}) - c\sum_{j=0}^{k}\cos^2\theta_j\,\|\nabla F(w^{(j)})\|^2.$$

Como F é limitada inferiormente em $\mathcal{L}(w^{(0)})$, seja F^* um tal limite inferior. Então,

$$c\sum_{j=0}^{k}\cos^2\theta_j\,\|\nabla f(w^{(j)})\|^2 \leq F(w^{(0)}) - F\left(w^{(k+1)}\right) \leq F(w^{(0)}) - F^*.$$

Portanto, a sucessão

$$\left\{\sum_{j=0}^{k}\cos^2\theta_j\,\|\nabla F(w^{(j)})\|^2\right\}$$

é crescente e limitada superiormente, e, portanto é convergente.

Então, considerando $k \to \infty$, obtém-se o resultado do teorema de Zoutendijk:

$$\sum_{k=0}^{\infty}\cos^2\theta_k\,\|\nabla F(w^{(k)})\|^2 < +\infty \qquad \text{(série convergente).} \qquad \square$$