Aprendizagem Automática
Laboratório 2: **Otimização de Funções**

N.º:90007 Nome: Alice Rosa
N.º:90026 Nome: Aprígio Malveiro

Turno: 3ª feira - 14h00

2.1 Minimização das funções com uma variável

Questão 2.1.1

Tabela 1

η	$a = 0.5$	$a = 1$	$a = 2$	$a = 5$
.001	>1000	>1000	>1000	990
.01	760	414	223	97
.03	252	137	73	31
.1	75	40	21	8
.3	24	12	5	8
1	6	1	Div	Div
3	6	Div	Div	Div
Mais rápido	η=2	η=1	η=0.5	η=0.2
Limiar de divergência	η=4	η=2	η=1	η=0.4

Questão 2.1.2

A relação entre os valores de a e η é dada por: $a \times \eta = 1$.

$$x^{(n+1)} = x^{(n)} - \eta \nabla f[x^{(n)}] \tag{1}$$

A partir equação (1) e tendo em consideração que a optimização mais rápida ocorre em apenas uma iteração tem-se,

$$x^{(1)} = x^{(0)} - \eta \nabla f[x^{(0)}].$$

Onde, para esta classe de funções se verifica,

$$\nabla f(x) = ax$$

logo tem-se um mínimo global em $x = 0$.

Podemos então chegar ao seguinte resultado para $x^{(0)} \neq 0$:

$$0 = x^{(0)} - \eta a x^{(0)} \iff \eta a = 1.$$

Questão 2.1.3

Pelos dados obtidos verifica-se que o limite de divergência é dado pela relação: $a\eta = 2$.

A condição necessária para o método divergir é dada por,

$$|x^{(n+1)}| \geq |x^{(n)}|$$

a partir da equação (1), deduz-se

$$|x^{(n)} - \eta \nabla f[x^{(n)}]| \geq |x^{(n)}| \iff |x^{(n)} - \eta a x^{(n)}| \geq |x^{(n)}| \iff$$

$$|x^{(n)}||1 - \eta a| \geq |x^{(n)}| \iff |1 - \eta a| \geq 1$$

Para este tipo de funções como $a > 0$ e $\eta > 0$ tem-se

$$\eta a - 1 \geq 1, \text{ com } \eta a > 1 \iff \eta a \geq 2, \text{ com } \eta a > 1$$

Conclui-se assim que o limite de divergência é dado por $a\eta = 2$, para valores de $a\eta$ acima o método diverge e para valores abaixo converge.

Questão 2.1.4

Para valores de η baixos o método demora muito tempo a convergir. À medida que $a\eta$ aumenta observa-se uma redução do número de iterações. Para valores de $a\eta$ superiores a 1, que corresponde ao melhor caso, o desempenho do método piora. Quando $a\eta$ ultrapassa o limite de convergência, neste caso $a\eta = 2$, o método diverge.

Questão 2.1.5

Para todas as funções de uma só variável, a otimização mais rápida corresponde a uma só iteração. Se esta função for diferenciável existe um η para um dado $x^{(0)}$ que otimiza a função numa iteração.

Partindo mais uma vez da equação (1) tem-se,

$$x^{(1)} = x^{(0)} - \eta \nabla f[x^{(0)}].$$

Com $x^{(1)} \neq x^{(0)}$ e sendo $x^{(1)}$ o mínimo global, obtemos

$$\eta = \frac{x^{(0)} - x^{(1)}}{\nabla f[x^{(0)}]}.$$

Fica assim provada a sua existência.

2.2 Minimização das funções com mais de uma variável

Questão 2.2.1

Tabela 2

η	$a = 2$	$a = 20$
.01	414	414
.03	137	137
.1	40	Div
.3	12	Div
1	Div	Div
3	Div	Div
Mais rápido	.57 (5 it.)	.091 (44 it.)
Limiar de divergência	1	0.1

Questão 2.2.2

A partir da tabela 2 verificou-se que para vales mais estreitos, ou seja onde o valor de a é maior, o número de iterações mínimas que se consegue alcançar, com o método do gradiente descente, é superior relativamente a um vale mais largo. Também se confirmou que, para a=20, o método de otimização utilizado começa a divergir para valores de η menores.

Isto acontece porque para um vale mais largo, a=2 por exemplo, as curvas de nível da função são aproximadamente circunferências, logo o vetor ortogonal a estas, que corresponde ao gradiente, tem a direção do mínimo da função.

Por outro lado, para vales estreitos, a=20 por exemplo, as curvas de nível começam a tomar a forma de elipses e, consequentemente, a direção do gradiente

não está alinhada com a do mínimo. Deste modo, os valores de x iterados não se deslocam diretamente na direção deste, e o "caminho" percorrido para encontrar o mínimo é mais longo, o que corresponde a um maior número de iterações. Os valores de η também têm de ser baixos para evitar que o método de otimização divirja.

Questão 2.2.3

Não é sempre realizável pois, com o método do gradiente descente apenas é possível atingir o minimizante com uma iteração se o vetor ortogonal à curva de nível no ponto $\mathbf{x}^{(0)}$ estiver alinhado com o mínimo da função. Como vimos anteriormente, para um vale mais estreito as curvas de nível têm a forma de uma elipse, logo para vários $\mathbf{x}^{(0)}$, independentemente do valor de η que escolhermos não vai ser possível alcançar o minimizante com apenas uma iteração.

3. Termo de Momento

Questão 3.1

Tabela 3

η	$\alpha = 0$	$\alpha = .5$	$\alpha = .7$	$\alpha = .9$	α=.95
.003	>1000	>1000	>1000	>1000	>1000
.01	414	411	406	382	338
.03	137	134	129	96	171
.1	Div	36	31	85	122
.3	Div	Div	31	67	148
1	Div	Div	Div	74	146
3	Div	Div	Div	Div	172
10	Div	Div	Div	Div	Div
Limiar de divergência	0.1	0.35	0.57	1.9	3.9

Questão 3.2

Comparando os resultados obtidos para o método do gradiente descente na questão 2.2, para a=20, com os apresentados na tabela 3 verifica-se que, conforme se aumenta o valor de α, maior é o valor de η no limiar de divergência. Isto

acontece porque o termo de momento atenua as oscilações, logo é necessário um maior valor de η para o método oscilar demasiado e divergir.

Também se conclui que para um valor de α adequado consegue-se chegar ao mínimo com um número muito menor de iterações, relativamente ao método do gradiente descente.

4. Tamanho de passo adaptativo

Questão 4.1

O melhor par de valores que corresponde ao menor número de iterações encontrado é (α,η)=(0.93; 0.04)

Tabela 4

N. de testes	α	$\eta \rightarrow$	-20%	-10%	Best	+10%	+20%
10	0.93	N. de iterações $\rightarrow$	179	142	75	169	107

Questão 4.2

Foi mais complicado encontrar os parâmetros que produziam um número relativamente pequeno de iterações para a função de Rosenbrok. Enquanto que para os casos anteriores, para o mesmo α tínhamos um maior intervalo de valores de η para o qual o método de otimização utilizado não divergia. Nesta situação não se verificou o mesmo, bastava alterar-se ligeiramente o valor de η, para haver divergência ou o número de iterações aumentar radicalmente. A função de Rosenbrok tem partes onde é praticamente plana e outras com vales bastantes estreitos o que a torna bastante difícil de otimizar.

Questão 4.3

Tabela 5

η	$\alpha = 0$	$\alpha = .5$	$\alpha = .7$	$\alpha = .9$	α=.95	α=.99
.001	401	215	171	101	160	158
.01	384	201	168	165	145	139
.1	575	306	159	149	138	144
1	522	305	169	135	132	123
10	470	292	190	146	113	108

Questão 4.4

Tabela 6

	N. de testes	η	α	N. de iterações
Sem o tamanho do passo adaptado	21	-10%	0.9523	> 1000
		final $\eta = 0.01$		132
		+10%		507
Com o tamanho do passo adaptado	13	-10%	0.99	242
		final $\eta = 0.011$		147
		+10%		216

Questão 4.5

Neste trabalho laboratorial testaram-se vários métodos de aceleração e a sua eficácia. Começou-se por estudar o método do gradiente descente, onde se verificou que para funções com apenas uma variável este método convergia para o minimizante em apenas uma iteração, para um valor de η adequado. Logo, para funções desta classe não é necessário utilizar métodos mais complexos.

Para funções de ordem superior, o método do gradiente descente não é o mais adequado porque a sua eficácia está muito dependente do valor de η e $\mathbf{x}^{(0)}$ escolhidos. Não só pode oscilar muito e divergir facilmente, especialmente em funções com vales estreitos, como também, nas funções com mais de um mínimo, não é capaz de identificar se está perante um mínimo local ou global.

O método do termo de momento é uma solução que permite acelerar a otimização em situações onde a função tem vales estreitos, porque diminui as oscilações, observadas no método do gradiente descente, utilizando em cada passo, uma fração

do anterior. Logo é mais eficiente que método anterior.

Com a introdução da função de Rosenbrock, uma função difícil de otimizar, confirmou-se que é complicado escolher um par de valores α e η que levem o método do termo de momento a convergir. Deste modo, analisou-se o método com tamanho de passo adaptativo que apresenta uma pouca dependência do valor de η, e consequentemente torna mais fácil escolher parâmetros que resultam na eficácia da otimização. No entanto tem o custo de levar mais tempo a convergir.

Resumindo, a escolha do método depende da complexidade da função e do objectivo do utilizador. Para funções simples basta utilizar o método do gradiente descente, para funções mais complexas depende se o utilizador quiser uma resposta rápida mas que não têm a certeza se o método converge, ou uma solução mais lenta mas que existe uma maior probabilidade do método convergir.