Um novo método de otimização estocástico baseado no conceito de gradiente dilacional

Aline C. Soterroni[1], **Fernando M. Ramos**[2], **Roberto Luiz Galski**[3]

[1]Programa de Pós-Graduação em Computação Aplicada
Instituto Nacional de Pesquisas Espaciais (INPE)

[2]Laboratório Associado de Computação e Matemática Aplicada
Instituto Nacional de Pesquisas Espaciais (INPE)

[3]Centro de Controle de Satélite
Instituto Nacional de Pesquisas Espaciais (INPE)

{aline,fernando}@lac.inpe.br, galski@ccs.inpe.br

Resumo. *Neste trabalho estendemos o conceito de vetor gradiente para definir o vetor gradiente dilacional e apresentamos um novo método de otimização estocástico denominado método do gradiente dilacional. Esse método é aplicado em algumas funções teste e são apresentadas curvas do valor da função objetivo versus o número de avaliações da função. Os resultados obtidos são preliminares, porém mostram que as propriedades geométricas do vetor gradiente dilacional permitem que o algoritmo escape de ótimos locais, e que o método vai ao encontro do mínimo global de cada uma das funções teste analisadas.*

1. Introdução

O método do gradiente dilacional é um algoritmo estocástico de busca global que utiliza a direção contrária à direção do vetor gradiente dilacional como orientação para a sua busca. Considerando uma função de n variáveis, o vetor gradiente dilacional é o vetor composto pelas q-derivadas parciais de primeira ordem da função em relação a cada uma de suas coordenadas. E a q-derivada é a generalização do conceito de derivada no contexto do q-cálculo desenvolvida pelo reverendo inglês Frank Hilton Jackson no início do século XX [Jackson 1908, Jackson 1909, Jackson 1910a, Jackson 1910b].

O cálculo da q-derivada é baseado em dilatações por meio de um parâmetro q, e quando esse parâmetro tende a 1, a derivada de Jackson tende à derivada tradicional. Dessa forma, quando $q = 1$ o vetor gradiente por definição é o tradicional e o método se assemelha ao método da máxima descida. O parâmetro de dilatação q é obtido de forma estocástica por meio da escolha de um número aleatório em um intervalo especificado. Essa escolha aleatória implica em uma busca de caráter global ao longo de toda a execução do algoritmo. Um refinamento desse método foi possível com a introdução de características do algoritmo de recozimento simulado. Com isso o caráter global é controlado por uma função de probabilidade dependente de T, onde T é a temperatura. No início muitas configurações são aceitas, mas a medida que a temperatura é reduzida as configurações indesejáveis são rejeitadas.

Neste trabalho apresentamos resultados preliminares do desempenho desse novo algoritmo por meio de sua aplicação em algumas funções teste, são elas Ackley, De Jong,

Griewangk e Rastringin. A seguir são apresentados a definição do vetor gradiente dilacional para funções de n variáveis (Seção 2), o algoritmo (Seção 3), os resultados obtidos (Seção 4) e, por fim, algumas considerações finais (Seção 5).

2. Gradiente Dilacional

A derivada tradicional avalia o quanto uma dada função $f(x)$ é sensível à pequenas *translações* em sua variável independente por meio da equação

$$\frac{df(x)}{dx} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x}. \tag{1}$$

A generalização do conceito de derivada desenvolvida por Jackson é baseada em *dilatações* na variável independente, ou seja, em vez da variável independente x ser transladada por uma quantidade Δx, ela é dilatada por uma quantidade qx. Dessa forma, a derivada de Jackson, ou q-derivada, ou ainda operador q-diferença, de uma função $f(x)$ é dada por

$$D_q f(x) = \frac{f(qx) - f(x)}{qx - x}, \tag{2}$$

e no limite quando $q \to 1$, a q-derivada tende à derivada tradicional

$$\lim_{q \to 1} D_q f(x) = \frac{df(x)}{dx}. \tag{3}$$

Observe na Equação (2) que para $x = 0$ o denominador se anula. Dessa forma, para funções diferenciáveis em $x = 0$ uma definição para a q-derivada é dada por [Koekoev and Koekoev 1993]

$$D_q f(x) = \begin{cases} \frac{f(x) - f(qx)}{(1-q)x}, & x \neq 0, \;\; q \neq 1 \\ \frac{df(0)}{dx}, & x = 0\,. \end{cases} \tag{4}$$

A partir das Equações (3) e (5), podemos definir a q-derivada como

$$D_q f(x) = \begin{cases} \frac{f(qx) - f(x)}{qx - x} & \text{, se } q \neq 1, \;\; x \neq 0 \\ \frac{df(x)}{dx} & \text{, se } q = 1 \\ \frac{df(0)}{dx} & \text{, se } x = 0\,. \end{cases} \tag{5}$$

Estendendo o conceito de derivada de Jackson de uma função $f(\mathbf{x})$ com $\mathbf{x} \in \mathbb{R}^n$ diferenciável em $\mathbf{x} = 0$, denomina-se aqui o gradiente dilacional como o seguinte vetor composto pelas n q-derivadas parciais

$$\nabla_{\mathbf{q}} f(\mathbf{x}) = \left(\frac{D_{q_1} f(\mathbf{x})}{D_{q_1} x_1}, \ldots, \frac{D_{q_j} f(\mathbf{x})}{D_{q_j} x_j}, \ldots, \frac{D_{q_n} f(\mathbf{x})}{D_{q_n} x_n} \right), \tag{6}$$

em que, para qualquer $j \in \{1, \ldots, n\}$, a q-derivada parcial é dada por

$$\frac{D_{q_j} f(\mathbf{x})}{D_{q_j} x_j} = \frac{f(x_1, \ldots, q_j x_j, \ldots, x_n) - f(x_1, \ldots, x_j, \ldots, x_n)}{q_j x_j - x_j}. \tag{7}$$

Note que o parâmetro $\mathbf{q}$ assume valores distintos para dilatar cada uma das n coordenadas de $\mathbf{x}$. Logo, para funções de n variáveis esse parâmetro é dado pelo vetor $\mathbf{q} = (q_1, \ldots, q_i, \ldots, q_n)$.

Para $q = 1$ na Equação (5), a derivada tradicional é recuperada. Analogamente, para qualquer $q_j = 1$ em (6), a q-derivada parcial é igual a derivada parcial tradicional

$$\frac{D_{q_j} f(\mathbf{x})}{D_{q_j} x_j} = \frac{\partial f(\mathbf{x})}{\partial x_j} = \lim_{\Delta x_j \to 0} \frac{f(x_1, ..., x_j + \Delta x_j, ..., x_n) - f(x_1, ..., x_j, ..., x_n)}{\Delta x_j}. \tag{8}$$

E quando todos os q_j são iguais a 1, o vetor gradiente dilacional se torna igual ao vetor gradiente tradicional

$$\nabla_{\mathbf{q}} f(\mathbf{x}) = \nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_j}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_n} \right). \tag{9}$$

Igualmente, para $x = 0$ na Equação (5) a q-derivada é dada pela derivada tradicional aplicada no ponto zero. Logo, para qualquer $x_j = 0$ em (6), a q-derivada parcial é dada por

$$\frac{D_{q_j} f(\mathbf{x})}{D_{q_j} x_j} = \lim_{\Delta x_j \to 0} \frac{f(x_1, ..., 0 + \Delta x_j, ..., x_n) - f(x_1, ..., 0, ..., x_n)}{\Delta x_j}. \tag{10}$$

3. Algoritmo

Nesta seção introduzimos o algoritmo do método do gradiente dilacional. Esse algoritmo é inspirado no método da máxima descida, pois a cada iteração a direção de busca é a direção contrária à direção do vetor gradiente dilacional. Dessa forma, considerando uma função de n variáveis, temos que a cada iteração k o parâmetro $\mathbf{q}_k$ é obtido de forma estocástica, a direção de busca no ponto $\mathbf{x}_k$ é dada por

$$\mathbf{d}_k = -\nabla_{\mathbf{q}} f(\mathbf{x}_k) / ||\nabla_{\mathbf{q}} f(\mathbf{x}_k)||, \tag{11}$$

e o tamanho do passo é

$$\alpha_k = ||\mathbf{q}_k \mathbf{x}_k - \mathbf{x}_k||. \tag{12}$$

Lembrando que para funções $f : \mathbb{R}^n \to \mathbb{R}$ o parâmetro $\mathbf{q}_k$, a direção $\mathbf{d}_k$ e o ponto $\mathbf{x}_k$ são vetores de n posições na iteração k.

Note que o tamanho do passo é exatamente a dilatação exercida pelo parâmetro $\mathbf{q}$ sobre o ponto $\mathbf{x}$. A estratégia utilizada para a obtenção desse parâmetro é dada pela geração de números aleatórios uniformemente distribuídos em um intervalo $[q_{min}, q_{max}]$. Seja $\mathbf{r}$ um número aleatório uniformemente distribuído no intervalo $[0, 1]$ e $\mathbf{r} \in \mathbb{R}^n$, então o parâmetro $\mathbf{q}$ a cada iteração é dado por

$$\mathbf{q}_k = q_{min} + \mathbf{r} \cdot (q_{max} - q_{min}). \tag{13}$$

Logo, o novo ponto $\mathbf{x}_{k+1}$ da busca é

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k. \tag{14}$$

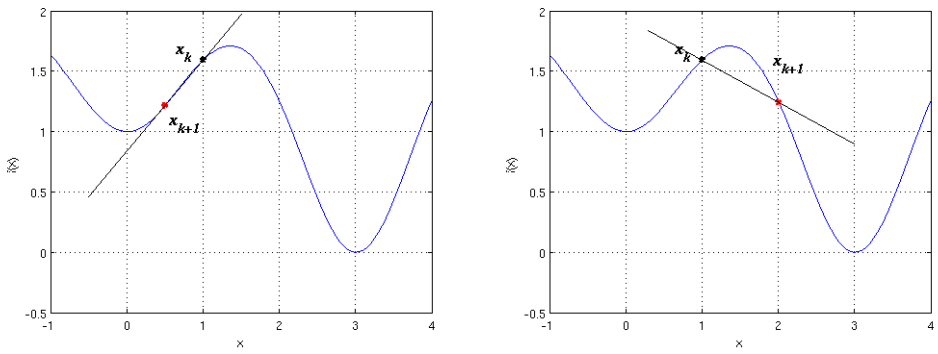
O ponto inicial $\mathbf{x}_0$ é obtido de forma aleatória no espaço de busca viável. O critério de parada pode ser dado pelo número de avaliações da função objetivo ou por um número máximo de iterações. E o ponto de mínimo será o ponto $\mathbf{x}_k$ que atingir o menor valor da função objetivo ao longo de todas as iterações.

A Figura 1 ilustra, geometricamente, como a propriedade de dilatação do vetor gradiente dilacional permite que os pontos da busca escapem de mínimos locais. Para isso, considere a função $f : \mathbb{R} \to \mathbb{R}$ dada por

$$f(x) = 2 - (e^{-x^2} + 2e^{-(x-3)^2}). \tag{15}$$

As retas secantes passando pelos pontos $(x_k, f(x_k))$ e $(x_{k+1}, f(x_{k+1}))$ na Figura 1, fornecem uma interpretação geométrica para a derivada de Jackson dada pela Equação (2). Para $x_k = 1.0$ e $q_k = 0.5$ (veja a Figura 1a) o próximo ponto da busca de acordo com as Equações (11), (12) e (14) é $x_{k+1} = 0.5$. Observe que a reta secante tem inclinação positiva e como se trata de uma função unidimensional, podemos dizer que o vetor gradiente dilacional em x_k aponta para a direita. Como a direção de busca é contrária à direção do vetor gradiente, então o ponto $x_{k+1} = 0.5$ encontra-se a esquerda de $x_k = 1.0$.

Analogamente, para $q_k = 2.0$ o próximo ponto é $x_{k+1} = 2.0$ localizado a direita de $x_k = 1.0$, pois a reta secante tem inclinação negativa (veja a Figura 1b). Assim, o ponto x_{k+1} pode se localizar tanto a direita quanto a esquerda do ponto x_k. Observe também que para $q_k = 0.5$ o ponto x_{k+1} se aproximou do mínimo local da função f. E para $q_k = 2.0$ o ponto x_{k+1} escapa da bacia de atração do mínimo local, e cai na bacia de atração do mínimo global dessa função. É a dilatação $q_k x_k$ que define a direção do vetor gradiente dilacional e que permite passos grandes ou pequenos nessa direção. Portanto, q é o parâmetro de ajuste fundamental para métodos de otimização que utilizem o conceito de vetor gradiente dilacional.



(a) $q_k = 0.5$ (b) $q_k = 2.0$

Figura 1. Interpretação geométrica da derivada de Jackson e da direção de busca do algoritmo para a Função (15) com $x_k = 1.0$ e diferentes valores de q_k.

Nem sempre a dilatação leva o ponto atual $\mathbf{x}_k$ para um novo ponto $\mathbf{x}_{k+1}$ de tal forma que o valor da função objetivo diminua. Porém permitir valores piores do ponto de

vista da minimização da função objetivo pode evitar que a busca fique presa aos mínimos locais do espaço de busca viável.

A estratégia proposta para o cálculo do parâmetro $\mathbf{q}$ é simples: gerar números aleatórios no intervalo $[q_{min}, q_{max}]$. Com isso, a direção de busca e o tamanho do passo podem variar muito de uma iteração para a outra. Ou seja, ao longo de toda a execução do algoritmo, a dilatação exercida por $\mathbf{q}_k$ pode levar o ponto $\mathbf{x}_{k+1}$ para qualquer lugar do espaço de busca, sem considerar melhoras ou pioras no valor da função objetivo.

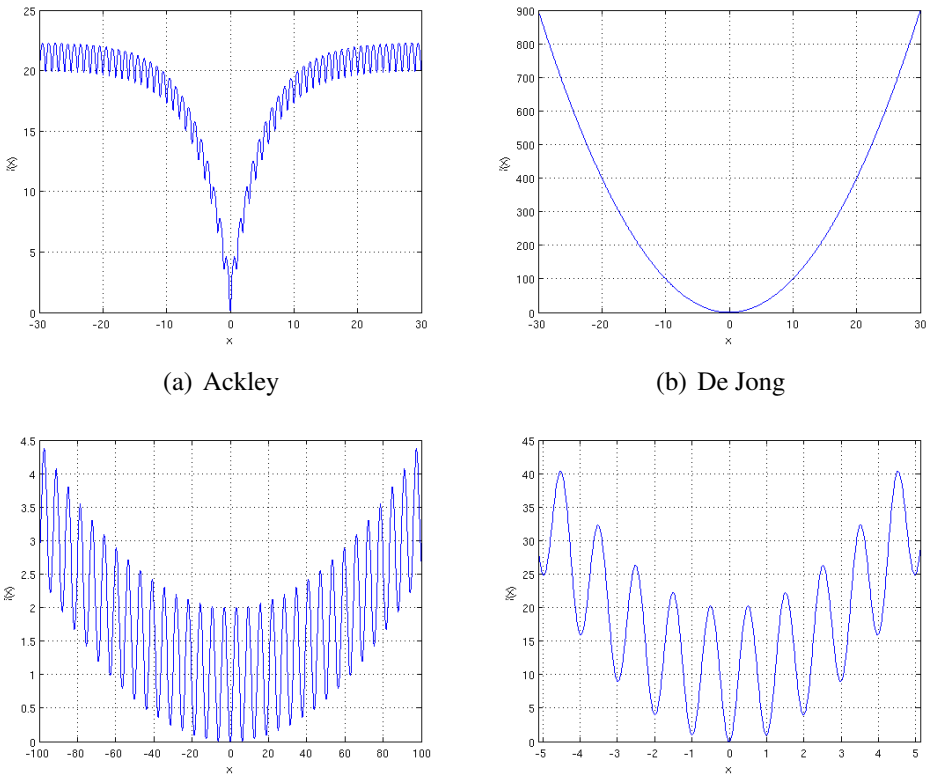
Um refinamento dessa idéias foi obtido com a introdução de características do método recozimento simulado [Kirkpatrick et al. 1983] no processo de busca. A cada iteração o parâmetro $\mathbf{q}$ é obtido de forma estocástica. Em seguida, verificamos se a dilatação $\mathbf{q}_k\mathbf{x}_k$ gera uma melhora no valor da função objetivo. Para isso, calculamos o novo ponto $\mathbf{x}_{k+1}$ e fazemos $\Delta f = f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)$. Se $\Delta f < 0$, aceitamos o novo ponto. Senão ($\Delta f > 0$), a probabilidade do novo ponto ser aceito é dada por $\exp(-\Delta f/T)$, em que T é a temperatura. Geramos um número aleatório r uniformemente distribuído no intervalo $[0, 1]$, e se $r < \exp(-\Delta f/T)$, então o novo ponto é aceito e T é reduzido por meio da relação $T = T \cdot \beta$. No início, o parâmetro de controle T possui um valor alto e a busca assume caráter global, pois todas as configurações são aceitas, mas a medida que T é reduzido com taxa β, as configurações indesejáveis são rejeitadas.

A seguir apresentamos um pseudo-código para o método do gradiente dilacional.

Gere uma solução inicial $\mathbf{x}_0$
Faça $\mathbf{x}_{melhor} = \mathbf{x}_0$
Faça $T = T_0$
Enquanto $k = 0, 1, \ldots,$ até atingir o critério de parada
    Calcule o parâmetro $\mathbf{q}_k$ (Equação 13)
    Calcule a direção $\mathbf{d}_k$ (Equação 11)
    Calcule o passo α_k (Equação 12)
    Calcule o novo ponto $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$
    Calcule $\Delta f = f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)$
    Se $\Delta f < 0$ ou $r < exp(-\Delta f/T)$
        $\mathbf{x}_k = \mathbf{x}_{k+1}$
        Se $f(\mathbf{x}_{k+1}) < f(\mathbf{x}_{melhor})$, então $\mathbf{x}_{melhor} = \mathbf{x}_{k+1}$
        $T = T \cdot \beta$
    Fim Se
    $k = k + 1$
Fim Enquanto
Apresente $\mathbf{x}_{melhor}$ e $f(\mathbf{x}_{melhor})$.

4. Resultados

Consideramos as funções teste Ackley, De Jong, Griewangk e Rastringin, para o caso unidimensional (veja a Figura 2).



(a) Ackley (b) De Jong

(c) Griewangk (d) Rastringin

Figura 2. Funções Teste - caso unidimensional.

A função Ackley é definida como [Potter and Jong 1994]:

$$f(\mathbf{x}) = -20\mathbf{exp}\left(-0.2\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}\right) - \mathbf{exp}\left(\frac{1}{n}\sum_{i=1}^{n}(2\pi x_i)\right) + 20 + \mathbf{exp}^1. \quad (16)$$

em que n é a dimensão e $-30 \leqslant x_i \leqslant 30$. A função De Jong é dada por [Pohlheim 1994]

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2 \quad (17)$$

em que n é a dimensão e $-5.12 \leqslant x_i \leqslant 5.12$. A função Griewangk é definida por [Potter and Jong 1994]:

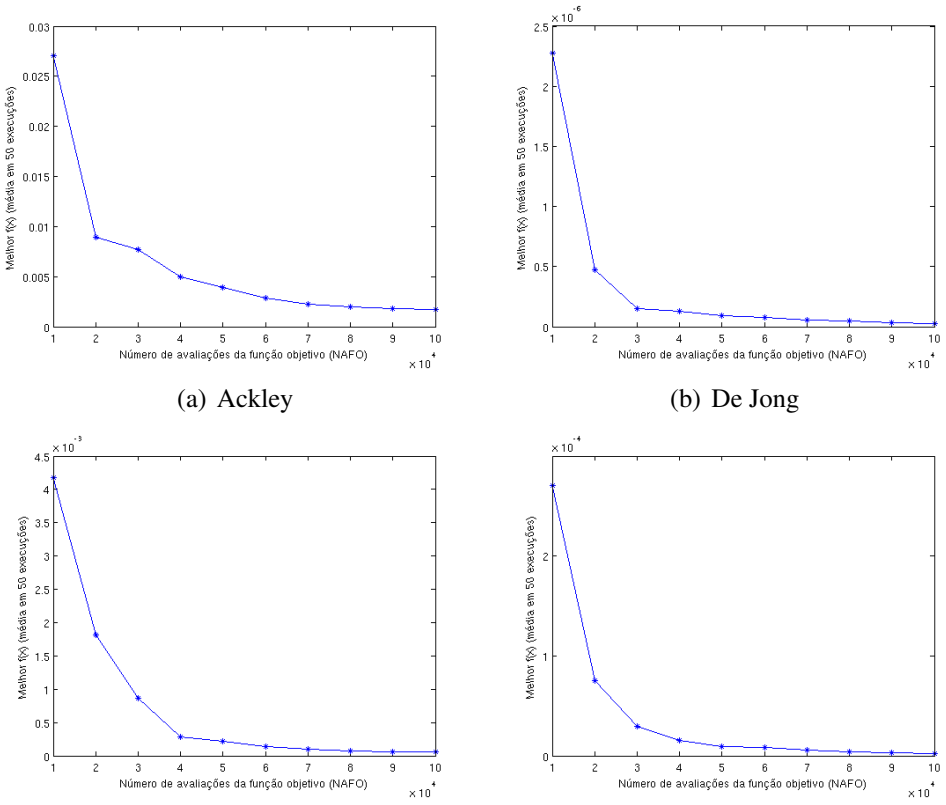
$$f(\mathbf{x}) = 1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} cos\left(\frac{x_i}{\sqrt{i}}\right), \quad (18)$$

em que n é a dimensão e $-600 \leqslant x_i \leqslant 600$. E a função Rastringin é dada por [Potter and Jong 1994]:

$$f(\mathbf{x}) = 3.0n + \sum_{i=1}^{n} x_i^2 - 3.0cos(2\pi x_i), \quad (19)$$

em que n é a dimensão e $-5,12 \leqslant x_i \leqslant 5,12$. Para todas essas funções o mínimo global é zero no ponto $\mathbf{x} = \mathbf{0}$.

Foram realizadas 50 execuções independentes do algoritmo para as funções teste selecionadas com $q \in [1/2, 2]$, $T = 2.5$, taxa de redução $\beta = 0.9$ e 100 mil avaliações da função objetivo (NAFO). A Figura 3 ilustra os resultados para os melhores valores da função objetivo versus o número de avaliações.



(a) Ackley (b) De Jong

(c) Griewangk (d) Rastringin

Figura 3. Curvas dos melhores valores da função objetivo versus o número de avaliações da função para diferentes funções teste.

A Tabela 1 sumariza os resultados obtidos para cada função teste por meio da média e do desvio padrão dos melhores valores de $f(x)$ ao longo de 50 execuções independentes do algoritmo.

Tabela 1. Resultados para as diferentes funções teste

Funções	Melhor Valor Médio	Desvio Padrão
Ackley	1.7027e-03	1.5059e-03
De Jong	2.8464e-08	9.4642e-08
Griewangk	5.4227e-05	1.0367e-04
Rastringin	2.3751e-06	5.0157e-06

5. Considerações Finais

Neste trabalho estendemos o conceito de gradiente, definimos o vetor gradiente dilacional para funções de n variáveis e apresentamos um algoritmo para o método do gradiente dilacional. Verificamos que a estratégia de escolha aleatória do parâmetro q implica em uma busca de cárater global ao longo das iterações do algoritmo. Assim, com a introdução de características do método de recozimento simulado foi possível refinar esse algoritmo e permitir que a busca assuma um caráter global apenas no início, pois a temperatura é alta e muitas configurações são aceitas, mas a medida que a temperatura diminui mais pontos são rejeitados e apenas configurações melhores são implementadas.

Para as funções teste analisadas a aplicação do algoritmo exibe resultados que se aproximam do mínimo global de cada função, exceto para a função Ackley que apresentou a pior precisão encontrada. Esses resultados são iniciais mas mostram que o algoritmo vai de encontro ao mínimo global de cada função objetivo e para que a precisão aumente, estratégias mais sofisticadas para o cálculo de q devem ser elaboradas.

Como vimos, para determinados valores do parâmetro q um ponto localizado na bacia de atração de um mínimo local pode dar um passo grande e escapar desse mínimo. É a dilatação exercida por esse parâmetro que define a direção de busca e também permite passos grandes ou pequenos nessa direção. Portanto, q é o parâmetro de ajuste fundamental para métodos de otimização baseados no conceito de vetor gradiente dilacional.

Referências

Jackson, F. H. (1908). On q-functions and a certain difference operator. volume 46, pages 253–281. Trans. Roy. Soc. Edinburg.

Jackson, F. H. (1909). A q-form of Taylor's theorem. volume 38, pages 62–64. Mess. Math.

Jackson, F. H. (1910a). On q-definite integrals. volume 41, pages 193–203. Quart. J. Pure Appl. Math.

Jackson, F. H. (1910b). q-difference equations. volume 32, pages 305–314. American J. Math.

Kirkpatrick, S., Jr., C. D. G., and Vecchi, M. P. (1983). Optimizing by simulated annealing. volume 220, pages 671–680. Science.

Koekoev, J. and Koekoev, R. (1993). A note on the q-derivative operator. volume 176, pages 627–634. Journal of Mathematical Analysis and Applications.

Pohlheim, H. (1994). Geatbx: Example functions (single and multi-objective functions) 2 parametric optimization. Disponível em http://www.geatbx.com/docu/fcnindex-01.html. Acesso em 12 ago. 2008.

Potter, M. A. and Jong, K. A. D. (1994). A cooperative coevolutionary approach to function optimization. pages 249–257. Anais do The Third Parallel Problem Solving From Nature, Springer-Verlag.