X Congresso Brasileiro de Inteligência Computacional (CBIC'2011), 8 a 11 de Novembro de 2011, Fortaleza, Ceará

ESTUDO DOS MECANISMOS DE FUNCIONAMENTO DO ALGORITMO DE EVOLUÇÃO DIFERENCIAL

Fillipe Goulart, Felipe Campelo, Jaime A. Ramírez
Departamento de Engenharia Elétrica, Universidade Federal de Minas Gerais
{fgsm-c,fcampelo,jramirez}@ufmg.br

Resumo – Neste trabalho são apresentadas considerações teóricas e estudos experimentais sobre os mecanismos de funcionamento do algoritmo de evolução diferencial (DE) para otimização não-linear. Uma relação constante entre a matriz de covariância da população em uma dada iteração do algoritmo e a covariância correspondente da distribuição de probabilidade da mutação diferencial, representada pelo conjunto de vetores-diferença possíveis, é obtida e utilizada para a análise do comportamento adaptativo do algoritmo.

Palavras-chave – otimização, algoritmos evolutivos, evolução diferencial.

Abstract – This work presents theoretical considerations and experimental studies on the working mechanisms of the differential evolution (DE) algorithm for nonlinear optimization. A constant relationship between the covariance matrix of the population at a given iteration of the algorithm, and the corresponding covariance matrix of the probability distribution for the differential mutation operator, represented by the set of all possible mutation vectors, is obtained. This relationship is then employed for analyzing the adaptive behavior of the differential evolution algorithm.

Keywords – optimization, evolutionary algorithms, differential evolution.

1. INTRODUÇÃO

Ao longo das últimas décadas, algoritmos evolutivos têm conquistado crescente atenção em suas aplicações para otimização em diversos campos de ciência e engenharia. Particularmente para o caso de problemas não-lineares com variáveis contínuas, o algoritmo de Evolução Diferencial (*differential evolution* - DE) [1, 2] tem se mostrado como uma ferramenta eficaz, robusta e versátil para a exploração eficiente de espaços de busca com características reconhecidamente desafiadoras para outras heurísticas evolutivas, como a existência de interações fortes entre variáveis ou espaços de busca de dimensão elevada.

Apesar da diversidade e popularidade das aplicações deste algoritmo, avanços proporcionais na compreensão de seus mecanismos de funcionamento não tem sido observados na literatura. Embora ponderações sobre o tema não sejam inéditas na literatura científica [3], há uma pronunciada escarsidade de investigações sobre a dinâmica e propriedades de adaptação dos mecanismos de variação da população do DE, responsáveis pela capacidade de exploração global e local deste método.

O presente trabalho propõe uma investigação rigorosa destes mecanismos de adaptação, e da dinâmica evolutiva da população nos algoritmos de evolução diferencial. Em particular, são investigadas as propriedades do operador de mutação diferencial, principal responsável pela capacidade autoadaptativa do DE a diferentes funções objetivo. Neste estudo inicial, serão utilizados problemas irrestritos com funções-objetivo quadráticas arbitrárias como problemas-modelo. Apesar de simples, estes problemas podem fornecer informações importantes sobre a dinâmica de adaptação dos vetores de mutação diferencial, possibilitando o estabelecimento de diretivas e procedimentos analíticos e experimentais para uma investigação mais sofisticada dos algoritmos de evolução diferencial.

Este artigo está organizado da seguinte forma: na seção 2 são descritos o algoritmo básico de evolução diferencial e a forma de implementação de seus operadores; a seção 3 inclui considerações teóricas e o resultado de investigações, tanto analíticas quanto experimentais, sobre a dinâmica do algoritmo. Algumas das consequências dos resultados obtidos são discutidas na seção 4, com a definição de algumas conjecturas e ideias para a continuidade desta investigação. Finalmente, a seção 5 traz as considerações finais e conclusões do trabalho.

2. O ALGORITMO EVOLUÇÃO DIFERENCIAL

Evolução diferencial é um algoritmo evolutivo bastante simples e extremamente eficiente para otimização contínua [2]. Ao contrário de muitos outros, ele não faz uso explícito [1] de distribuições de probabilidade para efetuar a mutação: os parâmetros empregados na variação são os próprios indivíduos da população, combinados por meio de operações simples de soma vetorial e multiplicação escalar.

[1] O aspecto das equações nada remete a distribuições de probabilidade. Entretanto, como será visto neste trabalho, o efeito da mutação no algoritmo de evolução diferencial é semelhante ao acréscimo de uma perturbação de distribuição conhecida a um dado ponto no espaço.

Assim como outros algoritmos evolutivos, o DE é passível de sofrer modificações, numa tentativa de adequar seu comportamento diante de cada tipo de problema. Não obstante, é verificado que essas mudanças podem acarretar resultados melhores para uma determinada classe de problemas, embora frequentemente ao custo de degradação do desempenho em outras. De forma geral, o DE básico se mostra suficientemente bom para problemas de otimização em espaços contínuos de problemas de otimização aplicada a engenharia, razão pela qual esta instância será utilizada para as investigações descritas neste texto. O leitor interessado nas variações deste método é remetido à vasta literatura disponível sobre o tema [2, 4].

A estrutura do algoritmo de evolução diferencial segue uma sequência relativamente simples: considere uma população de vetores $\mathbf{x}_i \in \mathbb{R}^n$, descrita por uma matriz $\mathbf{P} \in \mathbb{R}^{\mu \times n}$, ou seja, contendo μ vetores (*indivíduos*) n-dimensionais. Esta população de pontos distribuídos no espaço de variáveis de otimização é iterativamente modificada através da aplicação de operadores de seleção e variação, de forma a promover a exploração do espaço em diversas escalas, da busca global inicial a um refinamento local nas etapas finais do algoritmo.

A seguir, são apresentados os operadores utilizados pelo DE para esta modificação iterativa da população $\mathbf{P}$, na ordem em que geralmente são aplicados.

2.1 Mutação

O procedimento de mutação no algoritmo de evolução diferencial passa pela construção de um vetor diferencial $\mathbf{v}_i$ e sua adição a um ponto qualquer da população $\mathbf{P}$. Para a composição de $\mathbf{v}_i$, dois indivíduos quaisquer da população, $\mathbf{x}_{i2}$ e $\mathbf{x}_{i3}$, são subtraídos entre si:

$$\mathbf{v}_i \stackrel{\triangle}{=} \mathbf{x}_{i2} - \mathbf{x}_{i3}$$

Para a composição do ponto mutado $\mathbf{u}_i$ (também chamado frequentemente de *vetor de teste*), um terceiro indivíduo qualquer $\mathbf{x}_{i1}$ é adicionado a uma ponderação do vetor diferencial, na forma:

$$\mathbf{u}_i = \mathbf{x}_{i1} + F\mathbf{v}_i \tag{1}$$

onde $F \in \mathbb{R}$ é um dos parâmetros ajustáveis do DE, conhecido como *fator de escala*. Uma boa escolha de F depende um pouco do problema em questão. A literatura técnica sobre o DE tende a recomendar $0 < F < 1$, sendo $F = 0,6$ um valor considerado razoável para diversas situações [5], e $F > 1$ não recomendado [4]. Outra prática recomendada, e adotada neste trabalho, consiste garantir a utilização de vetores diferentes entre si para a geração de $\mathbf{u}_i$, ou seja, $\mathbf{x}_{i1} \neq \mathbf{x}_{i2} \neq \mathbf{x}_{i3} \neq \mathbf{x}_i$.

2.2 Recombinação

Este operador é o responsável pela criação de um novo indivíduo $\tilde{\mathbf{x}}_i$, que será posteriormente apresentado ao operador de seleção. Na geração de $\tilde{\mathbf{x}}_i$, são utilizados o i-ésimo ponto da população, $\mathbf{x}_i$; e o i-ésimo vetor de teste, $\mathbf{u}_i$. No tipo de cruzamento aqui utilizado, denominado *binomial*, cada componente de $\tilde{\mathbf{x}}_i$ tem uma dada probabilidade de ser proveniente de $\mathbf{x}_i$ ou do vetor de teste correspondente. Matematicamente, a j-ésima componente do vetor recombinado[2] será dada por

$$\tilde{x}_{i,j} = \begin{cases} u_{i,j} & \text{se} \quad U_j \leq CR \\ x_{i,j} & \text{caso contrário} \end{cases} \tag{2}$$

em que U_j é uma variável aleatória de distribuição uniforme no intervalo $(0, 1)$, e $CR \in (0, 1)$ é outro parâmetro do DE, denominado *fator de cruzamento*. Resumidamente, este parâmetro determina o "grau de parentesco" de $\tilde{\mathbf{x}}_i$ com $\mathbf{u}_i$ (CR grande) ou com $\mathbf{x}_i$ (CR pequeno). Em diversos casos, este parâmetro pode ser entendido como a "probabilidade de cruzamento". Estudos na literatura recomendam $CR \approxeq 0,8$ [5].

Existe um outro tipo de recombinação comumente utilizado em evolução diferencial, nomeado *exponencial*, que não será de interesse nesse trabalho. Como última nota, vê-se que a equação (2) pode ter seus sinais de desigualdade invertidos livremente, desde que a ideia descrita no parágrafo anterior seja similarmente ajustada.

2.3 Seleção

Mais simples que os outros dois operadores, a seleção no DE é puramente determinística: compara-se o desempenho do i-ésimo pai, $\mathbf{x}_i$, com o do indivíduo mutado, $\tilde{\mathbf{x}}_i$, selecionando aquele que apresentar melhor valor para compor a próxima população. Assumindo um problema de minimização, o i-ésimo indivíduo da $(t+1)$-ésima iteração é selecionado da forma:

$$\mathbf{x}_i(t+1) = \begin{cases} \mathbf{x}_i(t) & \text{se} \quad f(\mathbf{x}_i(t)) \leq f(\tilde{\mathbf{x}}_i(t)) \\ \tilde{\mathbf{x}}_i(t) & \text{caso contrário} \end{cases} \tag{3}$$

2.4 O DE básico

Com todos os blocos construtivos do DE discutidos, basta apenas uni-los. Isso é feito no algoritmo 1. Note que ele é válido tanto para minimização quanto para maximização de uma função.

[2]Uma vez que o ponto "mutado" $\tilde{\mathbf{x}}_i$ pode ser efetivamente criado durante processo de seleção, não é raro que o DE seja descrito de forma resumida, sem menção explícita aos operadores de mutação e recombinação.

Algorithm 1: Evolução Diferencial básico

início
  Inicialize o contador de iterações $t \leftarrow 0$
  Inicialize os parâmetros de controle F e CR
  Inicialize a população $\mathbf{P}(t)$ com μ pontos n-dimensionais
  Avalie o desempenho $f(\mathbf{x}_i(t))$ de todos os pontos $\mathbf{x}_i(t) \in \mathbf{P}(t)$
  enquanto *Condição de parada não for atingida* **faça**
    para todo $\mathbf{x}_i(t) \in \mathbf{P}(t)$ **faça**
      Crie o indivíduo de teste, $\mathbf{u}_i(t)$; /* Eq. (1) */
      Crie o indivíduo recombinado, $\tilde{\mathbf{x}}_i(t)$; /* Eq. (2) */
      Avalie o desempenho $f(\tilde{\mathbf{x}}_i(t))$
      se $f(\mathbf{x}_i(t))$ *é melhor que* $f(\tilde{\mathbf{x}}_i(t))$ **então** /* (Eq. (3)) */
        $\mathbf{x}_i(t+1) \leftarrow \mathbf{x}_i(t)$;
      senão
        $\mathbf{x}_i(t+1) \leftarrow \tilde{\mathbf{x}}_i(t)$
      fim
    fim
    $t \leftarrow t+1$
  fim
  Retorne o indivíduo com melhor desempenho como solução
fim

Para expressar as mais diversas instâncias nas quais o DE é normalmente utilizado, emprega-se a notação padrão DE/x/y/z, em que x representa o método de seleção o vetor base; y indica o número de vetores-diferença empregados na etapa de mutação; e z o método de cruzamento usado. Assim, pode-se resumir o algoritmo empregado neste trabalho como **DE/rand/1/bin**, que significa escolha aleatória (rand) do vetor base, um único vetor diferença utilizado (1), e método de cruzamento binomial (bin). Outras variações do DE podem ser encontradas na literatura [4, 5].

Uma vez delineada a estrutura do algoritmo tratado nesse trabalho, parte-se para um estudo mais aprofundado acerca de seus mecanismos de funcionamento.

3. CONSIDERAÇÕES TEÓRICAS

Apesar da vasta utilização da evolução diferencial em diversas aplicações de otimização, observa-se na literatura uma carência de trabalhos abordando o problema de modelagem estatística e matemática dos mecanismos de operação deste algoritmo, bem como da caracterização de sua dinâmica e comportamento em famílias de funções-objetivo. O presente trabalho propõe-se a investigar certas características comportamentais do DE, bem como seus mecanismos de funcionamento. Para isso, é necessário fixar um foco antes de partir para generalizações.

Neste trabalho, serão utilizadas funções-objetivo quadráticas, cujas superfícies de nível representem elipsoides rotacionados em relação aos eixos coordenados, ou seja, funções da forma:

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{H}\mathbf{x} + \mathbf{d}^T\mathbf{x} + c \tag{4}$$

com a matriz Hessiana H constante, simétrica e definida positiva. Embora esta classe trate de problemas relativamente simples, o estudo do comportamento do DE nesta família de funções pode fornecer informações importantes sobre os mecanismos de funcionamento deste algoritmo, bem como pistas para sua dinâmica em funções de maior complexidade. A evolução diferencial utilizada para este estudo será a instância mais comumente encontrada na literatura, **DE/rand/1/bin**, com parâmetros $F = 0,6$ e $CR = 0,8$. Os critérios de parada adotados, quando necessário, serão definidos como a execução de um número máximo de iterações ou de avaliações de função, ou ainda estabilização da população em torno de um dado ponto.

3.1 Considerações Iniciais

Através de observações preliminares da dinâmica do algoritmo em problemas com duas dimensões, percebeu-se uma tendência da população de se distribuir ao longo das direções dos semieixos principais dos elipsóides de nível do problema, independente da rotação da função objetivo[3]. Após este alinhamento inicial, a população tende a comprimir-se rumo ao ótimo, resultando no processo de convergência à solução do problema.

Os vetores de diferenças também apresentam a mesma característica, ilustrada na Fig. 1. Esta figura ilustra as curvas de nível de uma função quadrática rotacionada em relação aos eixos x e y, sobrepostas ao conjunto de todos os possíveis vetores-diferença de uma população normalizada no intervalo [0,1]. Pode-se perceber que, a partir de uma distribuição inicial sem uma estrutura definida, os vetores-diferença começam a apontar nas direções das curvas de nível com o passar das iterações. Além disto, pode-se notar uma redução progressiva em sua amplitude, indicando a ocorrência de passos de mutação menores à medida em que o algoritmo caminha rumo à convergência.

[3] O que está de acordo com o fato de o comportamento do DE ser invariante à rotação de coordenadas [4]

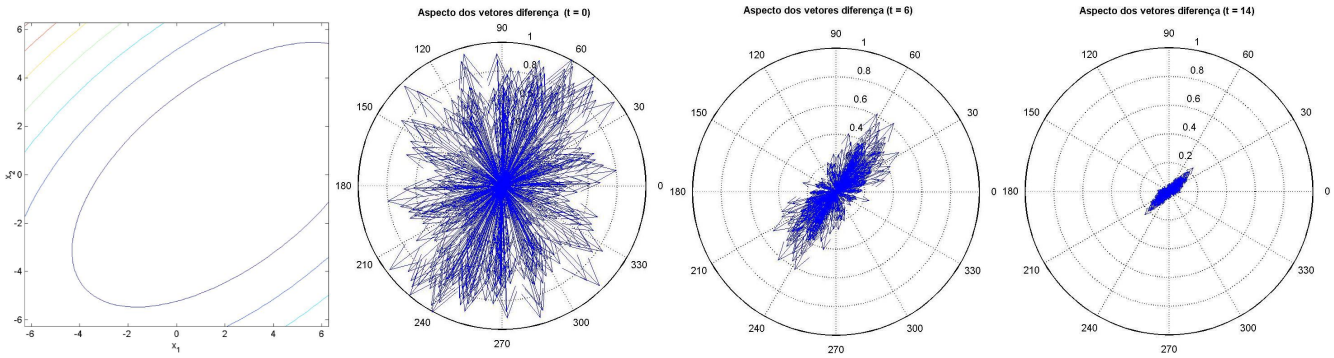

Figura 1: Curvas de nível de uma função quadrática, de duas variáveis, e comportamento dos vetores-diferença no decorrer do algoritmo. Em $t = 0$, as setas são dispostas ao acaso. Após $t = 6$ iterações, já é possível notar um certo alinhamento em relação às curvas de nível, e uma pequena redução da magnitude de cada vetor. Em $t = 14$, as setas já estão bem menores e mais alinhadas, indicando uma possível convergência da população. A partir da iteração vinte (em média), as setas eram tão pequenas que a escala adotada perdia a capacidade de discriminação entre os vetores.

Essa observação sugere um aspecto interessante do algoritmo. Examinando a equação (1), vemos que cada vetor de teste é obtido a partir de adição a um dado ponto da população de um vetor que, como observado, é obtido a partir de um conjunto de possíveis vetores-diferença que tende a se alinhar de acordo com as características da função objetivo, com uma magnitude que se ajusta ao longo das iterações de forma a explorar diferentes escalas desta função. Estas observações nos permitem construir a seguinte conjectura: o *DE funciona por meio da adaptação do mecanismo de variação (mutação diferencial) à forma da função objetivo*. Note o cuidado que deve ser tomado ao expressar esta conjectura: nada é dito a respeito da forma da função objetivo. Neste trabalho, isso é verificado para funções quadráticas, conforme explicitado anteriormente. Há suspeitas que este comportamento também se verifique para problemas mais gerais, e isso é mencionado na seção de discussão.

3.2 Relação entre as distribuições da população e dos vetores-diferença

As imagens da Fig. 1, juntamente com as equações que regem os operadores do DE, permitem suspeitar que há uma forte relação entre a distribuição da população e a dos vetores-diferença. A abordagem utilizada para tratar isso é por meio de suas propriedades estatísticas, mais precisamente, suas matrizes de covariância, $\mathbf{C}_P$ (da população), e $\mathbf{C}_V$ (dos vetores-diferença). De fato, há uma relação muito simples entre elas, descrita pelo teorema 3.1.

Teorema 3.1. (Relação entre $\mathbf{C}_P$ e $\mathbf{C}_V$) *Seja $\mathbf{P} \in \mathbb{R}^{\mu \times n}$ uma matriz contendo μ vetores n-dimensionais, e $\mathbf{V} \in \mathbb{R}^{\mu(\mu-1) \times n}$ uma matriz contendo todas os possíveis vetores-diferença gerados a partir de pontos distintos contidos em $\mathbf{P}$. Sejam ainda $\mathbf{C}_P \in \mathbb{R}^{n \times n}$ e $\mathbf{C}_V \in \mathbb{R}^{n \times n}$ as matrizes de covariância amostral obtidas a partir dos vetores em $\mathbf{P}$ e $\mathbf{V}$, respectivamente. Dadas estas matrizes, pode-se mostrar que ambas são relacionadas por um escalar dependente apenas do valor de μ:*

$$\mathbf{C}_V = \frac{2\mu(\mu-1)}{\mu(\mu-1)-1}\mathbf{C}_P = \kappa \mathbf{C}_P$$

A prova deste teorema é descrita no Apêndice A. O efeito desta relação, no caso do DE, é uma relação direta entre a população e a distribuição dos vetores-diferença, que por sua vez influenciam fortemente na distribuição da população nas iterações posteriores, o que sugere um sistema de autoadaptação dos mecanismos de variação do algoritmo de evolução diferencial. Outra consequência desta relação é a possibilidade de se investigar certos aspectos estatísticos do operador de mutação diferencial a partir da dinâmica evolutiva da população do DE, e vice-versa. Esta propriedade pode ser explorada, por exemplo, no estudo do alinhamento da população com as hipercurvas de nível de uma função objetivo.

3.3 Alinhamento com as Hipercurvas de Nível da Função Objetivo

Como ilustrado na Fig. 1, observa-se que os vetores-diferença (e, consequentemente, os pontos da população) exibem uma certa tendência ao alinhamento com as hipercurvas de nível de uma função objetivo quadrática. Este alinhamento pode ser expresso em termos dos autovetores de $\mathbf{C}_V$ (ou $\mathbf{C}_P$), que tendem a ficar paralelos aos eixos principais das elipses.

Considerando que as funções estudadas neste artigo são descritas por (4), tem-se que as mesmas possuem uma matriz Hessiana $\mathbf{H}$ constante e invertível, com os autovetores desta matriz apontando na direção dos eixos principais das hipercurvas de nível da função. Assim sendo, estes podem ser utilizados como referência no alinhamento entre a população e a função objetivo. Entretanto, os autovalores da Hessiana são proporcionais à curvatura da função objetivo, enquanto que as hipercurvas de nível tendem a ser inversamente proporcionais a esta característica. Desta forma, para fins desta análise, é preferível utilizar os autovetores e autovalores da inversa da Hessiana, $\mathbf{H}^{-1}$: os primeiros ainda formam um eixo que acompanha as hipercurvas de nível, enquanto que os últimos escalonam corretamente os autovetores de acordo com a "suavidade" da função na direção apontada.

Uma ilustração desse processo em duas dimensões pode ser observada na Fig. 2, que exibe os autovetores escalonados da inversa da Hessiana (fixos nas quatro figuras), e os da matriz de covariâncias da população[4] . Vale lembrar que, como $\mathbf{H}^{-1}$ e $\mathbf{C}_P$ são simétricas, seus respectivos autovetores serão perpendiculares entre si, como se vê na figura. Nota-se que inicialmente os vetores característicos de $\mathbf{C}_P$ estão dispostos ao acaso, mas que rapidamente tendem a ficar paralelos aos de $\mathbf{H}^{-1}$, oscilando em torno desse "estado de equilíbrio". Pode-se supor que estas oscilações representem um efeito de população finita, como sugerido em [3].

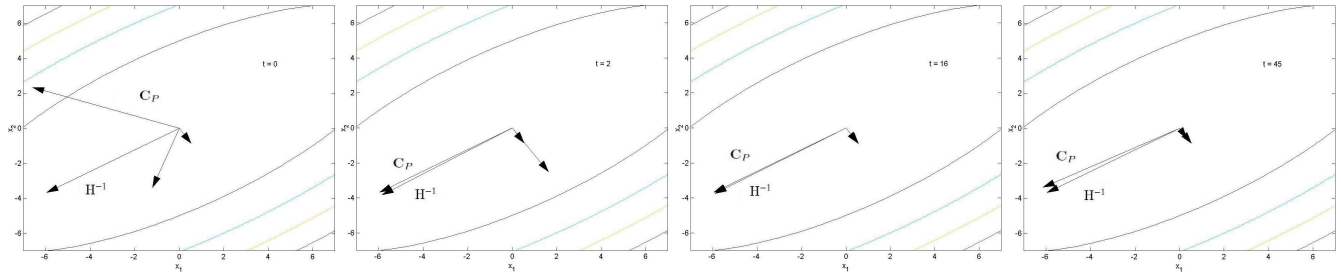

Figura 2: Observação do alinhamento dos autovetores de $\mathbf{C}_P$ e de $\mathbf{H}^{-1}$, nas iterações 0, 2, 16 e 45 de um DE básico com população $\mu = 20$, na minimização da função $f(x_1, x_2) = x_1^2 + 3x_2^2 - 2x_1x_2$. Os autovetores da inversa da Hessiana correspondem às setas fixas. Os vetores são proporcionais a seus respectivos autovalores. Note que na iteração 16 os vetores aparentam estar exatamente paralelos, mas em iterações posteriores ocorre uma oscilação em torno desse "equilíbrio", como mostra a figura da iteração 45, na qual os vetores se mostram menos alinhados que em $t = 16$. Os autovalores correspondentes a cada autovetor também oscilam em sua proporção, ao redor daquela definida pelos autovalores da inversa da Hessiana.

Para funções em maiores dimensões, é necessário criar outro procedimento para avaliar tal alinhamento, uma vez que o método gráfico se torna inviável. Uma alternativa é a utilização de uma métrica de dissimilaridade capaz de mensurar adequadamente a distância entre dois conjuntos de vetores ortogonais. A *dissimilaridade* entre um par qualquer de vetores normalizados é definida como:

$$d_{ij} \triangleq 1 - |\mathbf{u}_i \cdot \mathbf{u}_j| \tag{5}$$

onde $\mathbf{u}_i \cdot \mathbf{u}_j$ retorna o produto escalar dos vetores. Note que esta métrica de dissimilaridade atribui uma distância máxima unitária a vetores ortogonais entre si, e considera dois vetores em paralelo (ou antiparalelo) como tendo distância zero.

Para o cálculo da distância total entre os dois conjuntos de autovetores de interesse, o seguinte procedimento é utilizado: os autovetores de ambas as matrizes $\mathbf{H}^{-1}$ e $\mathbf{C}_P$ são ordenados em ordem decrescente de seus autovalores correspondentes. Seja d_{ii}^{HC} a dissimilaridade entre o i-ésimo autovetor ordenado da matriz $\mathbf{H}^{-1}$ e seu correspondente da matriz $\mathbf{C}_P$. A distância total entre os dois conjuntos de autovetores é então definida como:

$$D_{HC} \triangleq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(d_{ii}^{HC}\right)^2} \tag{6}$$

Esta métrica é aqui utilizada como uma medida do desalinhamento total entre os pontos da população do DE em uma dada iteração e as hipercurvas de nível da função objetivo sendo minimizada.

A Fig. 3 mostra o gráfico de dissimilaridade média para a função de teste $f(\mathbf{x}) = nx_1^2 + (n-1)x_2^2 + \ldots + x_n^2$ com dimensões variando de 2 a 100 [5] , e tamanho de população do DE definido como $\mu = 2n$. O comportamento observado para a dissimilaridade média entre a distribuição dos vetores de diferenças e as hipercurvas de nível desta função objetivo revela um padrão interessante, onde um aumento linear no tamanho da população é suficiente para contrabalancear o aumento de complexidade resultante da elevação da dimensão do problema. Experimentos com outras constantes de linearidade para a relação $\mu = an$ sugerem que esta propriedade se mantém, com variação apenas do valor de dissimilaridade no qual o algoritmo se estabiliza.

4. DISCUSSÃO

Um primeiro ponto a ser observado é o fato de todas as funções utilizadas como exemplos na seção anterior apresentarem curvas de nível com um dos eixos mais "alongado" que os outros (i.e., não-esféricas). A razão para isso é permitir uma melhor visualização do alinhamento da população e dos vetores-diferença (bem como seus autovalores) à forma da função[6] . Embora esta escolha possa parecer um tanto quanto arbitrária (e, de fato, testes com funções esféricas como $f(x_1, x_2) = x_1^2 + x_2^2$ não exibem o comportamento de alinhamento esperado), esta observação não deve ser tratada como um contraexemplo à hipótese básica desse

[4]Recorde que não faz diferença escolher a matriz da população ou de vetores-diferença. A escolha foi feita visando a uma facilidade maior na determinação de $\mathbf{P}$ que na determinação de $\mathbf{V}$.

[5]O fato de $f(\mathbf{x})$ possuir semieixos principais alinhados com os eixos coordenados não afeta a generalidade dos testes, uma vez que o DE é insensível à rotação [4].

[6]Por isso o uso da função não tão difundida na seção anterior, em que os coeficientes foram convenientemente escolhidos tanto para facilitar a construção da Hessiana como para permitir eixos com autovalores diferenciados.

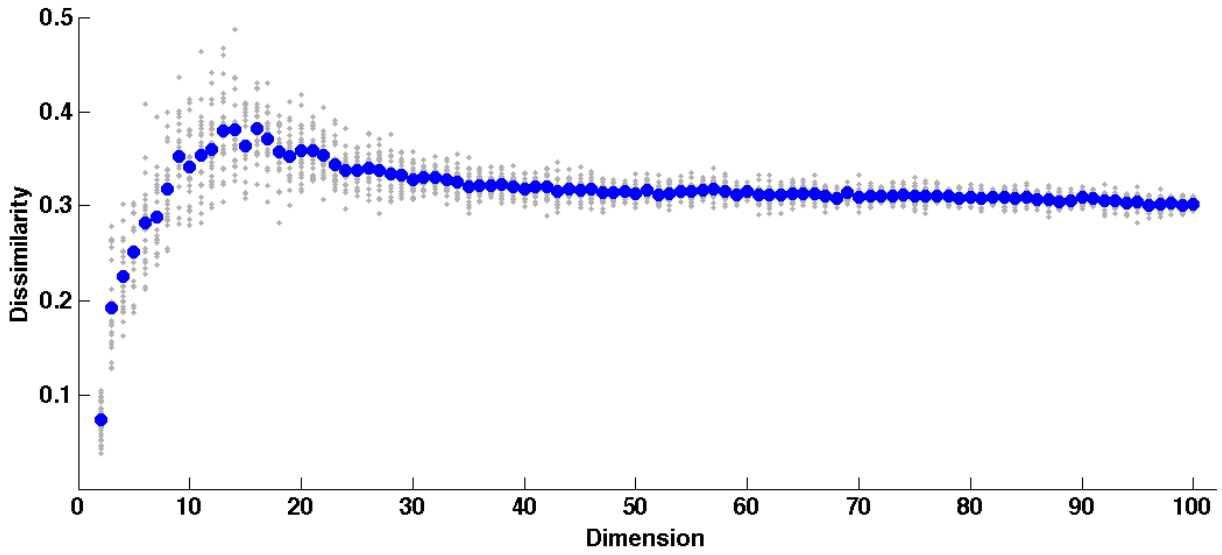

Figura 3: Dissimilaridade para o DE na otimização da função $f(\mathbf{x}) = \sum_{i=1}^{n}(n+1-i)x_i^2$. Para cada dimensão foram executadas 25 replicações do experimento, com os pontos em cinza simbolizando a dissimilaridade média em cada replicação e o ponto central (em azul) a média geral para a dimensão. O tamanho de população utilizado neste caso foi $\mu = 2n$, mas comportamentos similares foram observados para outros valores de μ variando linearmente com a dimensão do problema.

trabalho. Relembrando a conjectura inicialmente levantada, *o DE funciona por meio da adaptação do mecanismo de variação à forma da função objetivo*, temos que para o caso de funções com simetria esférica em suas hipercurvas de nível qualquer direção escolhida será satisfatória. Essa situação pode ser considerada, na verdade, um caso trivial. Contudo, é necessário perceber que a métrica de distância adotada, a dissimilaridade, serve para mensurar a distância entre os autovetores da população e os da inversa da Hessiana da função, e não a distância entre a forma da população (ou dos vetores-diferença) e a forma da função. Para essa situação, tal métrica pode não ser adequada, sendo necessário o desenvolvimento de outra ferramenta para uma eventual aplicação destes conceitos em, por exemplo, variações do mecanismos de autoadaptação do DE.

Um segundo ponto de interesse emerge a partir do comportamento observado e o mecanismo de mutação do DE. Relembrando a definição deste operador, explicitada na equação (1), vimos que é possível nomear o fator $F\mathbf{v}$ apropriadamente de *passo de mutação*. Tendo isso em vista, e efetuando um empréstimo da terminologia do algoritmo *estratégias evolutivas* com adaptação de matrizes de covariância, abreviado por CMA-ES (*covariance matrix adaptation - evolution strategies*) [6], é tentador rescrever (1) como

$$\mathbf{u}_i = \mathbf{x}_{i1} + \sigma\mathcal{N}(\mathbf{0}, \mathbf{C}_V) \tag{7}$$

onde $\mathcal{N}$ indica um termo retirado de um distribuição normal, σ um escalar (como o fator multiplicativo F) usado para representar o tamanho do passo de mutação, e $\mathbf{C}_V$ é a matriz de covariâncias dos vetores-diferença[7] Verificações experimentais preliminares utilizando análise de componentes principais e teste de Lilliefors para avaliação de normalidade [7] confirmam que a distribuição dos vetores de diferenças é realmente modelável como uma variável Gaussiana multidimensional [8] . Esta alteração na notação do algoritmo, juntamente com um mecanismo autorrealimentado de adaptação das matrizes de covariância envolvidas na geração de novos pontos ao longo do processo evolutivo, sugerem uma conexão próxima com os mecanismos evolutivos do CMA-ES. Se confirmada, esta semelhança pode possibilitar a utilização do grande corpo de conhecimento analítico já desenvolvido para este último algoritmo para a análise dos mecanismos de funcionamento do DE.

Finalmente, este trabalho limitou-se a funções quadráticas, tirando proveito de suas características (e.g., existência de matriz Hessiana constante) na execução dos testes e análise do comportamento do DE. A análise do comportamento do algoritmo de evolução diferencial em problemas com funções objetivo de maior complexidade é um trabalho de continuidade cujo desenvolvimento certamente será facilitado pelos resultados obtidos no presente artigo.

5. CONCLUSÕES

O presente trabalho apresentou considerações teóricas sobre os mecanismos de funcionamento e adaptação do algoritmo de evolução diferencial. A relação constante entre os indivíduos da população em uma dada iteração e a distribuição de vetores-diferença, responsáveis pela direção de exploração do espaço de busca, foi demonstrada e caracterizada. O comportamento de ajuste desta distribuição às curvas de nível da função objetivo também foi investigada para funções quadráticas de matriz Hessiana constante, indicando uma forte tendência à ocorrência deste alinhamento para esta classe de funções. Estas observações formam

[7] Conforme visto, é possível permutá-la com a matriz de covariâncias da população, $\mathbf{C}_P$, com o fator de proporcionalidade κ incorporado em σ. O uso de $\mathbf{C}_P$ tende a ser mais conveniente para análise.

[8] De fato, pode-se argumentar que o conjunto de vetores-diferença provavelmente seria melhor modelado por uma distribuição T multidimensional. Entretanto, o número de diferenças possíveis é geralmente alto o suficiente para permitir a utilização da aproximação normal sem prejuízos à análise.

a base para a elaboração de experimentos de sequência para a caracterização do comportamento do DE em funções mais gerais e de maior complexidade.

AGRADECIMENTOS

Este trabalho foi financiado pelo Conselho Nacional de Pesquisa (CNPq, projetos 472446/2010-0 e 306910/2006-3); e pela Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG, Pronex APQ 01075/09).

APÊNDICE A: PROVA DO TEOREMA 3.1

Demonstração. Sejam $\mathbf{P}$ e $\mathbf{V}$ definidas como:

$$\mathbf{P} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\mu,1} & x_{\mu,2} & \cdots & x_{\mu,n} \end{bmatrix} \qquad \mathbf{V} = \begin{bmatrix} (x_{1,1} - x_{2,1}) & (x_{1,2} - x_{2,2}) & \cdots & (x_{1,n} - x_{2,n}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{1,1} - x_{\mu,1}) & (x_{1,2} - x_{\mu,2}) & \cdots & (x_{1,n} - x_{\mu,n}) \\ (x_{2,1} - x_{1,1}) & (x_{2,2} - x_{1,2}) & \cdots & (x_{2,n} - x_{1,n}) \\ \vdots & \vdots & \vdots & \vdots \\ (x_{\mu,1} - x_{\mu-1,1}) & (x_{\mu,2} - x_{\mu-1,2}) & \cdots & (x_{\mu,n} - x_{\mu-1,n}) \end{bmatrix}$$

i.e., com a matriz $\mathbf{V}$ formada pela subtração dois a dois dos μ indivíduos. Consequentemente, o número de linhas nesta matriz é dado por:

$$\mu' = 2\binom{\mu}{2} = 2\frac{\mu!}{2!(\mu-2)!} = \mu(\mu-1)$$

com o fator 2 aparecendo devido aos termos simétricos. As matrizes de covariância podem ser expressas como:

$$\mathbf{C}_P = \begin{bmatrix} s_P(1,1) & \cdots & s_P(1,n) \\ \vdots & \ddots & \vdots \\ s_P(n,1) & \cdots & s_P(n,n) \end{bmatrix} \qquad \mathbf{C}_V = \begin{bmatrix} s_V(1,1) & \cdots & s_V(1,n) \\ \vdots & \ddots & \vdots \\ s_V(n,1) & \cdots & s_V(n,n) \end{bmatrix}$$

em que $s(a,b)$ representa a covariância amostral entre a a-ésima e a b-ésima variáveis das respectivas matrizes, para quaisquer $a, b \in \{1, \ldots, n\}$ [9] . A expressão $\mathbf{C}_V = \kappa \mathbf{C}_P$ pode ser deduzida mostrando-se que $s_V(a,b) = \kappa s_V(a,b)$, pois isso prova a relação para todos os elementos das matrizes. Para tal, considere que x e y representam termos das colunas a e b, respectivamente, de $\mathbf{P}$; e u e v representam termos correspondentes em $\mathbf{V}$. As covariâncias amostrais podem ser calculadas a partir de:

$$\begin{aligned} s_P(a,b) &= \frac{1}{\mu-1}\sum_{i=1}^{\mu}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{\mu-1}\left(\sum_{i=1}^{\mu} x_i y_i - \mu\bar{x}\bar{y}\right) \\ s_V(a,b) &= \frac{1}{\mu'-1}\sum_{i=1}^{\mu'}(u_i - \bar{u})(v_i - \bar{v}) = \frac{1}{\mu(\mu-1)-1}\sum_{i=1}^{\mu(\mu-1)} u_i v_i \end{aligned} \tag{8}$$

onde as médias amostrais $\bar{u}$ e $\bar{v}$ são iguais a zero por definição, uma vez que cada vetor-diferença possui seu vetor simétrico. Expandindo o somatório do termo $s_V(a,b)$:

$$\begin{aligned} \sum_{i=1}^{\mu(\mu-1)} u_i v_i &= (x_1 - x_2)(y_1 - y_2) + \cdots + (x_1 - x_\mu)(y_i - y_\mu) + (x_2 - x_1)(y_2 - y_1) + \cdots + (x_\mu - x_{\mu-1})(y_\mu - y_{\mu-1}) \\ &= \sum_{i=1}^{\mu}\sum_{\substack{j=1 \\ j\neq i}}^{\mu}(x_i - x_j)(y_i - y_j) \end{aligned}$$

O termo $j \neq i$ no somatório tem como função indicar que, na construção dos vetores-diferença, não se realiza a diferença de um vetor com ele próprio. Este termo, entretanto, pode ser facilmente eliminado considerando-se que a diferença entre um vetor e ele mesmo é nula, nada modificando na soma final. Assim, é possível escrever:

$$s_V(a,b) = \frac{1}{\mu(\mu-1)-1}\sum_{i=1}^{\mu}\sum_{j=1}^{\mu}(x_i - x_j)(y_i - y_j)$$

[9] Evidentemente, $s(i,i) = s_i^2$ representa a variância marginal da i-ésima coordenada.

Expandindo os produtos acima, temos:

$$\begin{aligned} s_V(a,b) &= \frac{1}{\mu(\mu-1)-1}\sum_{i=1}^{\mu}\sum_{j=1}^{\mu}(x_iy_i - x_jy_i - x_iy_j + x_jy_j) \\ &= \frac{1}{\mu(\mu-1)-1}\sum_{i=1}^{\mu}\left(\underbrace{\sum_{j=1}^{\mu}x_iy_i}_{\mu x_iy_i} - \underbrace{y_i\sum_{j=1}^{\mu}x_j}_{\mu\bar{x}y_i} - \underbrace{x_i\sum_{j=1}^{\mu}y_j}_{\mu\bar{y}x_i} + \sum_{j=1}^{\mu}x_jy_j\right) \\ &= \frac{1}{\mu(\mu-1)-1}\left(\mu\sum_{i=1}^{\mu}x_iy_i - \underbrace{\mu\bar{x}\sum_{i=1}^{\mu}y_i}_{\mu^2\bar{x}\bar{y}} - \underbrace{\mu\bar{y}\sum_{i=1}^{\mu}x_i}_{\mu^2\bar{x}\bar{y}} + \underbrace{\sum_{i=1}^{\mu}\sum_{j=1}^{\mu}x_jy_j}_{\mu\sum_{j=1}^{\mu}x_jy_j}\right) \end{aligned}$$

Da equação (8), temos que $\sum_{k=1}^{\mu}x_ky_k = (\mu-1)s_{a,b}^P + \mu\bar{x}\bar{y}$. Somando-se os termos semelhantes, temos:

$$\begin{aligned} s_V(a,b) &= \frac{1}{\mu(\mu-1)-1}\left[\mu(\mu-1)s_{a,b}^P + \mu^2\bar{x}\bar{y} - 2\mu^2\bar{x}\bar{y} + \mu(\mu-1)s_{a,b}^P + \mu^2\bar{x}\bar{y}\right] \\ &= \frac{2\mu(\mu-1)}{\mu(\mu-1)-1}s_P(a,b) \end{aligned}$$

Portanto, para qualquer elemento das matrizes de covariância:

$$\kappa \triangleq \frac{s_V(a,b)}{s_P(a,b)} = \frac{2\mu(\mu-1)}{\mu(\mu-1)-1} \tag{9}$$

□

REFERÊNCIAS

[1] R. M. Storn and K. V. Price. "Differential Evolution - a Simple and Efficient Heuristic for Global Optimization over Continuous Spaces". *Journal of Global Optimization*, vol. 11, pp. 341–359, 1995.

[2] K. V. Price, R. M. Storn and J. A. Lampinen. *Differential Evolution: A Pratical Approach to Global Optimization*. Springer, 2005.

[3] L. S. Batista, F. Campelo, F. G. Guimarães and J. A. Ramírez. "A New Self-Adaptive Approach for Evolutionary Multiobjective Optimization". In *Proceedings of the 2010 Congress on Evolutionary Computation*, pp. 1–8, Barcelona, Spain, 2010.

[4] F. G. Guimarães. *Algoritmos de Evolução Diferencial para Otimização e Aprendizado de Máquina*, volume 1, pp. 1–17. Sociedade Brasileira de Redes Neurais, 2009.

[5] A. P. Engelbrecht. *Computational Intelligence - An Introduction*. John Wiley & Sons, 2007.

[6] N. Hansen and A. Ostermeier. "Completely derandomized self-adaptation in evolution strategies". *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.

[7] D. J. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman & Hall/CRC, fourth edition, 2007.