Descenso de Partículas mediante el Estimador del Gradiente Esperado

Segovia-Domínguez Ignacio - jsegovia@itesg.edu.mx - Instituto Tecnologico Superior de Guanajuato

Munguía-Gutiérrez Sandra - smunguia@itesg.edu.mx - Instituto Tecnologico Superior de Guanajuato

Resumen

Este documento introduce un nuevo método de optimización que mezcla los conceptos de optimización evolutiva y descenso de gradiente. Así, esta investigación se enfoca en problemas en el dominio continuo mediante múltiples partículas. Es importante resaltar que la actualización de la población considera una estimación de gradiente poco utilizada en la literatura. Adicionalmente se presentan resultados en diversos problemas para mostrar la capacidad del método propuesto.

Abstract

This paper introduces a new optimization method in which two main topics are mixed: 1) evolutionary optimization and, 2) gradient descent. Thus, this research focuses in solving continuous domain problems by updating several particles at each iteration. It is worth pointing out the population is updated by considering an unusual gradient estimator from literature. Aditionally, the proposed method is tested on several benchmark problems.

Palabras Clave: Optimización, Algoritmo Evolutivo, Estimación de Gradiente.

I. Introducción

Hallar el optimo global de una función, cuya expresión matemática es desconocida, es un problema retador y ampliamente discutido en la literatura (Larranaga, 2002) (Schwefel, 1995) (Deb, 2001). Debido a la diversidad de ideas propuestas para resolver este problema, existe una gran variedad de paradigmas dirigidos a alcanzar el optimo global de una función objetivo $\mathcal{F}\ x$; por ejemplo:

optimización estocástica, algoritmos evolutivos, ajuste de modelos paramétricos, superficies de respuesta, entre otros. Aunque cada uno de estas propuestas tiene cierto éxito en subconjuntos de problemas, mas y mejores métodos pueden ser desarrollados al extraer mayor información de $\mathcal{F}(x)$. Particularmente las derivadas de distintos ordenes proveen información valiosa acerca del comportamiento de la función a resolver (Nocedal, 2006). En este sentido, el vector gradiente $\nabla\mathcal{F}(x)$ brinda información local acerca de la dirección que debiera seguir una partícula para encontrar una posición con mejor valor de la función objetivo (Snyman, 2005).

Los métodos basados en la información del gradiente (descenso de gradiente, BFGS, gradiente conjugado, entre otros) han demostrado ser eficientes en la optimización de funciones (Nocedal, 2006). Siendo que los algoritmos basados en estos métodos tienen baja complejidad computacional, es común que estas técnicas consigan alcanzar un optimo de la función en pocas iteraciones y poco tiempo de computo. La información que el campo gradiente aporta es, indiscutiblemente, significativa para conseguir que una partícula avance a una mejor posición. Sin embargo, el paradigma basado en el gradiente es propenso a estancarse en óptimos locales (Salomon, 1998).

Algunos algoritmos evolutivos, inspirados en la optimización clásica, proponen utilizar vectores dirigidos hacia donde se espera que los individuos alcancen en mejor valor en $\mathcal{F}\ x$. La mayoría de estas propuestas basan su funcionamiento en vectores de desplazamiento. Se ha observado que los vectores construidos con este propósito favorecen la localización de óptimos de la función, a pesar de no contar con la información de gradiente (Clerc, 2010) (Hansen, 2006) (Storn, 1997).

En general, se ha observado que los algoritmos basados en múltiples partículas (población) son menos susceptibles de estancarse en óptimos locales (Salomon,

1998) (Schwefel, 1995). En este contexto, una idea interesante consiste en fusionar ambos conceptos: optimización basada en poblaciones y la información del gradiente de la función objetivo. En este sentido, las siguientes interrogantes cobran especial importancia: ¿Cómo podemos utilizar la información del gradiente en los algoritmos basados en poblaciones? y ¿Qué ventajas podríamos obtener con propuestas de esta naturaleza?. En la literatura se pueden encontrar algunas propuestas de solución a estas interrogantes (Bosman, 2006) (Chen, 2009).

La principal dificultad para crear algoritmos, con las ideas discutidas anteriormente, radica en que los algoritmos basados en poblaciones (algoritmos evolutivos) se enfocan en resolver problemas de "caja negra". En otras palabras, resuelven problemas asumiendo que la expresión matemática de la función es desconocida. Considerando el anterior supuesto, no es posible emplear la expresión matemática de $\nabla\mathcal{F}(x)$ en un algoritmo evolutivo. Sin embargo, este obstáculo puede evadirse al calcular una aproximación numérica del gradiente.

El presente trabajo de investigación se enfoca en construir un algoritmo evolutivo que emplee una aproximación de $\nabla\mathcal{F}(x)$ para desplazar la población a través del espacio de búsqueda. La aproximación numérica empleada se basa en la estimación de gradiente esperada; EGE, por sus siglas en ingles. Adicionalmente se muestran algunos resultados en diversos problemas ampliamente utilizados en la literatura.

El resto de este documento justifica y especifica un algoritmo de optimización para funciones, en dominio continuo, cuya expresión matemática es desconocida. La sección II presenta la estimación de gradiente utilizada, así como una comparativa contra otra aproximación comúnmente utilizada en la literatura. Además, se detallan los pasos necesarios para programar el algoritmo propuesto. En la sección III se muestran algunos resultados obtenidos al aplicar el algoritmo de optimización. Finalmente, la sección IV argumenta algunas conclusiones y posibles áreas de oportunidad para trabajo futuro.

II. Metodología

Para desarrollar un algoritmo de optimización utilizando la idea descrita en la sección I, es necesario establecer el método de aproximación de gradiente a utilizar. Además, se requiere establecer la metodología de actualización de partículas a emplear. Ambos aspectos son presentados y estudiados en las siguientes secciones.

A. Estimación del gradiente de una función

Los algoritmos evolutivos asumen que la expresiones matemáticas de $\mathcal{F}\ x$ y $\nabla\mathcal{F}(x)$ son desconocidas. Para poder tomar ventaja de la información de gradiente es necesario considerar una aproximación numérica. Esta sección presenta dos estimadores encontrados en la literatura. Sin embargo, solo uno de ellos será empleado en la propuesta descrita e la siguiente sección.

El vector gradiente $\nabla\mathcal{F}(x)$ indica la dirección en la cual $\mathcal{F}(x)$ varia mas rápidamente en el punto x. A fin de construir una aproximación del vector gradiente es necesario considerar observaciones de $\mathcal{F}(x)$ en distintas posiciones del espacio. Considérese una estimación de gradiente $\nabla\mathcal{F}(x^{(i)})$ para la partícula $x^{(i)}$. Para calcular $\nabla\mathcal{F}(x^{(i)})$ al menos se requiere el valor de la función objetivo $\mathcal{F}(x^{(i)})$ e información sobre un conjunto de otras partículas en el espacio de búsqueda; es decir $x^{(1)}, \dots, x^{(k)}, \dots, x^{(N)}$ y $\mathcal{F}(x^{(1)}), \dots, \mathcal{F}(x^{\ k}), \dots, \mathcal{F}(x^{\ N})$, donde $x^{(i)} \neq x^{(k)}$ y N representa la cantidad de partículas a considerar. En este contexto, uno de los estimadores mas antiguos aproxima el gradiente mediante un ajuste por hiper-plano. Así, una estimación puede obtenerse mediante el método de mínimos cuadrados ordinarios (Hazen, 2009). Aunque esta técnica crea una aproximación de gradiente de manera intuitiva, en algunos contextos podría ser inadecuada (por ejemplo, cuando no existe la cantidad de observaciones suficientes para crear un sistema de ecuaciones sobre determinado). Es por ello que en este trabajo se utiliza una propuesta diferente fundamentada en dos conceptos: la derivada direccional y la esperanza estadística.

Definición 1. Sean $x^{(1)}, \ldots, x^{(k)}, \ldots, x^{(N)}$ un conjunto de partículas vecinas de $x^{(i)}$, todas localizadas en el dominio de búsqueda para $\mathcal{F}(x)$. Entonces el Estimador del Gradiente Esperado (EGE) para $x^{(i)}$ se calcula mediante

$$\nabla\mathcal{F}\left(x^{(i)}\right) = \frac{1}{N}\sum_{k=1}^{N} \frac{\mathcal{F}\left(x^{(k)}\right) - \mathcal{F}\left(x^{(i)}\right)}{\left\| x^{(k)} - x^{(i)} \right\|^{2}} \left(x^{(k)} - x^{(i)}\right)$$

donde N es el tamaño del vecindario y $\mathcal{F}(x)$ representa la función objetivo (Segovia-Dominguez, 2014).

La definición 1 muestra un estimador de gradiente (EGE) con dos notorias ventajas sobre la común de las aproximaciones: 1) basta con al menos una partícula vecina para obtener una aproximación del gradiente y 2) es computacionalmente fácil de implementar. Adicionalmente, el tiempo de computo que se requiere para estimar el gradiente es bastante bajo, lo cual hace aun mas atractiva esta propuesta (Segovia-Dominguez, 2014). La siguiente sección introduce un algoritmo evolutivo que utiliza una aproximación de gradiente mediante la técnica de EGE.

B. Descenso de Partículas mediante el Gradiente Esperado

{

, ,

Mejor valor en

Particula de con valor

Calcular utilizando , y

, Mejor valor en

Particula de con mejor valor

{

reemplaza al peor individuo en

reemplaza al peor valor en

}

{ }

}

Retornar como la solucion del problema

Figura 1. Pseudocódigo del metodo DPGE

Esta sección se enfoca en la construcción de un algoritmo de optimización mediante la generación iterativa de múltiples partículas en el espacio de búsqueda. Así, un algoritmo de esta naturaleza persigue encontrar cada vez mejores soluciones conforme transcurran las iteraciones.

La Figura 1 muestra el pseudocódigo del método propuesto: Descenso de Partículas mediante el Gradiente Esperado (DPGE). La idea central de esta propuesta consiste en simular nuevos individuos, $\mathcal{P}_t^S$, a partir de una densidad normal multivariada y reemplazar el mejor individuo encontrado, x_{best}, cuando surja una mejor partícula, x_{best}^S, de entre las generadas. A fin de brindar mayor estabilidad al algoritmo, la estimación del gradiente, EGE, se realiza mediante un conjunto resultado de la concatenación de dos poblaciones diferentes: $\mathcal{P}_t^H$ y $\mathcal{P}_t^S$. El conjunto $\mathcal{P}_t^H$ contiene las ultimas N^H partículas que históricamente han sido los mejores individuos; mientras que el conjunto $\mathcal{P}_t^S$ contiene nuevas partículas generadas en el tiempo t. Los parámetros de la normal multivariada a utilizar se adaptan mediante los parámetros reales λ y σ. La idea central del método de adaptación consiste en incrementar la exploración cuando se actualiza el mejor individuo y favorecer la convergencia del método cuando no se localiza un mejor individuo que el existente en el tiempo t.

La estimación de gradiente mediante la formulación de la técnica EGE permite crear un algoritmo computacionalmente eficiente. Adicionalmente, nótese que el algoritmo DPGE adquiere robustez ante óptimos locales al generar nuevos individuos mediante una función de densidad (Larranaga, 2002) (Pelikan, 2007) (Schwefel, 1995). La siguiente sección esta dedicada analizar algunos resultados obtenidos con base en las ideas discutidas anteriormente.

III. Resultados

Esta sección presenta dos resultados relacionadas con el algoritmo propuesto (DPGE). El primero de estos muestra la diferencia angular entre la estimación de gradiente empleada y otro estimador clásico en la literatura. Acto seguido se discute el comportamiento del algoritmo DPGE en diferentes problemas.

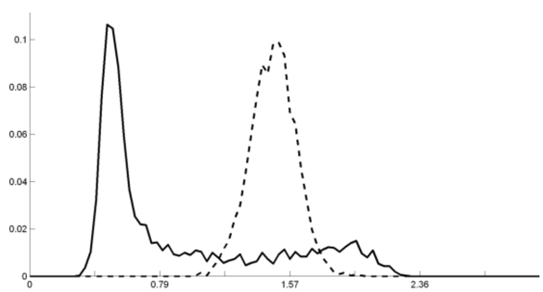


Figura 2. Comparativa angular en 30 dimensiones entre el verdadero gradiente y EGE (linea solida)/OLS (linea punteada)

A. Comparativa angular entre estimaciones de vectores gradiente

Considérese un conjunto de 300 partículas generadas en el dominio $(-600,300)$ mediante una secuencia pseudoaleatoria Halton. Sea $d = 30$ la dimensión del problema y $N = d + 1$ el tamaño del vecindario, cuyos vecinos son aquellas partículas mas cercanas al individuo x^{i}, la Figura 2 muestra un histograma del ángulo entre un estimador de gradiente y el verdadero vector gradiente para la función objetivo $\mathcal{F}(x) = -\sum_{i=1}^{d} x_i^2$.

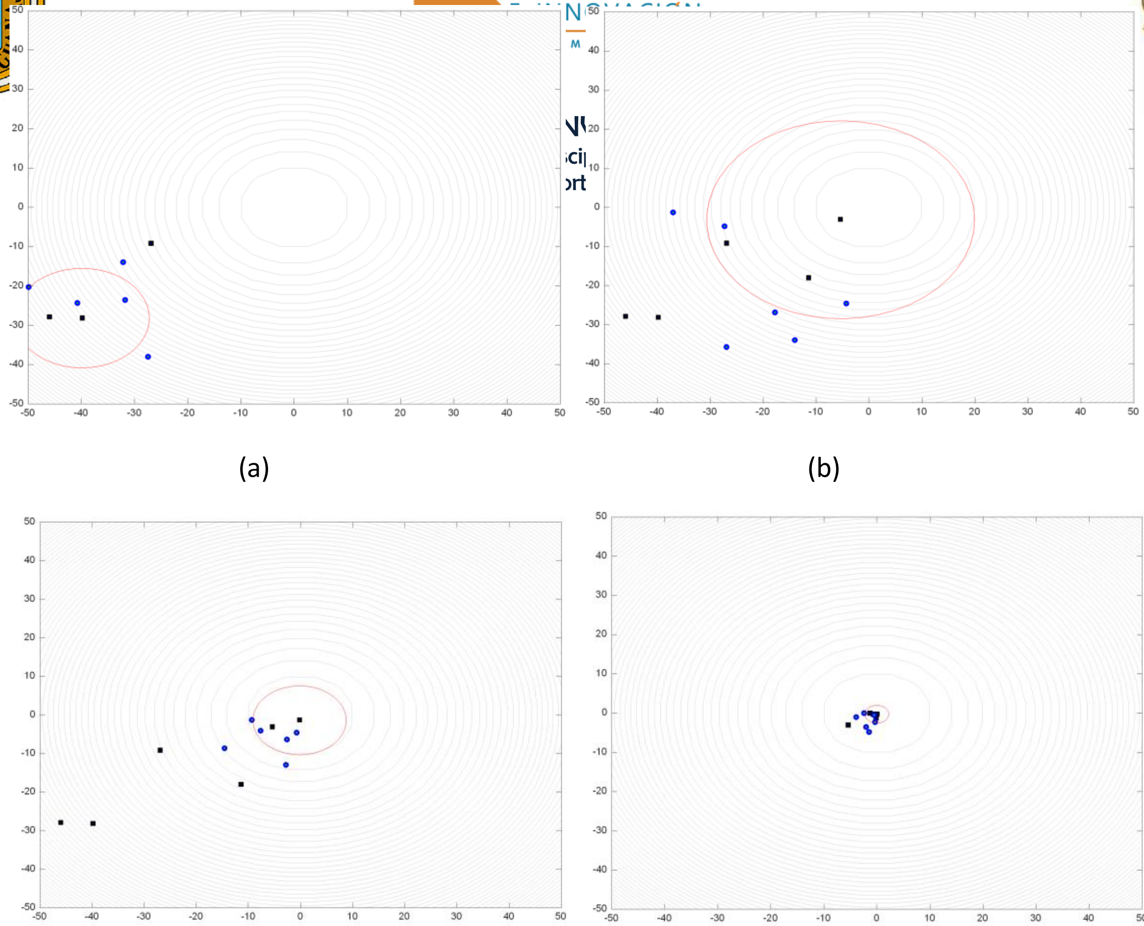
Así, un valor cercano a cero implica una mejor predicción de la dirección. Los estimadores comparados en la Figura 2 son la técnica EGE y un ajuste de hiper-plano mediante mínimos cuadrados ordinarios (OLS, por sus siglas en ingles). Así, de este grafico podemos concluir que la técnica EGE se aproxima mejor, en valor angular, la dirección a la que apunta el verdadero gradiente.

Este experimento demuestra el potencial que tiene la aproximación de gradiente mediante el EGE. Adicionalmente, estos resultados sugieren que la actualización de las partículas empleada por el algoritmo DPGE es adecuada para el contexto de optimización.

B. Comportamiento del algoritmo en distintos problemas

Esta sección esta dedicada a estudiar el comportamiento del algoritmo propuesto (DPGE) en diferentes funciones objetivo ampliamente utilizadas en la literatura. En el presente trabajo de investigación se emplean las funciones: 1) Sphere, 2) Zakharov, 3) Exponential, 4) Rosenbrock y 5) Ackley; véase (Jamil, 2013). Es pertinente hacer notar que estas funciones son problemas de minimización, por lo que es necesario utilizar el negativo de la estimación ofrecida por el método EGE.

La Figura 3 muestra el comportamiento del algoritmo propuesto en la función objetivo Sphere. En cada imagen pueden observarse los conjuntos de individuos $\mathcal{P}_t^H$ y $\mathcal{P}_t^S$; adicionalmente se muestra una curva de nivel (circunferencia) de la densidad normal multivariada utilizada en cada iteración. Nótese que el algoritmo DPGE es capaz de adaptar la localización y apertura de la función de densidad. La Figura 4 muestra algunos estadísticos sobre las evaluaciones de función requeridas por el algoritmo, considerando 20 corridas del algoritmo en cada dimensión y cada función objetivo. Estos gráficos muestran que el algoritmo es capaz de alcanzar el optimo global de la función utilizando una cantidad razonable de evaluaciones de función.

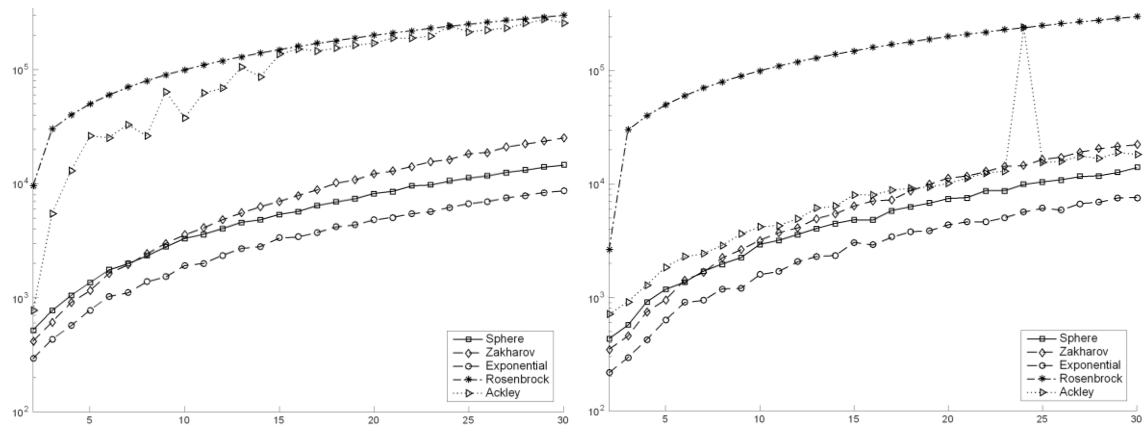


(a) (b) (c) (d)

Figura 3. Comportamiento del algoritmo DPGE en la funcion Sphere. Cuadrados negros: . Circulos azules: . Circunferencia roja: . (a) . (b) . (c) . (d)

IV. Conclusiones

La creación de algoritmos de optimización que mezclen ideas de optimización clásica y optimización evolutiva es una área de investigación prometedora. En este sentido, el algoritmo DPGE une las ideas de descenso de gradiente y simulación



(a) (b)

Figura 4. Comportamiento del algoritmo DPGE en 5 funciones objetivo desde 2 hasta 30 dimensiones. (a) Media del numero de evaluaciones de funcion requeridas. (b) Minima cantidad del numero de evaluaciones de funcion requeridas.

aleatoria de individuos. Aunque la actual propuesta no considera dependencia entre las variables, esta idea puede desarrollarse como trabajo futuro. Así, la propuesta presentada en el presente trabajo muestra que es posible mezclar ambos conceptos para construir propuestas eficientes y con baja complejidad computacional.

V. Bibliografía

Bosman, P. A. (2006). Combining gradient techniques for numerical multi-objective evolutionary optimization. *Proceedings of the 8th annual conference on Genetic and evolutionary computation* (pp. 627-634). ACM.

Chen, X. L. (2009). Combining evolution strategy and gradient descent method for discriminative learning of bayesian classifiers. *11th Annual conference on Genetic and evolutionary computation* (pp. 507-514). ACM.

Clerc, M. (2010). *Particle swarm optimization* (Vol. 93). John Wiley & Sons.

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms* (Vol. 16). John Wiley & Sons.

Hazen, M. &. (2009). Gradient estimation in global optimization algorithms. *Evolutionary Computation, 2009. CEC'09. IEEE Congress on* (pp. 1841-1848). IEEE.

Hansen, N. (2006). The CMA evolution strategy: a comparing review. In *Towards a new evolutionary computation* (pp. 75-102). Springer Berlin Heidelberg.

Jamil, M. &. (2013). A literature survey of benchmark functions for global optimisation problems. *International Journal of Mathematical Modelling and Numerical Optimisation* , 150-194.

Larranaga, P. &. (2002). *Estimation of distribution algorithms: A new tool for evolutionary computation* (Vol. 2). Springer Science & Business Media.

Nocedal, J. &. (2006). *Numerical optimization.* Springer Science & Business Media.

Pelikan, M. S.-P. (2007). *Scalable optimization via probabilistic modeling: From algorithms to applications* (Vol. 33). Springer.

Salomon, R. (1998). Evolutionary algorithms and gradient search: similarities and differences. *Evolutionary Computation, IEEE Transactions on* (pp. 45-55). IEEE.

Schwefel, H. P. (1995). *Contemporary evolution strategies.* Springer Berlin Heidelberg.

Segovia-Dominguez, I. H.-A. (2014). A New EDA by a Gradient-Driven Density. *Parallel Problem Solving from Nature–PPSN XIII* (pp. 352-361). Springer International Publishing.

Snyman, J. (2005). *Practical mathematical optimization: an introduction to basic optimization theory and classical and new gradient-based algorithms* (Vol. 97). Springer Science & Business Media.

Storn, R. &. (1997). Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization* , 341-359.