27 Congreso Nacional de Estadística e Investigación Operativa
Lleida, 8–11 de abril de 2003

UN ALGORITMO DE OPTIMIZACIÓN ESTOCÁSTICA CON INCORPORACIÓN DE INFORMACIÓN A PRIORI

Gilberto Pérez Lechuga, Jorge A. Rojas Ramírez, Juan Carlos Seck Tuoh Mora

Centro de Investigación Avanzada en Ingeniería Industrial
Universidad Autónoma del Estado de Hidalgo, 42184 Pachuca Hidalgo, México

RESUMEN

En este documento se desarrolla una algoritmo de búsqueda aleatoria con incorporación de información a priori para optimizar algunos modelos de programación matemática no lineal con estructura estocástica. El algoritmo propuesto incorpora la información en forma de esperanzas matemáticas y analiza diversas distribuciones de probabilidad. La propuesta se acompaña de la correspondiente prueba de convergencia y se ilustran algunas aplicaciones.

Palabras y frases clave: Optimización estocástica, Simulación Montecarlo, Teoría de la información.

1 Introducción

La teoría de la inferencia estadística y las medidas de información juegan un papel importante en el desarrollo de algoritmos de búsqueda aleatoria para la optimización de modelos estocásticos aunque constituyen un campo muy poco explorado.
La optimización por búsqueda aleatoria fue originada con los trabajos pioneros de Robins y Monro (1951), quienes sugirieron una técnica iterativa para encontrar las raíces de una función de regresión en $\mathbb{R}^1$ medida con ruido. Posteriormente, Kiefer y Wolfowitz (1952) sugieren un método para encontrar el valor óptimo de una función no restringida en $\mathbb{R}^1$. Las técnicas Robins-Monroe, y Kiefer-Wolfowitz fueron generalizadas por Dvoretzky (1956).
En los modelos estocásticos, la minimización tiene sentido en términos de una esperanza matemática, es decir, si $x \in \mathbb{R}^n$ es el vector de variables de decisión, W es el vector de variables aleatorias o ruido que acompañan a la función, entonces los modelos pueden caracterizarse como sigue:

$$\mathbf{P(1)} = \begin{cases} \text{Minimizar}_x \; g_0(x) \;=\; \min_x \int \psi_0(x,w) f_W(w,x) dw \;=\; \min_x \; \mathbf{E}[\psi_0(x,W)] \\ \\ \text{Sujeto a} \\ \\ g_j(x) \;=\; \int \psi_j(x,w) f_W(w,x) dw \;=\; \mathbf{E}\left[\psi_j(x,W)\right] \leq 0, \quad j = 1,2,\ldots M, \end{cases}$$

donde $x \in \mathcal{D}$ y w son vectores n y m dimensionales respectivamente, y $f_W(w, x)$ es una función de densidad de probabilidad (f.d.p.) de una variable aleatoria W que depende de un vector de parámetros x (en el caso particular, la densidad de W se puede escribir como una función que no depende de x en la forma $F_W(w)$).
Note que si la f.d.p $f_W(w, x)$ es conocida, y las esperanzas matemáticas $\mathbf{E}[\psi_j(x, W)]$, $j = 0, 1, \ldots M$, pueden encontrarse analíticamente para todas las x, entonces $\mathbf{P}(\mathbf{1})$ puede reducirse a un modelo no lineal el cual puede ser resuelto por los métodos convencionales de la programación matemática. Sin embargo, si $\mathbf{E}[\psi_j(x, W)]$, $j = 0, 1, \ldots M$, no puede evaluarse explícitamente, (esto es, solamente pueden estimarse algunas de sus realizaciones a través de la simulación Montecarlo) o bien, si la f.d.p, $f_W(w, x)$ es desconocida entonces no hay ninguna forma analítica de resolver a $\mathbf{P}(\mathbf{1})$.
En la práctica, la mayoría de los métodos de búsqueda estocástica, funcionan a partir de una muestra de puntos obtenidos aleatoriamente sobre un conjunto $\mathcal{D} \subset \mathbb{R}^n$. Bajo condiciones ideales de la función objetivo f, se puede hacer que la probabilidad de encontrar el óptimo global sea mayor conforme aumenta el número de iteraciones. En general, es muy deseable que los métodos de búsqueda aleatoria aporten cierta información sobre la rapidez y tipo de convergencia así como la calidad de los resultados esperados.
Estos métodos son generalmente construidos a partir de una sucesión de distribuciones de probabilidad que convergen para una amplia clase de funciones a una distribución límite centrada en el mínimo global. Los algoritmos de búsqueda aleatoria empiezan en algún punto $x_0 \in \mathcal{D}$ y generan una sucesión $x_1, x_2 \ldots \in \mathcal{D}$. Por su naturaleza, se les puede considerar como algoritmos de descenso en el sentido de que las sucesiones $g(x_0), g(x_1), \ldots$ son monótonas decrecientes. Asimismo x_{k+1} se genera a partir de x_k por técnicas de búsqueda aleatoria usando la idea de los métodos de descenso.
La forma de operar en un algoritmo de búsqueda es determinar una dirección de descenso en forma aleatoria usando una distribución de prueba alrededor de x_i, obtener un tamaño de paso apropiado de acuerdo a las mejoras sucesivas que se consiguen en cada paso con objeto de minimizar f y aplicar repetidamente el proceso hasta, eventualmente lograr la convergencia (ver Dorea (1983)).
La aleatoriedad de la búsqueda consiste en proponer un estimador del vector subgradiente que sirva para aportar una dirección de descenso; el estimador es generado mediante la técnica Montecarlo y el problema más importante consiste en probar que bajo ciertas condiciones los valores esperados de los números aleatorios coinciden con las funciones aleatorias evaluadas en esos puntos, asimismo, el valor esperado del estimador del subgradiente debe ser una expresión del tipo

$$\nabla g_0(x) = \mathbf{E}[\nabla \psi_0(x, W)] = 0,$$

finalmente se debe probar que la sucesión se mantiene dentro de la región factible $\mathcal{D}$, o en su defecto se puede encontrar $x_k \in \mathcal{D}$ tal que $x_k = \pi_{\mathcal{D}}(y_k)$. Aquí $\pi_D(y_k)$ denota al operador proyección del punto y_k sobre el conjunto $\mathcal{D}$, esto es, para cada $x \in \mathbb{R}^n, \pi(x) \in \mathcal{D}$ y $\parallel x - \pi(x) \parallel = \min_{z \in \mathcal{D}} \parallel x - z \parallel$. Si el punto límite $x^\star$ de la

sucesión es tal que su convergencia coincide con el mínimo buscado, se dice que tal sucesión es un algoritmo que lo resuelve.
A partir de los años 70's, la técnica ha recibido una considerable atención surgiendo importantes contribuciones a la misma a través de libros (ver Nevelson y Hasminskii (1973), Ermoliev (1988), Kushner y Clark (1978), y Pflug. Ch. G. (1996)) y publicaciones en revistas especializadas en el tema, tratando primordialmente la convergencia y velocidad de las sucesiones aleatorias, propuestas a través de algoritmos construidos para cierta clase de problemas con estructura estocástica.
En esta propuesta se desarrolla un algoritmo de optimización por búsqueda estocástica el cual incorpora información a priori en forma de esperanzas matemáticas acerca de la dirección de descenso más probable durante la búsqueda. Con esto, se unen las teorías de la medida de información y la optimización matemática. Otras aplicaciones se pueden encontrar en Pérez–Lechuga (1993).
Para su presentación, este trabajo está organizado como sigue: En la sección 2, se muestra la técnica de incorporación de información a priori y se aplica en la construcción de estimadores subgradientales a través de la máxima entropía y mínima entropía cruzada. En la sección 3 se desarrolla un algoritmo de búsqueda aleatoria que incorpora los resultados obtenidos en la sección 2 y se demuestra su convergencia. Finalmente, en la sección 4 se discuten las estrategias de implantación de nuestra propuesta.

2 Estimación de subgradientes a través de la máxima entropía y la mínima entropía cruzada

El concepto de subgradiente de una función convexa está fuertemente relacionado con el epígrafe y la teoría de los hiperplanos soporte en la teoría de la optimización. En el caso de la optimización estocástica, las direcciones de descenso se pueden establecer a través de la búsqueda aleatoria por medio de un subgradiente el cual, normalmente debe aproximarse a través de estimadores estadístico insesgados. En la mayoría de los casos, la función objetivo se encuentra también en esta condición. Sea $g(x)$, $x \in \mathbb{R}^n$ una función convexa no necesariamente diferenciable. El vector $\hat{\nabla} g(x)$ se llama subgradiente de g en x si se satisface que para toda $y \in \mathbb{R}^n$.

$$g(y) - g(x) \geq \langle \hat{\nabla} g(x), y - x \rangle. \tag{1}$$

Aquí, $\langle \cdot, \cdot \rangle$ denota al producto interno en el espacio Euclidiano $\mathbb{R}^n$
La construcción de estimadores estadísticos de los subgradientes constituyen un área de gran interés en la optimización por búsqueda aleatoria. Puede obtenerse un estimador estadístico $\xi(x)$ del subgradiente de g en x de la siguiente manera. Sea el vector aleatorio $\theta^t = (\theta_1, \theta_2, \ldots, \theta_n)$, (el supra índice t denota la operación transpuesta), con componentes independientes e idénticamente distribuidas. Sea

$$\theta_i^t = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_n^{(i)}), \;\; i = 1, 2, \ldots, \rho,$$

una muestra aleatoria de θ, y sea $\Delta > 0$. Un estimador $\xi(x)$ del subgradiente de g en x está dado por

$$\xi(x) = \frac{1}{\rho} \sum_{i=1}^{\rho} \frac{g(x + \Delta\theta_i) - g(x)}{\Delta} \theta_i.$$

En este promedio, la cantidad $g(x + \Delta\theta_i) - g(x)/\Delta$ evalúa la tasa de cambio de la función g con respecto a x, mientras que θ_i es un vector aleatorio que toma valores en la dirección de descenso más probable de acuerdo a la información a priori con que se cuenta.
Cuando se incorpora la teoría de la información en el diseño de algoritmos de búsqueda estocástica, es posible determinar direcciones de movimiento que simplifican la misma al desplazarse en la dirección que conlleva a la trayectoria esperada de descenso más probable. Suponga que el conocimiento previo de la dirección de descenso está dada en términos de la siguiente medida de información

$$\mathcal{I}(\theta) : \int_{\Theta} a_k(\theta)\, \pi(\theta) d\theta = \bar{a}_k, \quad k = 1, 2, \ldots, m, \tag{2}$$

el principio de la máxima entropia (Jaynes (1957)) aporta un método general basado en la inferencia estadística para evaluar la densidad desconocida, $\pi(\theta)$, cuando existe información actualizada sobre ésta en términos de esperanza matemática. El principio de mínima entropía cruzada establece que cuando hay un estimador *a priori* de una densidad e información adicional en términos de valores esperados, se debería tomar como un estimador a *posteriori*, aquella densidad que minimice la entropía cruzada con la *a priori* y que sea compatible con la información adicional .
Este principio es equivalente a la minimización de la entropía cruzada (Kullback (1956)) en el caso especial de espacios discretos y estimadas inicialmente unformes. El principio de máxima entropía establece que entre todas las distribuciones que sean compatibles con información adicional dada en términos de valores esperados, y en ausencia de una densidad *a priori*, se debería tomar como estimador *a posteriori* aquella que minimice la entropía.

2.1 Mínima entropía cruzada

Para encontrar un estimador a posteriori de una función de densidad cuando existe un estimador a priori $p(\theta)$ e información adicional en términos de valores esperados se tiene entonces que (2) conlleva a resolver el problema variacional (ver Venegas, M. F. (1990) (a), y Venegas, M. F. (1990) (b)).

$$\text{Minimizar } H(\pi, p) = \int_{\Theta} \pi(\theta)\, \log \frac{\pi(\theta)}{p(\theta)}\, d\theta,$$

$$\text{Sujeto a : } \begin{cases} \int_{\Theta} \pi(\theta)\, d\theta = 1 \\ \\ \int_{\Theta} a_k(\theta)\pi(\theta) d\theta = \bar{a}_k, \quad k = 1, 2, \ldots, m, \end{cases}$$

la condición necesaria de primer orden de la formulación anterior está dada por

$$\begin{cases} \pi^*(\theta) = p(\theta)\exp\{-\lambda_0 - \sum_{k=1}^m \lambda_k a_k(\theta)\}, \\ 1 - \int_\Theta \pi^*(\theta)d\theta = 0, \\ \int_\Theta (\bar{a} - a_k(\theta))\pi^*(\theta)d\theta = 0, \;\; k = 1, 2, \ldots, m, \end{cases} \tag{3}$$

donde $\lambda_0, \lambda_1, \ldots, \lambda_m$, son los multiplicadores de Lagrange asociados a las restricciones. Sustituyendo π^* en las condiciones de primer orden restantes en (3) se tiene que

$$\begin{cases} 0 = \lambda_0 - \log\{\int_\Theta \, p(\theta)\, \Pi_{k=1}^m \, e^{-\lambda_k a_k(\theta)} d\theta\} \\ 0 = \int_\Theta [a_k(\theta) - \bar{a}_k]\, p(\theta)\, \Pi_{k=1}^m e^{-\lambda_k a_k(\theta)} d\theta, \;\; k = 1, 2, \ldots, m \end{cases} \tag{4}$$

el cual es un sistema no lineal homogéneo en las variables $\lambda_0, \lambda_1, \ldots, \lambda_m$. Cuando la integral que define a λ_0 puede resolverse, entonces los demás multiplicadores pueden encontrarse a partir de las siguientes relaciones

$$\frac{\partial\, \lambda_0}{\partial \lambda_k} = -\bar{a}_k, \;\; k = 1, 2, \ldots, m. \tag{5}$$

2.2 Máxima entropía

Para encontrar un estimador a posteriori de una función de densidad cuando no existe un estimador a priori, pero se cuenta con información adicional en términos de valores esperados, el principio de máxima entropía conlleva al siguiente problema variacional

$$\text{Maximizar } H(\pi) = -\int_\Theta \pi(\theta) \log\, \pi(\theta) d\theta,$$

$$\text{Sujeto a : } \begin{cases} \int_\Theta \, \pi(\theta)\, d\theta = 1, \\ \int_\Theta \, a_k(\theta)\, \pi(\theta)\, d\theta = \bar{a}_k, \;\; k = 1, 2, \ldots, m... \end{cases}$$

La condición de primer orden del problema anterior está dada por

$$\begin{cases} \pi^*(\theta) = \exp\left\{\lambda_0 + \sum_{k=1}^m \lambda_k a_k(\theta)\right\}, \\ 1 - \int_\Theta \pi^*(\theta)d\theta = 0, \\ \int_\Theta \left[(\bar{a} - a_k(\theta))\right]\, \pi^*(\theta)d\theta = 0, \;\; k = 1, 2, \ldots, m, \end{cases} \tag{6}$$

nuevamente, $\lambda_0, \lambda_1, \ldots, \lambda_m$ son los multiplicadores de Lagrange asociados a las restricciones. Así, sustituyendo π^* en las condiciones de primer orden restantes en (6)

se tiene que

$$\begin{cases} 0 = \lambda_0 - \log\{\int_\Theta \Pi_{k=1}^m e^{\lambda_k a_k(\theta)} d\theta\} \\ \\ 0 = \int_\Theta [a_k(\theta) - \bar{a}_k]\, p(\theta)\, \Pi_{k=1}^m e^{\lambda_k a_k(\theta)} d\theta, \quad , k = 1, 2, \ldots, m, \end{cases}$$

el cual es nuevamente un sistema no lineal homogéneo en las variables $\lambda_0, \lambda_1, \ldots, \lambda_m$. Nuevamente, cuando la integral que define a λ_0 puede resolverse, entonces los demás multiplicadores pueden encontrarse a partir de las relaciones definidas en (5).
Sin pérdida de generalidad, y con objeto de simplificar la representación de (2), en lo sucesivo será denotado de la siguiente manera

$$I(\theta) = \{\Theta; a_1(\theta), a_2(\theta), \ldots, a_m(\theta);\, \bar{a}_1, \bar{a}_2, \ldots, \bar{a}_m\}.$$

Teorema 1. *Sea g una función convexa definida sobre $\mathbb{R}^n$, no necesariamente diferenciable en x, sea además $\Delta > 0$. Si*

$$\xi(x) = \frac{1}{\rho} \sum_{i=1}^{\rho} \frac{g(x + \Delta\theta_i) - g(x)}{\Delta}\, \theta_i,$$

entonces existe una matriz simétrica definida positiva (de información a priori) $\mathcal{G}(I)$, y un vector $g(x, I) \in \mathbb{R}^n, g(x, I) \geq 0$ tal que

$$\mathbf{E}\{\xi(x) \mid x, I(\theta)\} = \mathcal{G}(I)\hat{\nabla} g(x) + g(x, I),$$

donde $\mathbf{E}\{\cdot\}$ es el operador esperanza, y $\mathcal{G}(I)$ tiene la forma

$$\mathcal{G}_{ij} = \begin{cases} \int_\Theta \theta^2 \exp\{\lambda_0 + \sum_{k=1}^m \lambda_k a_k(\theta)\}\, d\theta, \ i = j, \\ \\ \left[\int_\Theta \theta \exp\{\lambda_0 + \sum_{k=1}^m \lambda_k a_k(\theta)\}\, d\theta\right]^2, \ i \neq j, \end{cases} \tag{7}$$

donde existe información a priori $I(\theta)$. Asimismo, $\mathcal{G}(I)$ toma la forma

$$\mathcal{G}_{ij} = \begin{cases} \int_\Theta \theta^2 \exp\{-\lambda_0 - \sum_{k=1}^m \lambda_k a_k(\theta)\}\, d\theta, \ i = j, \\ \\ \left[\int_\Theta \theta \exp\{-\lambda_0 - \sum_{k=1}^m \lambda_k a_k(\theta)\}\, d\theta\right]^2, \ i \neq j, \end{cases} \tag{8}$$

cuando existe un estimador a priori $p(\theta)$ e información a priori sobre $I(\theta)$.

La prueba se presenta en el anexo A de este documento.
Del teorema anterior, es inmediato que

$$\mathbf{E}\{\mathcal{G}^{-1}(I)\, \xi(x) \mid x, I(\theta)\} = \hat{\nabla} g(x) + h(I),$$

donde $h(x, I) = \mathcal{G}^{-1}(I)\, g(x, I)$, es el sesgo condicional de $\xi(x)$. Para ilustrar la aplicación de este resultado a continuación se muestran algunos ejemplos.

Ejemplo 1. En los siguientes casos, no existe información a priori.

1. Suponga que $I(\theta) = \{(-\infty, \infty); \theta, (\theta - \mu_0)^2; \mu_0,\ \sigma_0^2\},\ \sigma_0 > 0$, entonces la matriz asociada tiene la forma

$$\mathcal{G}_{ij}(I) = \sigma_0^2 + \mu_0^2, \quad i = j, \ \text{ y } \ \mathcal{G}_{ij}(I) = \mu_0^2, \quad i \neq j.$$

2. Si la información a priori está dada por $I(\theta) = \{(0, \infty); \theta, \log\theta; \alpha\beta^{-1}, \psi(\alpha) - \log\beta\}\ \alpha,\ \beta > 0$, donde ψ es la función digamma, entonces la matriz asociada tiene la forma

$$\mathcal{G}_{ij}(I) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}, \quad i = j, \ \text{ y } \ \mathcal{G}_{ij}(I) = \left(\frac{\alpha}{\beta}\right)^2, \quad i \neq j.$$

3. Si la información a priori tiene la forma $I(\theta) = \{(0, \infty); \log\theta, \theta^\delta; \delta^{-1}[\psi(\alpha) - \log\beta], \beta^{-1}\},\ \alpha,\ \beta,\ \delta > 0$, la matriz de información tiene la forma

$$\mathcal{G}_{ij}(I) = \frac{\Gamma(\alpha\beta + \frac{1}{\delta})}{\Gamma(\alpha)} (\alpha\beta)^{\alpha - \alpha\beta - (2/\delta)}, \quad i = j,$$

y

$$\mathcal{G}_{ij}(I) = \left(\frac{\Gamma(\alpha\beta + (1/\delta))}{\Gamma(\alpha)} (\alpha\beta)^{\alpha - \alpha\beta - (1/\delta)}\right)^2, \quad i \neq j.$$

Ejemplo 2. Suponga ahora que la información a priori es que θ se encuentra en alguna región $\Theta = (b_1, b_{m+1})$. Suponga también que se asignan pesos, $\gamma_1, \gamma_2, \ldots, \gamma_m \geq 0$, tales que, $\sum_{k=1}^m \gamma_k = 1$, entonces, θ se encuentra en las subregiones $A_k = (b_k, b_{k+1}],\ k = 1, 2, \ldots, m-1$ y $A_m = (b_m, b_{m+1})$, con $b_1 < b_2 \ldots < b_{m+1}$, $m > 2$, la cual constituye una partición de $\Theta = (b_1, b_{m+1})$. Así, la información a priori puede escribirse como

$$\int_\Theta I_{A_k}(\theta)\, \pi(\theta) d\theta = \gamma_k > 0, \quad k = 1, 2, \ldots, m, \quad \sum_{k=1}^m \gamma_k = 1, \tag{9}$$

Aplicando el principio de mínima entropía, las condiciones necesarias dadas en (4), se transforman ahora en

$$\begin{cases} 0 = \lambda_0 - \log\{\int_\Theta \Pi_{k=1}^m e^{-\lambda_k I_{A_k}(\theta)} d\theta\} \\ \\ 0 = \lambda_0 + \lambda_k - \log\left\{\gamma_k^{-1} \int_\Theta I_{A_k}(\theta) d\theta\right\}, \quad k = 1, 2, \ldots, m. \end{cases} \tag{10}$$

Haciendo el siguiente cambio de variable $\omega_0 = e^{\lambda_0}$ y $\omega_k = e^{-\lambda_k},\ k = 1, 2, \ldots, s$, (10) se convierte ahora en el siguiente sistema lineal homogéneo

$$\begin{pmatrix} -1 & u_1 & u_2 & \ldots & u_m \\ -1 & v_1 & 0 & \ldots & 0 \\ -1 & 0 & v_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 0 & 0 & \ldots & v_m \end{pmatrix} \begin{pmatrix} \omega_0 \\ \omega_1 \\ \omega_2 \\ \vdots \\ \omega_m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{11}$$

donde $u_k = \int_\Theta I_{A_k}(\theta)d\theta$, y $v_k = \gamma_k^{-1}u_k,\ k = 1, 2, \ldots, m$. La matriz (11) juega un papel muy importante en la mínima entropía cruzada, y en lo sucesivo será denotada por $M(\gamma_1, \gamma_2, \ldots, \gamma_m)$. Se puede demostrar que el determinante de la matriz (11) viene dado por

$$\left(\frac{\sum_{k=1}^m \gamma_k^{-1}}{\Pi_{k=1}^m \gamma_k}\right) \Pi_{k=1}^m u_k.$$

Note que como $\sum_{k=1}^m \gamma_k = 1$ entonces se garantiza la existencia de una solución no trivial. Así, solucionando el sistema y usando (8) se tiene que

$$\mathcal{G}_{ij} = \begin{cases} \frac{1}{3}\sum_{k=1}^n \gamma_k(b_k^2 + b_k b_{k+1} + b_{k+1}^2),\ i = j, \\ \\ \left(\frac{1}{2}\sum_{k=1}^n \gamma_k(b_k + b_{k+1})\right)^2,\ i \neq j, \end{cases}$$

Suponga ahora que existe un estimador a priori, digamos $p(\theta) = \sum_{k=1}^m \beta_k\, u_k^{-1}\, I_{A_k}(\theta)$, $\theta \in \Theta$, donde $\beta_k > 0,\ k = 1, 2, \ldots, m$, $\sum_{k=1}^m \beta_k = 1$, e información a priori como en (9), la cual se expresa como cambios en los pesos. Usando el mismo cambio de variable anteriormente mencionado, se tiene el siguiente sistema homogéneo lineal en términos de $M(\gamma_1, \gamma_2, \ldots, \gamma_m)$

$$M(\gamma_1, \gamma_2, \ldots, \gamma_m) = \text{Diag}\ (1, \bar{b}_1 u_1^{-1}, \bar{b}_2 u_2^{-1}, \ldots, \bar{b}_m u_m^{-1})\ \Omega = \mathbf{0},$$

donde $\Omega = (\omega_0, \omega_1, \ldots, \omega_n)^t$ y $\mathbf{0}$ es el vector cero. Aquí, la información aportada por la estimada inicial se incorpora a través de una matriz diagonal. La solución está dada por

$$\Omega^* = (1, \bar{b}_1^{-1}\bar{a}_1, \bar{b}_2^{-1}\bar{a}_2, \ldots, \bar{b}_m^{-1}\bar{a}_m)^t.$$

Note que en este caso el estimador a priori, no tiene ningún efecto sobre el estimador a posteriori.

Ejemplo 3. (Un caso de información redundante). Suponga ahora que se tiene información adicional sobre θ en términos de

$$I(\theta) = \{(0, \infty); \theta, \beta^{-1}\},\ \beta > 0. \tag{12}$$

$$\mathcal{G}_{ij} = \frac{2}{\beta^2},\ i = j, \quad \text{y} \quad \mathcal{G}_{ij} = \frac{1}{\beta^2},\ i \neq j. \tag{13}$$

Si ahora se incorpora información en (12) en términos de

$$I(\theta) = \{(0, \infty); \theta, \log\theta; \beta^{-1}, -\mathcal{K} - \log\beta\},\ \ \beta > 0,$$

donde $\mathcal{K} \approx 0.5772$ es la constante de Euler, aquí tampoco se tiene cambio en $\mathcal{G}(I)$, por lo tanto la información agregada es redundante. No obstante los resultados obtenidos existen algunos casos donde esto no sucede. Suponga ahora que se agrega información en (12) como sigue

$$I(\theta) = \{(0, \infty); \theta, \log\theta; \beta^{-1}, \psi(\alpha) - \log\alpha\beta\},\ \ \alpha, \beta > 0.$$

$$\mathcal{G}_{ij} = \frac{1}{\alpha\beta^2} + \frac{1}{\beta^2},\ i = j \quad \text{y} \quad \mathcal{G}_{ij} = \frac{1}{\beta^2},\ i \neq j. \tag{14}$$

Ejemplo 4. Acerca de la sensibilidad de la incorporación de la información
Suponga ahora que se tiene información inicial sobre θ, que está dada por

$$I(\theta) = \{(0,\infty); \theta; \beta^{-1/\alpha}\,\Gamma(1+\alpha^{-1})\},\ \ \alpha,\ \beta > 0,$$

la solución de máxima entropía para este caso es

$$\pi^*(\theta) = \{\beta^{\frac{-1}{\alpha}}\,\Gamma(1+\alpha^{-1})\}^{-1}\ \exp\left\{-\{\beta^{\frac{-1}{\alpha}}\ \Gamma(1+\alpha^{-1})\}^{-1}\,\theta\right\},\ \theta > 0,$$

Aquí, la forma de la matriz $\mathcal{G}(I)$ es como se definió en (13).
Un caso interesante de este ejemplo se obtiene cuando se agrega información a priori sobre θ en (14) de tal forma que

$$I(\theta) = \{(0,\infty); \theta, \log\theta; \beta^{\frac{1}{\alpha}}\ \Gamma(1+\alpha^{-1})\}^{-1},\ \psi[\Gamma(1+\alpha^{-1})\}] - \alpha^{-1}\log\beta\},\ \ \alpha,\ \beta > 0.$$

La solución para el problema variacional de máxima entropía $\pi^*(\theta)$, es la distribución Gamma con parámetros $\beta^{\frac{1}{\alpha}}$ y $\Gamma(1+\alpha^{-1})$. La matriz $\mathcal{G}(I)$ es similar a la obtenida en el punto 2. del ejemplo 1.
Otro caso importante se presenta cuando en (14) la información viene dada por

$$I(\theta) = \{(0,\infty); \theta, \log\theta, \theta^{\alpha}; \beta^{\frac{-1}{\alpha}}\ \Gamma(1+\alpha^{-1}),\ -\alpha^{-1}(\mathcal{K} - \log\beta), \beta^{-1}\},\ \ \alpha, \beta > 0,$$

Aquí la solución de máxima entropía está dada por la distribución Weibull de dos parámetros α y β

$$\pi^*(\theta) = \alpha\,\beta\,\theta^{\alpha-1}\,e^{-\beta\theta^{\alpha}},\ \theta > 0,$$

y la matriz de información a priori es

$$\mathcal{G}_{ij}(I) = \left(\frac{1}{\beta}\right)^{\frac{2}{\alpha}}\ \Gamma(1+\frac{2}{\alpha}),\ i = j, \quad \text{y} \quad \mathcal{G}_{ij}(I) = \left[\left(\frac{1}{\beta}\right)^{\frac{1}{\alpha}}\ \Gamma(1+\alpha)^{-1}\right]^2,\ i \neq j$$

Con los ejemplos anteriores, ahora se procederá a incorporar la información obtenida en el diseño del algoritmo de búsqueda aleatoria.

3 Construcción del algoritmo de búsqueda aleatoria con información a priori

Algunos algoritmos de búsqueda estocástica utilizan superficies de prueba (tales como esferas o símplex en $\mathbb{R}^n$) para localizar una direción de descenso. En esta técnica se deben generar aleatoriamente $\mathcal{M}$ puntos sobre la superficie de prueba, y, posteriormente, determinar la trayectoria de descenso con base a aquella dirección que

aporta el mejor valor para este propósito. La eficiencia de la búsqueda se relaciona de manera directa con el número de puntos $\mathcal{M}$ explorados (ver Pérez-Lechuga (1993)) sobre la superficie de prueba. Una forma de eficientar esta búsqueda consiste en proporcionar al agoritmo la dirección más probable de descenso en términos de información a priori. Esta información es proporcionada a través de la matriz $\mathcal{G}(\mathcal{I})$. Sea $g(x)$, $x \in \mathbb{R}^n$ una función convexa no necesariamente diferenciable, en esta sección, el objetivo es minimizar a $g(x)$ a través de la generación de una sucesión aleatoria de puntos $x_s \in \mathbb{R}^n$ tales que

$$\mathbf{P}_\theta \{ \lim_{s\to\infty} g(x_s) = g(x^*) = \min g(x)\} = 1.$$

Para lograr lo anterior, en cada iteración s defina al vector aleatorio

$$x_{s+1} = x_s - \varphi_s \, \mathcal{G}(I)^{-1} \, \xi(x_s), \;\; s = 0, 1, 2 \ldots \tag{15}$$

donde x_0 es un valor arbitrario en $\mathbb{R}^n$, y además

$$\mathbf{E}\{\mathcal{G}^{-1}(I)\xi(x) \mid x, I(\theta)\} = \hat{\nabla} g(x_s) + h(I),$$

donde $h(x, I) = \mathcal{G}^{-1}(I) g(x, I)$. Aquí, se tomará la muestra $\theta_i^t = (\theta_1^{(i)}, \ldots, \theta_n^{(i)})$, $i = 1, \ldots, \rho$, de un vector aleatorio $\theta^t = (\theta_1, \ldots, \theta_n)$, cuya distribución es compatible con la información a priori $I(\theta) = \{\Theta; a_1(\theta), \ldots, a_m(\theta); \bar{a}_1, \ldots, \bar{a}_m\}$.

Teorema 2. *Considere la sucesión aleatoria definida en* (15)*, suponga además que se cumplen las siguientes condiciones*

1. *Existe una sucesión de constantes φ_s tales que $\sum_{s=0}^{\infty} \varphi_s = \infty$, $\varphi_s \geq 0$*
2. *Existe una constante $\mathcal{A}$ tal que $\mathbf{E}\{\| x_s \| \mid x_s\} \leq \mathcal{A}$, $\forall\ s = 0, 1, 2, \ldots$*
3. *Existe una constante $\mathcal{B}$ y una sucesión $\{\beta_s\}_{s=0}^{\infty}$ tal que $\mathbf{E}\{\| \xi(x_s) \|^2 \mid x_s, I(\theta)\} \leq \beta_s \leq \mathcal{B}$, $\forall\ s = 0, 1, 2, \ldots$*
4. *Existe una sucesión $\{\gamma_s\}_{s=0}^{\infty}$ tal que $\mathbf{E}\{\| h(x_s, I) \| \mid x_s\} \leq \gamma_s$, con $\sum_{s=0}^{\infty} \alpha_s(\alpha_s + \gamma_s) < \infty$.*

Entonces $\mathbf{P}_\theta\{\lim_{s\to\infty} g(x_s) = g(x^\star)\} = 1$.

Demostración. Primero note que

$$\mathbf{E}\left\{\| \xi(x_s) \|^2 \mid x_s, I(\theta)\right\} = \sum_{j=1}^{n} (\mathbf{Var}\{\xi_{sj}(x_s) \mid x_s, I(\theta)\} + \mathbf{E}^2\{\xi_{sj}(x_s) \mid x_s, I(\theta))$$

donde $\xi_{sj}(x_s)$ es la j-ésima componente de $\xi(x_s)$. Así, si $\sum_{j=1}^{n} \mathbf{Var}[\xi \mid x_s, I(\theta)]$ es acotada y si $\| \hat{\nabla} g(x_s) \|$ es también acotada, entonces la condición (2) se cumple. Sea x^* una solución arbitraria del problema, entonces

$$\| x^* - x_{s+1} \|^2 \leq \| x^* - x_s + \varphi_s \mathcal{G}(I)^{-1} \xi(x_s) \|^2 =$$

$$\| x^* - x_s \|^2 + 2\varphi_s \langle \mathcal{G}(I)^{-1}\xi(x_s), x^* - x_s\rangle + \varphi_s^2 \| \mathcal{G}(I)^{-1}\xi(x_s) \|^2 .$$

Tomando la esperanza matemática condicional en ambos lados de la desigualdad se tiene que

$$\begin{aligned}
&\mathbf{E}\{\| x^* - x_{s+1} \|^2 \mid x_s\} \\
&\quad \leq \| x^* - x_s \|^2 + 2\varphi_s (x^* - x_s)[\hat{\nabla} g(x_s) + h(I)] + \varphi_s^2 \mathbf{E}\{\| \mathcal{G}(I)^{-\infty} \xi(x_s) \|^2 | \, x_s\} \\
&\quad = \| x^* - x_s \|^2 + 2\varphi_s \langle \hat{\nabla} g(x_s), x^* - x_s\rangle + 2\varphi_s^2 \mathbf{E}\{\| \mathcal{G}(I)^{-1} \xi(x_s) \|^2 | \, x_s\}, \qquad (16)
\end{aligned}$$

En este caso, el valor esperado de la norma del vector $\tilde{x} = (x^* - x_{s+1}) \in \mathbb{R}^n$ se define como

$$\mathbf{E}\{\| \tilde{x} \|\} = \int_\Omega [\tilde{x}_1^2(\omega) + \tilde{x}_2^2(\omega) + \ldots + \tilde{x}_n^2(\omega)]^{\frac{1}{2}} \mathcal{P}(d\omega),$$

donde Ω es el espacio muestral asociado al espacio de probabilidades $(\Omega, \mathcal{F}, \mathcal{P})$. De la convexidad de g se tiene que $g(x^*) - g(x_s) \leq 0$. Así, usando (1) y por la la desigualdad de Cauchy-Schwartz, (16) toma la forma

$$\begin{aligned}
\mathbf{E}\{\| x^* - x_{s+1} \|^2 \mid x_s\} \leq \| x^* - x_s \|^2 + 2\varphi_s \ \| h(I) \| [\| x^* \| + \| x_s \|] \\
+ \ 2\varphi_s[g(x^*) - g(x_s)] + \varphi_s^2 \mathbf{E}\{\| \mathcal{G}(I)^{-1} \xi(x_s) \|^2 | \, x_s\},
\end{aligned}$$

por las condiciones (1), (3) y (4) se tiene que

$$\mathbf{E}\{\| x^* - x_{s+1} \|^2 \mid x_s\} \ \leq \| x^* - x_s \|^2 + 2\varphi_s \gamma_s [\| x^* \| + \| x_s \|] + \varphi_s^2 \mathcal{B}.$$

Simplificando el proceso anterior se tiene que

$$\begin{aligned}
\mathbf{E}\{\| x^* - x_{s+1} \|^2 \mid x_s\} \ &\leq \| x^* - x_s \|^2 + \sum_{s=0}^{\infty} \varphi_s \mathbf{E}\{\langle \nabla g(x_s), x^* - x_s \mid x_s\rangle \mid x_s\} \\
&\quad + 2\sum_{s=0}^{\infty} \varphi_s \gamma_s [\| x^* \| + \| x_s \|] + \sum_{s=0}^{\infty} \varphi_s^2 \mathcal{B}.
\end{aligned}$$

De lo anterior se sigue que

$$\sum_{s=0}^{\infty} \varphi_s \mathbf{E}\{\langle \hat{\nabla} g(x_s, x^* - x_s\rangle \mid x_s\} \geq -\infty,$$

pero por la condición (1) se tiene que $\mathbf{E}\{\langle \hat{\nabla} g(x_s, x^* - x_s\rangle \mid x_s\} \to 0$ cuando $s \to \infty$. Note que existe una subsucesión $\{S_t\}$, tal que $[\{\langle \hat{\nabla} g(x_s, x^* - x_s\rangle \mid x_s\} \to 0$ con probabilidad 1 cuando $t \to \infty$. Para casi todas las ω, la sucesión $\{\| x_s(\omega) \|\}$ es acotada, esto es, para casi todas las ω, $[\{\langle \hat{\nabla} g(x_{s_t}(\omega)), x^* - x_{s_t}(\omega)\rangle\} \to 0$ con probabilidad 1 cuando $t \to \infty$, y por lo tanto la sucesión converge al óptimo. $\square$

4 Estrategias de implantación

La sucesión (15) presupone la existencia de puntos $x_s \in \mathcal{D}$ que convergen con probabilidad 1 hacia el óptimo de la función g. Así, la región que rodea a cada punto de la sucesión constituye una región aleatoria (en realidad es una esfera de dimensión n) por donde la sucesión x_s atraviesa. Cuando se incorpora la información, la trayectoria se mueve ahora a través de una esfera deformada, y la deformación creada depende de la densidad asociada a la información incorporada.
Así, en la implantación del algoritmo, es conveniente "crear" una superficie de prueba donde la dirección del vector apunte hacia el valor esperado de dicha distribución. La selección de la densidad asociada constituye por sí misma una línea de investigación, pues se relaciona directamente con la eficiencia del algoritmo y la concentración de masa de los puntos candidatos a definir la nueva dirección de la trayectoria. En la práctica, resulta benéfico considerar dicha distribución como normal (para el caso de $\mathbb{R}^2$ la normal bivariada).

5 Conclusiones

En este trabajo se desarrolló un algoritmo de búsqueda aleatoria para optimizar algunos modelos con funciones estocásticas no restringidas usando información a priori. Se demuestra la convergencia del mismo y se ilustran varios ejemplos de aplicación en diversas distribuciones de probabilidad. Algunas líneas de investigación derivadas de este trabajo son: a) determinar la eficiencia de búsqueda del algoritmo propuesto para cada densidad asociada a las medidas de información, b) determinar el efecto de la varianza de la densidad considerada en las medida de la información en la rapidez y precisión del algoritmo, c) diseño de superficies de prueba (símplex) que incorporen la información a priori en su estructura, así como la determinación de las densidades apropiadas a cada caso.

6 Anexo A

Prueba del Teorema 1

Demostración. Usando (1) se tiene que para cada $i = 1, 2, \ldots, \rho$,

$$\frac{g(x + \Delta\theta_i) - g(x)}{\Lambda}\theta_i \geq \langle \hat{\nabla} g(x), \theta_i \rangle \theta i$$

entonces

$$\mathbf{E}\left\{ \frac{g(x + \Delta\theta_i) - g(x)}{\Lambda}\theta_i \mid x, I(\theta) \right\} \geq \mathbf{E}\left\{ \langle \hat{\nabla} g(x), \theta_i \rangle \theta_i \mid x, I(\theta) \right\}.$$

Donde el producto $\langle \hat{\nabla} g(x), \theta_i \rangle \theta_i$, puede escribirse como $\mathbf{H}(\theta_i)\hat{\nabla} g(x)$, con $H_{lj}(\theta_i) = \theta_l^i \theta_j^i$, $l, j = 1, 2 \ldots, n$. Entonces

$$\mathbf{E}\left\{ \frac{g(x + \Delta\theta_i) - g(x)}{\Lambda} \theta_i \mid x, I(\theta) \right\} \geq \mathbf{E}\left\{ \mathbf{H}(\theta_i) \mid I(\theta) \right\} \hat{\nabla} g(x).$$

Note que como la matriz $\mathbf{E}\{\mathbf{H}(\theta_i) \mid I(\theta)\}$ es independiente de i, entonces se puede escribir como $\mathbf{E}\{\mathbf{H}(\theta_i) \mid I(\theta)\} = \mathcal{G}(I)$, $\forall\, i$. De (3) y (6), se obtiene (7) y (8) respectivamente. El resultado final se obtiene sumando sobre i y dividiendo por ρ. □

7 Reconocimientos

Gran parte del material expuesto en este trabajo fue obtenido de los resultados de investigación generados por los autores en forma conjunta con el Dr. Francisco Venegas Martínez de la División de Estudios de Posgrado de la Universidad Nacional Autónoma de México, y de los resultados publicados por el propio Dr. Venegas en Venegas, M. F. (1990) (a), y Venegas, M. F. (1990) (b).

Referencias

Dorea, C. C. Y. (1983), *Expected Number of Steps of a Random Optimization Method*, Journal of Optimization Theory and Applications , **39**, pp. 165-171.

Dvoretzky, M. N. (1956), *On Stochastic Approximation in Proceeding of the Third Berkely Symposium On Mathematical Statistical and Probability*, Vol **1**, pp. 39-55.

Ermoliev, Y., and Wets, J-B. (Eds.) (1988), *Numerical Techniques for Stochastic Optimization*, Springer-Verlag.

Jaynes, E.T. (1957), *Information Theory and Statistical Mechanics I*, Phis. Rev., **106**, pp. 620-630.

Kiefer, J., and Wolfowitz, J. (1952), *Stochastic Estimation of the Maximum of Regresion Function*, Ann. Math. Stat., Vol. **23**, pp. 462-466.

Kullback, S. (1956), *Information Theory and Statistics*. New York, Wiley.

Kushner, C.Y., and Clark., D.S. (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems,* Springer Verlag, New York.

Nevelson, M. B. and, R. Z. Hassminskii, (1973), *Stochastic Approximation and Recursive Estimation*, American Mathematical Society.

Pfluj, Ch. G., (1996), *Optimization of Stochastic Models: The Interface Between Simulation and Optimization*, Kluwer Academic Publishers, London.

Pérez, L. G. (1993), *Un Algoritmo Para la Optimización Estocástica de Algunos Modelos Dinámicos*, Tesis Doctoral, UNAM, DEPFI. Depto. de Sistemas.

Robbins, H.S., and Monro, S. (1951), *Stochastic Approximation Methods*, Ann. Math. Stat., Vol **22**, pp. 400-407.

Shore, J. E., and Johnson, R. W. (1980), *Axiomatic Derivation of the Principle of Maximun Entropy and the Principle of Minimum Cross Entropy*, IEEE Trans. Inform. Theory, **IT-26**, pp. 26-37.

Shore, J. E. and Johnson, R. W. (1980), *Properties of Cross- Entropy Minimization Entropy Minimization*, IEEE Trans. Inform. Theory, **IT-27**, pp. 472-482.

Venegas, M. F. (1990) (a), *On Regularity and Optimality Conditions for Maximum Entropy Priors*, REBRAPE, **4**, pp. 105-136.

Venegas, M. F. (1990) (b), *Información Suplementaria a Priori, Aspectos Computacionales y Clasificación*, Estadística, IASI, **42**, No. 139, pp 64-80.