

# MAT-468: Sesión 11, Optimización Monte Carlo

Felipe Osorio

<http://fosorios.mat.utfsm.cl>

Departamento de Matemática, UTFSM



## Objetivo:

Se desea un enfoque **basado en simulación** para resolver el problema:

$$\max_{\theta \in \Theta} h(\theta).$$

La diferencia con el enfoque numérico está en el tratamiento de la función  $h$ , en este caso desde el punto de vista probabilístico.



Si  $\Theta$  es acotado (lo que puede ser conseguido mediante reparametrización), entonces un primer enfoque es simular desde una uniforme sobre  $\Theta$ .

Es decir, considere  $u_1, \dots, u_m \sim U(\Theta)$ , y usamos la aproximación:

$$h_m^* = \max\{h(u_1), \dots, h(u_m)\}.$$

Este método converge (cuando  $m \rightarrow \infty$ ) pero puede ser muy lento dado que no toma en cuenta **ninguna característica específica** de  $h$ . Otras distribuciones en lugar de la uniforme podrían ser mejores.

## Observación:

- ▶ Note que para funciones de verosimilitud costosas de evaluar, este método es poco práctico.
- ▶ Exploración puede ser difícil cuando  $\Theta$  no es convexo, en cuyo caso métodos basados en simular una muestra  $\theta_1, \dots, \theta_m$ , pueden ser útiles.



## Ejemplo:

Considere la función:

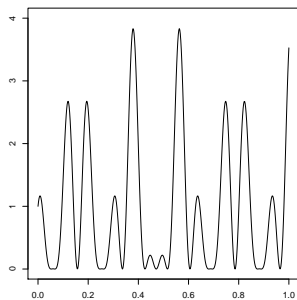
$$h(x) = \{\cos(50x) + \sin(20x)\}^2, \quad 0 < x < 1.$$

El máximo de  $h(x)$  es 3.832544. Mientras que usando 5000 puntos desde  $U(0, 1)$ , obtenemos

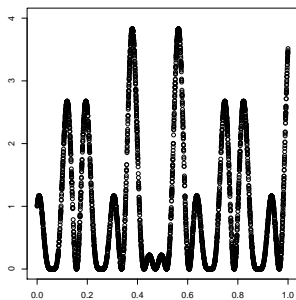
$$h_m^* = 3.832466,$$

que es una aproximación muy buena para el máximo.





(a)



(b)

Suponga que  $h$  está relacionado con una función de probabilidad. Por ejemplo, si  $h$  es positivo y si

$$\int_{\Theta} h(\theta) d\theta < +\infty,$$

entonces el problema de  $\max_{\theta \in \Theta} h(\theta)$  es equivalente a hallar las modas de la densidad  $h$ .

En ocasiones es posible transformar  $h(\theta)$  en otra función  $H(\theta)$  tal que

- (i)  $H$  es no negativa y  $\int H < +\infty$ .
- (ii) las soluciones de  $\max_{\theta} h(\theta)$  son aquellas que maximizan  $H(\theta)$  en  $\Theta$ .

La idea es escribir  $H(\theta)$  como

$$H(\theta) = \exp(h(\theta)/T),$$

y se escoge  $T$  de manera de acelerar la convergencia o evitar máximos locales.



## Métodos gradiente estocásticos

Estos procedimientos producen una secuencia de soluciones  $\{\boldsymbol{\theta}^{(k)}\}$  que converge a una solución exacta del problema  $\max_{\boldsymbol{\theta}} h(\boldsymbol{\theta})$  digamos  $\boldsymbol{\theta}^*$ . La secuencia es construida como

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \gamma_k \nabla h(\boldsymbol{\theta}^{(k)}), \quad \gamma_k > 0, \quad (1)$$

donde  $\nabla h$  es el gradiente de  $h$ , mientras que la secuencia  $\{\gamma_k\}_{k \geq 1}$  es escogida de manera apropiada para asegurar la convergencia del algoritmo.

Cuando  $h$  es una función ruidosa el algoritmo anterior puede ser modificado mediante perturbaciones estocásticas y aún alcanzar convergencia.

(a) Suponga que la secuencia  $\{\gamma_k\}_{k \geq 1}$  satisface la condición

$$\sum_{k=1}^{\infty} \gamma_k = \infty, \quad \sum_{k=1}^{\infty} \gamma_k^2 < \infty. \quad (2)$$

(b) Substituir el gradiente  $\nabla h$  por la **diferencia**, como

$$\nabla h_k(\boldsymbol{\theta}) \approx \frac{h(\boldsymbol{\theta} + \lambda_k z_k) - h(\boldsymbol{\theta} - \lambda_k z_k)}{2\lambda_k} z_k = \frac{\Delta(\boldsymbol{\theta}, \lambda_k z_k)}{2\lambda_k} z_k$$

donde  $\{\lambda_k\}_{k \geq 1}$  es una segunda secuencia como en (2) y  $z_j$  es uniforme distribuido sobre la esfera unitaria  $\|z\| = 1$ .



Esto lleva a la actualización:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \frac{\gamma_k}{2\lambda_k} \Delta(\boldsymbol{\theta}^{(k)}, \lambda_k z_k) z_k.$$

Aunque el método no sigue la dirección de búsqueda dada por el gradiente permite poder escapar de óptimos locales.

La convergencia  $\{\boldsymbol{\theta}^{(k)}\}$  a la solución depende de la elección de las secuencias  $\{\gamma_k\}$  y  $\{\lambda_k\}$ . Se necesita la condición (2), mientras que  $\{\lambda_k\}$  debe decrecer más lentamente, tal que la serie

$$\sum_{k=1}^{\infty} \left( \frac{\gamma_k}{\lambda_k} \right)^2 < \infty.$$





## Ejemplo:

Considere minimizar la función en  $\mathbb{R}^2$

$$h(x, y) = (x \sin(20y) + y \sin(20x))^2 \cosh(\sin(10x)x) \\ + (x \cos(10y) - y \sin(10x))^2 \cosh(\cos(20y)y), \quad (x, y) \in [-1, 1]^2,$$

cuyo mínimo global es 0 en el punto (0,0). Esta función tiene muchos mínimos globales y es bastante desafiante para los algoritmos estándar de minimización.

Considere la siguiente tabla:<sup>1</sup>

$\gamma_k$	$\lambda_k$	$\theta^{(K)}$	$h(\theta^{(K)})$	iter ( $K$ )
$1/(10k)$	$1/(10k)$	(-0.1660, 1.020)	1.28700	50
$1/(100k)$	$1/(100k)$	( 0.2690, 0.786)	0.00013	93
$1/(10 \log(1 + k))$	$1/k$	( 0.0004, 0.245)	$4.2 \cdot 10^{-6}$	58

---

<sup>1</sup>Estimación inicial: (0.65, 0.80)

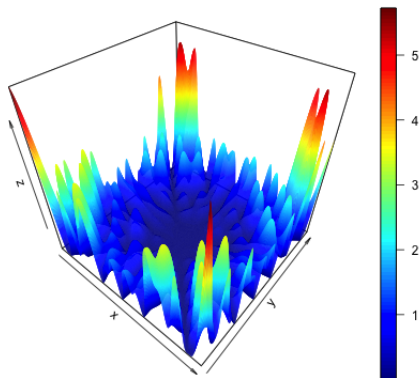


Figure: Gráfico de la función  $h(x, y)$ .

## Tarea:

Implementar el [algoritmo gradiente estocástico](#) con  $h(x, y)$  definida en el ejemplo anterior, y considere los siguientes escenarios:

1.

$$\gamma_k = \frac{1}{\log(1+k)}, \quad \lambda_k = \frac{1}{\log(1+k)^{1/10}}$$

2.

$$\gamma_k = \frac{1}{100 \log(1+k)}, \quad \lambda_k = \frac{1}{100 \log(1+k)}$$

3.

$$\gamma_k = \frac{1}{1+k}, \quad \lambda_k = \frac{1}{(1+k)^{1/2}}$$

4.

$$\gamma_k = \frac{1}{1+k}, \quad \lambda_k = \frac{1}{(1+k)^{1/10}}$$



# Simulated Annealing

El procedimiento de **simulated annealing** construye una secuencia  $\{\boldsymbol{\theta}^{(k)}\}$  simulando desde una densidad instrumental  $\pi_k$ .

Una elección estándar para  $\pi_k$  está basado en la transformación de **Boltzman-Gibbs** de  $h$ ,

$$\pi_k(\boldsymbol{\theta}) \propto \exp(h(\boldsymbol{\theta})/T_k),$$

donde  $\{T_k\}$  es una secuencia decreciente de *temperaturas*.

## *Observación:*

Conforme  $T_k \downarrow 0$  los valores simulados se concentran en un área cercana la máximo local de  $h$ .



El siguiente algoritmo fue propuesto por Metropolis et al., (1953)<sup>2</sup> y está relacionado con el [algoritmo Metropolis-Hastings](#)

En efecto, para actualizar  $\theta^{(k)}$  a  $\theta^{(k+1)}$  se debe generar una dirección  $\delta$  desde una densidad simétrica  $g$  y el nuevo valor  $\theta^{(k+1)}$  se construye como:

$$\theta^{(k+1)} = \begin{cases} \theta^{(k)} + \delta, & \text{con probabilidad } \rho = \min\{\exp(\Delta h/T_k), 1\} \\ \theta^{(k)}, & \text{con probabilidad } 1 - \rho \end{cases},$$

donde  $\Delta h = h(\theta^{(k)} + \delta) - h(\theta^{(k)})$ .

---

<sup>2</sup>The Journal of Chemical Physics 21, 1087-1092.



Es decir, el método propone una **perturbación simétrica** del valor actual  $\theta^{(k)} + \delta$ .

- ▶ Si la perturbación **incrementa**  $h$  es decir, si

$$h(\theta^{(k)} + \delta) \geq h(\theta^{(k)}),$$

el paso es **automáticamente aceptado**.

- ▶ Por otro lado, si

$$h(\theta^{(k)} + \delta) < h(\theta^{(k)}),$$

aún podemos aceptar una dirección candidata con probabilidad  $\rho > 0$ .

## *Observación:*

Permitiendo movimientos en los que  $h$  puede decrecer llevan a que el método de simulated annealing tenga la habilidad de **escapar de óptimos locales**.

Obviamente el desempeño del algoritmo depende de  $g$  y  $\{T_k\}$ .<sup>3</sup>

---

<sup>3</sup>Una elección común es  $T_k = 1 / \log(1 + k)$ .



## Tarea:

Considere la función:

$$h(x) = \{\cos(50x) + \sin(20x)\}^2, \quad 0 < x < 1.$$

Implementar el algoritmo **Simulated Annealing (SA)** considerando  $g$  como  $U(-\zeta, \zeta)$ ,  $T_k = 1/\log(1+k)$  y  $\zeta = \sqrt{T_k}$ .

