Curso:
Métodos de Monte Carlo Unidad 2, Sesión 6:
Integración por Monte Carlo

Departamento de Investigación Operativa
Instituto de Computación, Facultad de Ingeniería
Universidad de la República, Montevideo, Uruguay

1er semestre - 2021

Integración en múltiples variables

En esta sesión discutimos como emplear el método de Monte Carlo para calcular la integral de una función φ sobre una región $\mathcal{R}$ acotada (que supondremos ha sido escalada de manera que $\mathcal{R} \subseteq \mathcal{J}^m$).

Formalmente, consideremos la integral en el sentido de Lebesgue

$$\zeta(\mathcal{R}) = \int_{\mathcal{R}} \varphi(\mathbf{x}) d\mathbf{x},$$

donde φ es una función Lebesgue integrable en múltiples variables, definida en una región $\mathcal{R}$.

El concepto de integral de Lebesgue es una de las generalizaciones clásicas fundamentales en la evolución del concepto de integral, y las funciones Lebesgue integrables incluyen funciones acotadas y no acotadas que cumplen ciertas condiciones de teoría de la medida (ver `https://www.encyclopediaofmath.org/index.php/Lebesgue_integral` y

`http://en.wikipedia.org/wiki/Lebesgue_integral` por más información).

En algunos casos (muy particulares), es posible evaluar $\zeta(\mathcal{R}) = \int_{\mathcal{R}} \varphi(\mathbf{x}) d\mathbf{x}$ por métodos analíticos. En la enorme mayoría de las ocasiones, esto no es posible, y es necesario aplicar métodos numéricos. La evaluación de estas integrales es uno de los problemas clásicos del análisis numérico, que si bien posee ya una amplia literatura, sigue siendo objeto de investigación y de propuesta de nuevos métodos para clases particulares de funciones.

Entre los métodos determinísticos aplicables, se encuentran las fórmulas de cuadratura y el uso de secuencias equi-distribuidas. Como alternativa, tenemos la opción de emplear Monte Carlo. Cada alternativa tiene sus ventajas y sus limitaciones.

Por comodidad, en la discusión subsecuente supondremos que $\varphi(\mathbf{x}) = 0$ para todo $\mathbf{x} \in \mathcal{J}^m \backslash \mathcal{R}$, esto hace que los valores de las integrales sobre $\mathcal{R}$ y sobre $\mathcal{J}^m$ coincidan.

Fórmulas de cuadratura

Las fórmulas de cuadratura más frecuentemente empleadas en el caso multi-dimensional son extensiones directas de las fórmulas de cuadratura para el caso uni-dimensional.

Una fórmula de cuadratura consiste en una secuencia de pares $\{(w_i^{(j)}, x_i^{(j)}) : j = 1, \ldots, n_i; i = 1, \ldots, m\}$ donde $\{0 \leq x_i^{(j)} < x_i^{(j+1)} \leq 1; j = 1, \ldots, n\}$ es la secuencia de puntos de evaluación en dimensión i, y los $w_i^{(j)}$ son los pesos asociados. La aproximación utilizando dicha fórmula (conocida como producto cartesiano) se calcula como una suma ponderada de un conjunto de evaluaciones de la función a integrar:

$$\bar{\zeta}(\mathcal{R}) = \sum_{j_1=1}^{n_1} \cdots \sum_{j_m=1}^{n_m} w_1^{(j_1)} \cdots w_m^{(j_m)} \varphi(x_1^{(j_1)}, \cdots, x_m^{(j_m)}).$$

Si tomamos $n_1 = \ldots = n_m = n^{1/m}$, entonces esta expresión requiere n

evaluaciones de φ.

Respecto al error cometido, si, para algún k, $\partial^k\varphi/\partial x_1^k, \ldots, \partial^k\varphi/\partial x_m^k$ existen y son acotadas en $\mathcal{J}^m$, entonces existen fórmulas de cuadratura tales que

$$|\bar{\zeta}(\mathcal{R}) - \zeta(\mathcal{R})| = O(n^{k/m}).$$

Además, es posible demostrar que la tasa $n^{k/m}$ es la mejor posible para fórmulas de cuadratura.

Secuencias equidistribuidas

La malla m-dimensional (que vimos en la sesión 3) pertenece a la familia de secuencias equidistribuidas, que pueden ser utilizadas para definir los puntos de evaluación sobre los cuales derivar una aproximación del volumen $\lambda(\mathcal{R})$. Una secuencia $\mathcal{X} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots\}$ es una secuencia equidistribuida en $\mathcal{J}^m$ si, para $S(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}) = \sum_{i=1}^{n} \phi(\mathbf{x}^{(i)})$, se cumple que

$$\lim_{n\to\infty} |S(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})/n - \lambda(\mathcal{R})| = 0$$

para todo $\mathcal{R} \subseteq \mathcal{J}^m$.

Si φ es integrable Riemann en $\mathcal{R}$ (una condición más fuerte que la de integrable Lebesgue, por más información ver `https://www.encyclopediaofmath.org/index.php/Riemann_integral` y `http://en.wikipedia.org/wiki/Riemann_integral`), entonces para

toda secuencia equidistribuida $\mathcal{X} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \in \mathcal{J}^m$ se cumple que

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varphi(\mathbf{x}^{(i)}) = \zeta(\mathcal{R}).$$

La clase de funciones integrables Riemann no admite funciones no acotadas; sólo funciones acotadas continuas, o discontinuas en las cuales la medida de Lebesgue del conjunto de los puntos de discontinuidad es 0.

Para poder evaluar el error, es necesario contar con restricciones adicionales sobre φ.

Si $\partial^m \varphi / \partial x_1 \ldots \partial x_m$ y todas las derivadas parciales "previas" son continuas y acotadas en $\mathcal{J}^m$, entonces existen secuencias equidistribuidas en $\mathcal{J}^m$ que garantizan una convergencia del error de orden $O(n^{-1}(\ln n)^m)$.

Estimación por Monte Carlo

Sea $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(n)}$ una secuencia de vectores aleatorios uniformemente distribuidos en $\mathcal{J}^m$. Entonces

$$\bar{\zeta}_n(\mathcal{R}) = \frac{1}{n} \sum_{i=1}^{n} \varphi(\mathbf{X}^{(i)})$$

es un estimador insesgado de $\zeta(\mathcal{R})$, con error estándar $\sqrt{(\int_{\mathcal{R}} \varphi^2(\mathbf{x}) d\mathbf{x} - \zeta^2(\mathcal{R}))/n}$. Siempre que $\int_{\mathcal{R}} \varphi^2(\mathbf{x}) d\mathbf{x} < \infty$, el error es de $O(n^{-1/2})$.

Las fórmulas de puntos equidistribuidos, así como las de cuadratura cuando $k > m/2$ poseen un error de orden más pequeño, por lo que cuando son aplicables, resultan preferibles. Sin embargo, es justamente esta aplicabilidad la que no está garantizada o no es fácil de demostrar en una cantidad muy importante de casos, especialmente cuando m crece, y que hace que el método de Monte Carlo sea el preferido en esas instancias,

ya que es aplicable para funciones acotadas y sin acotar, con la única restricción de que $\int_{\mathcal{R}} \varphi^2(\mathbf{x}) d\mathbf{x} < \infty$ (en lugar de condiciones en derivadas de mayor orden).

El siguiente seudocódigo presenta el esquema más básico del método Monte Carlo aplicado a la integración de una función.

Procedimiento IntegraciónMonteCarlo (función φ, entero n, real $1-\delta$)
Entrada: función a integrar φ, n tamaño de la muestra, $1-\delta$ nivel de confianza
Salida: $\bar{\zeta}_n$ estimación de la integral, $V[\bar{\zeta}_n]$ estimación de la varianza

1. $S = 0$; $T = 0$;/* Inicialización */
2. For $j = 1, \ldots, n$ do
   2.1 Sortear $\mathbf{X}^{(j)}$ con distribución uniforme en $\mathcal{J}^m$;
   2.2 If $j > 1$ then $T = T + (1 - 1/j)\left(\varphi(\mathbf{X}^{(j)}) - S/(j-1)\right)^2$;
   2.3 $S = S + \varphi(\mathbf{X}^{(j)})$. /* Acumular en S y T*/
3. $\bar{\zeta}_n = S/n$; /*Estimador puntual de $\zeta(\mathcal{R})$*/
4. $\hat{\sigma}_n^2 = T/(n-1)$; /*Estimador puntual de la varianza de $\varphi(\mathbf{X}^{(j)})$/
5. $V[\bar{\zeta}_n] = \hat{\sigma}_n^2/n$; /*Estimador puntual de la varianza de $\bar{\zeta}_n$*/
5. Calcular $[I_1(S, n, \delta), I_2(S, n, \delta)]$ /*un intervalo de confianza de nivel $(1-\delta)$ para $\zeta(\mathcal{R})$ */

Este procedimiento tiene el mismo esquema general que en el caso del cálculo de un volumen, con algunos cambios puntuales.

Hacemos notar por un lado que se acumula el valor de la función φ en cada punto sorteado; recordamos que si queremos integrar sobre $\mathcal{R} \subset \mathcal{J}^m$, definimos $\varphi(\mathbf{x}) = 0$ para todo $\mathbf{x}$ que pertenezca al hipercubo pero no a $\mathcal{R}$.

Los puntos 2.3, 3 y 4 corresponden a una manera alternativa para calcular un estimador de la varianza. Las dos fórmulas clásicas son $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(\varphi(\mathbf{X}^{(j)}) - \bar{\zeta}_n)^2$, y $\hat{\sigma}_n^2 = \frac{1}{n-1}\left(\sum_{j=1}^{n}\varphi^2(\mathbf{X}^{(j)}) - n\bar{\zeta}_n^2\right)$. La primera tiene el inconveniente de que es necesario acumular todos los valores de $\varphi(\mathbf{X}^{(j)})$ hasta el final del programa para hacer el cálculo, la segunda tiene problemas numéricos, dado que acumula valores elevados al cuadrado y puede producirse pérdida de precisión. La fórmula recursiva mostrada en el seudocódigo es equivalente a estas dos, y es más robusta del punto de vista numérico (existe también una fórmula recursiva similar para la esperanza).

Lecturas adicionales:

- Demostración de la fórmula recursiva en el libro del curso, pag. 68; o en la enciclopedia MathWorld,entrada "Sample Variance Computation", página `http://mathworld.wolfram.com/SampleVarianceComputation.html` (último acceso: 2020-03-26).

- Fórmula de Welford (una manera alternativa de calcular la varianza a medida que se realiza el muestreo, sin almacenar todos los valores generados), discutida en el blog de John D.Cook, página `http://www.johndcook.com/blog/standard_deviation/` (último acceso: 2020-03-26).

Cálculo de intervalo de confianza empleando la aproximación normal

Bajo un conjunto de condiciones bastante generales, que incluyen $\int_{\mathcal{R}} \varphi^4(\mathbf{x})d\mathbf{x} < \infty$, es posible demostrar que $(\bar{\zeta}_n - \zeta)/(\hat{\sigma}_n^2/n)^{1/2}$ converge en distribución a una v.a. normal $N(0,1)$ cuando $n \to \infty$. Esto es lo mismo que afirmar que

$$\lim_{n\to\infty} \text{Prob}\left(\frac{\bar{\zeta}_n - \zeta}{(\hat{\sigma}_n^2/n)^{1/2}} \leq \beta\right) = F_{N(0,1)}(\beta) = \Phi(\beta),$$

y por lo tanto si queremos un nivel de confianza de $1-\delta$, podemos construir el intervalo

$$\left(\bar{\zeta}_n - \Phi^{-1}(1-\delta/2)(\hat{\sigma}_n^2/n)^{1/2}, \bar{\zeta}_n + \Phi^{-1}(1-\delta/2)(\hat{\sigma}_n^2/n)^{1/2}\right),$$

que es un intervalo de confianza asintóticamente válido (cuando $n \to \infty$).

Es muy común usar este intervalo en la práctica, aunque se mantienen los comentarios realizados en el caso de la estimación de volúmenes: dificultad para conocer la tasa de convergencia a la distribución normal, que puede no ser uniforme en ζ; error introducido al emplear $\hat{\sigma}_n^2$ en lugar de σ^2 (desconocido); existencia de correlaciones entre las estimaciones de la media y de la varianza.

Para salvar estos problemas se requiere información o condiciones adicionales sobre la función a integrar; cuando estas no se conocen o están disponibles, la aproximación normal sigue siendo la herramienta básica para evaluar el error, aunque es necesario tener en cuenta que suele ser optimista (resultando en un intervalo de confianza que no es siempre suficientemente ancho para que la cobertura efectiva alcance la nominal).

Determinación del número de replicaciones empleando la aproximación normal

Cuando se tiene una especificación de error (ϵ, δ) predeterminada, y se desea calcular el número de replicaciones para alcanzar este error, se debe proceder en estas etapas:

1. realizar un conjunto de n' pruebas preliminares, para estimar la varianza (estimada por $\dot{\sigma}^2_{n'}$);

2. calcular el tamaño de muestra requerido de acuerdo a la aproximación normal, que es $\dot{n}_N(\epsilon, \delta) = \lceil \left(\Phi^{-1}(1-\delta/2)\right)^2 \dot{\sigma}^2_{n'}/\epsilon^2 \rceil$,

3. realizar un conjunto de $N \geq n_N(\epsilon, \delta)$ nuevas pruebas, con semillas diferentes (para asegurar la independencia entre la estimación del valor de N y los experimentos en sí).

Estimación de integrales de Lebesgue-Stieltjes

Si bien toda la presentación la hemos hecho para integrales de Lebesgue, definidas en un dominio acotado (supuesto por simplicidad incluido en $\mathcal{J}^m$), el método de Monte Carlo puede emplearse en un cuadro más general, a través de la observación que cuando existe la integral de Lebesgue $\zeta(\mathcal{R}) = \int_{\mathcal{R}} \varphi(\mathbf{x}) d\mathbf{x}$, su valor coincide con el de una integral de Lebesgue-Stieltjes

$$\zeta = \int_{\mathcal{Z}} \kappa(\mathbf{z}) dF(\mathbf{z}),$$

donde $\mathcal{Z}$ es una región incluida en R^m (eventualmente no acotada), $\kappa()$ es una función medible (eventualmente no acotada) en $\mathcal{Z}$, y $F()$ es una distribución en los conjuntos medibles de $\mathcal{Z}$ (es decir, $0 \leq F(\mathcal{A}) \leq F(\mathcal{B}) \leq 1$ para todo $\mathcal{A}$, $\mathcal{B}$ conjuntos medibles tales que $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{Z}$), cuyas formas están determinadas por $\mathcal{R}$ y φ. (ver `https://www.encyclopediaofmath.org/index.php/Lebesgue-Stieltjes_integral`,`https://www.encyclopediaofmath.org/index.php/Stieltjes_integral`,`http://www.math.utah.edu/~li/L-S%20integral.pdf` - ver sección 6.1)

Al ser F una función de distribución, ζ puede interpretarse como la esperanza de $\kappa(\mathbf{Z})$, donde $\mathbf{Z}$ es un vector aleatorio de dimensión m y distribución F.

A través de esta propiedad, vemos que es posible emplear el esquema de Monte Carlo realizando n muestras independientes $\mathbf{Z}^{(i)}$ con distribución F, y tomando $\ddot{\zeta}_n = n^{-1} \sum_{i=1}^{n} \kappa(\mathbf{Z}^{(i)})$ como estimador de ζ.

Dependiendo de las varianzas de $\bar{\zeta}_n$ y $\ddot{\zeta}_n$, puede ser más ventajoso emplear uno u otro de los dos esquemas.

El siguiente seudocódigo corresponde al método Monte Carlo aplicado para calcular una integral de Lebesgue-Stieltjes.

Procedimiento MonteCarlo-Lebesgue-Stieljes (κ, dF, n, $1-\delta$)
Entrada: función a integrar κ, función densidad de probabilidad dF, n tamaño de la muestra, $1-\delta$ nivel de confianza
Salida: $\ddot{\zeta}_n$ estimación de la integral, $V[\ddot{\zeta}_n]$ estimación de la varianza

1. $S = 0$; $T = 0$;/* Inicialización */
2. For $j = 1, \ldots, n$ do
   2.1 Sortear $\mathbf{Z}^{(j)}$ con densidad de probabilidad dF;
   2.2 If $j > 1$ then $T = T + (1 - 1/j)\left(\kappa(\mathbf{Z}^{(j)}) - S/(j-1)\right)^2$;
   2.3 $S = S + \kappa(\mathbf{Z}^{(j)})$. /* Acumular en S y T*/
3. $\ddot{\zeta}_n = S/n$; /*Estimador puntual de $\zeta(\mathcal{R})$*/
4. $\hat{\sigma}_n^2 = T/(n-1)$; /*Estimador puntual de la varianza de $\kappa(\mathbf{Z}^{(j)})$/
5. $V[\ddot{\zeta}_n] = \hat{\sigma}_n^2/n$; /*Estimador puntual de la varianza de $\ddot{\zeta}_n$*/
5. Calcular $[I_1(S, n, \delta), I_2(S, n, \delta)]$ /*un intervalo de confianza de nivel $(1-\delta)$ para $\zeta(\mathcal{R})$ */

Material adicional de lectura

- Integración por Monte Carlo en Mathworld (de Wolfram Research): `http://mathworld.wolfram.com/MonteCarloIntegration.html`. (último acceso: 2020-03-26)

- Integración por Monte Carlo en Wikipedia: `http://en.wikipedia.org/wiki/Monte_Carlo_Integration`.(último acceso: 2019-03-26)

- Módulo 12 del tema Numerical Integration de las notas de curso "Numerical Analysis - Numerical Methods", por John H. Mathews, disponible en `http://web.archive.org/web/20140307021638/http://mathfaculty.fullerton.edu/mathews//n2003/MonteCarloMod.html`, (último acceso 2020-03-26).

Preguntas para auto-estudio

- ¿Que métodos determinísticos para evaluar integrales de Lebesgue multi-dimensionales conoce? ¿Qué precisión alcanzan? (en términos de orden en función del número de puntos y la dimensión del problema).

- ¿Cuál es el seudocódigo de un Método Monte Carlo para estimar el valor de una integral de Lebesgue multi-dimensional? ¿Cuál es el orden del error que alcanza este método?

- ¿En qué casos es preferible un método determinístico a uno Monte Carlo?

- ¿Cómo es posible determinar el número de replicaciones a realizar para alcanzar una precisión prefijada?

- ¿Cómo es posible generalizar el uso del método Monte Carlo para calcular una integral de Lebesgue-Stieltjes en un dominio no acotado?

Ejercicios - Entrega 3

Ejercicio 6.1: [individual]

Problema: se idealiza una montaña como un cono inscrito en una región cuadrada de lado 1 km. La base de la montaña es circular, con centro en $(0.5, 0.5)$ y radio $r = 0.4$km, y la altura es $H = 8$km. La altura de cada punto (x, y) de la montaña está dada por la función $f(x, y) = H - H/r \times \sqrt{(x - 0.5)^2 + (y - 0.5)^2}$, en la zona definida por el círculo, y 0 fuera del círculo. El volumen total de la montaña (en km cúbicos) puede verse como la integral de la función altura en la región.

- Parte a: escribir un programa para calcular el volúmen por Monte Carlo. Realizar 10^6 replicaciones y estimar el valor de ζ y el error cometido (con nivel de confianza 0.95), utilizando como criterio la aproximación normal.

- Parte b: en base al valor estimado en la parte a, calcular el número de

replicaciones necesario para obtener un error absoluto menor a 10^{-3} (con nivel de confianza 0.95).

- Parte c: realizar esa cantidad de replicaciones y estimar ζ y su intervalo de confianza.

Comentario - en un sólido regular como el del ejercicio, este volumen podría calcularse también analíticamente. En el caso de una montaña real, la altura para cada punto no va a surgir de una fórmula, sino que se debe usar información de altura procedente de mediciones reales, por ejemplo de información satelital.

Ejercicio 6.2: [en grupo]

Problema: se desea estimar la integral de la función $x_1 x_2^2 x_3^3 x_4^4 x_5^5$ sobre el hipercubo $\mathcal{J}^m$ de dimensión $m = 5$.

- Parte a: revisar los códigos preparados para el ejercicio 6.1, elegir uno de ellos como punto de partida. Sobre esa base, modificarlo para

realizar cálculo por Monte Carlo de la integral planteada en el ejercicio 6.2. realizar 10^6 replicaciones y estimar el valor de ζ. Calcular analíticamente el valor exacto de la integral.

- Parte b: en base al valor estimado en la parte a, calcular el número de replicaciones necesario para obtener un error menor a 10^{-4} (con nivel de confianza 0.95).

- Parte c: Decimos que un intervalo de confianza cubre el valor exacto cuando este último pertenece al intervalo.

  Realizar $L = 500$ experimentos con semillas diferentes, cada uno consistente en estimar por Monte Carlo con el nro. de replicaciones de la parte b el valor de la integral, así como intervalos de confianza de nivel 0.9, 0.95 y 0.99. Para cada nivel de confianza, calcular el nivel de cobertura empírico (en que porcentaje de los 500 experimentos el intervalo de confianza calculado cubrió el valor exacto).

  Discutir los resultados, comparando la cobertura empírica con la especificada.

Fecha entrega: Ver cronograma y avance del curso.