1. Implementacion del algoritmo Adaptive Rejection Sampling (ARS) y simulación de una $Gama(2,1)$ 10,000 muestras.

   Teoría

   El algoritmo Adaptive Rejection Sampling (ARS) fue propuesto por W.R Gilks en 1992 en Bayesian Statistics 4 Bernardo, Berger en la cuarta reunion Bayesiana en Valencia, España en su articulo "Derivative Free Adaptive Rejection Sampling for Gibbs Sampling". Gilks propone en 1992 una versión del ARS que incluye el uso de derivadas. Si bien el primer articulo quería hacer énfasis en Gibbs sampling, el método puede emplearse para poder extraer muestras de una densidad cualquiera, con la condición que esta sea log concava. Comenzaremos con una breve explicación teórica del algoritmo.

   Sea f una densidad log concava en $\mathbb{R}$. El ARS es un algoritmo que genera muestras i.i.d de la densidad f. El algoritmo utiliza la log concavidad para poder construir envolventes superiores e inferiores para la densidad objetivo (basado en un conjunto de puntos iniciales especificados). Se puede utilizar la envolvente superior como densidad propuesta para ejecutar un paso del algoritmo de aceptación rechazo. Mas aún, podemos actualizar las envolventes utilizando la muestra seleccionada, tomando en cuenta el espacio que existe entre la envolvente inferior y la función.

   Para $u < v$ sea $h = \log f$ y $\mathbb{L}_{uv}$ (es decir $y = \mathbb{L}_{uv}$) la recta que pasa por $(u, h(u)$ y $(v, h(v))$. Por tanto para $x \in (u, v)$ la curva h esta por sobre la recta $\mathbb{L}_{uv}$ y para los $x \notin (u, v)$ la curva esta por bajo de $\mathbb{L}_{uv}$. Sea $\mathbb{S} = \{x \in \mathbb{R} : f(x) > 0\}$ y consideremos el subconjunto de $\mathbb{S}$ dado por

   $$\{a = x_1 < x_2 < ... < x_M = b\}.$$

   Estos serán los puntos iniciales del algoritmo. Suponga que $a, b \in \mathbb{S}$ son tales que a es menor que el percentil uno de f y b es al menos el percentil 99 de f. Sea $y = \mathbb{L}_{i,i+1}(x)$ la recta que pasa por los puntos

   $$(x_i, h(x_i)), (x_{i+1}, h(x_{i+1}))$$

   Entonces para $x_i \leq x \leq x_{i+1}, 2 \leq i \leq M-2$ se tiene

   $$\mathbb{L}_{i,i+1}(x) \leq h(x) \leq \min\{\mathbb{L}_{i-1,i}(x), \mathbb{L}_{i+1,i+2}(x)\}.$$

   Sea $\mathbb{L}_{0,1}(x) = \mathbb{L}_{2,3}(x)$ y $\mathbb{L}_{M,M+1} = \mathbb{L}_{M-2,M-1}(x)$, entonces la desigualdad anterior es valida para $x_i \leq x \leq x_{i+1}, 1 \leq i \leq M-1$.

   Ahora para $x \leq x_1$ tenemos

   $$h(x) \leq \mathbb{L}_{1,2}(x)$$

   y para $x \geq x_M$ tenemos

   $$h(x) \leq \mathbb{L}_{M-1,M}(x).$$

   Por lo tanto definiendo para $x_i \leq x_{i+1}, 1 \leq i \leq M-1$

   $$\begin{aligned} g(x) &= \exp\{\min\{\mathbb{L}_{i-1,i}(x), \mathbb{L}_{i+1,i+2}(x)\} \\ i(x) &= \exp\{\mathbb{L}_{i,i+1}(x)\}, \end{aligned}$$

   para $x \leq x_1$

$$g(x) = \exp\{\mathbb{L}_{1,2}(x)\}$$
$$i(x) = 0,$$

y para $x \geq x_M$

$$g(x) = \exp\{\mathbb{L}_{M-1,M}(x)\}$$
$$i(x) = 0,$$

tenemos

$$i(x) \leq f(x) \leq g(x) \forall x.$$

Suponemos que $\{x_i : 1 \leq i \leq M\}$ (o en realidad x_1, x_2, x_{M-1}, x_M) han sido seleccionados de tal forma que la pendiente de $\mathbb{L}_{1,2}$ es positiva y la pendiente de $\mathbb{L}_{M-1,M}$ es negativa. Asi, la funcion $g(x)$ es integrable.

Sea $K = \int_{\mathbb{R}} g(x)dx$ y sea $\tilde{g} = \frac{1}{K} g(x)$, entonces $\tilde{g}$ es una densidad y por elección, $\tilde{g}$ es una envolvente para h. Mas aun podemos de forma adaptativa agregar puntos a la partición inicial para poder mejorar la envolvente. No es complicado generar puntos de la envolvente $\tilde{g}$, para esto notemos que si (w_i, z_i) son los puntos de intersección de las rectas $y = \mathbb{L}_{i-1,i}(x)$ e $y = \mathbb{L}_{i+1,i+2}(x)$ para $2 \leq i \leq M-2$ y escribiendo $y_i = h(x_i)$ podemos ver que $\log \tilde{g}(x)$ es una función lineal a trozos obtenida al unir

$$(x_2, y_2), (w_2, z_2), (x_3, y_3), (w_4, z_4), ..., (x_{M-2}, y_{M-2}).$$

y para $x \leq x_2$ es la recta $y = \mathbb{L}_{1,2}(x)$ mientras que para $x \geq x_M$ es la recta $y = \mathbb{L}_{M-1,M}(x)$. Notemos de esto que $\tilde{g}$ es una mezcla de funciones.

De lo anterior se tiene que para poder utilizar la envolvente es necesario poder extraer muestras de la mezcla que tenemos. Tenemos cuatro casos de interés. Para cada caso mostraremos la función de densidad $g(x)$ a ser utilizada, su constante normalizadora (pues en principio no son densidades), la inversa de su CDF. Estos los elementos necesarios para poder llevar a cabo el algoritmo de aceptación rechazo. Las constantes normalizadoras serán utilizadas para asignar los pesos de la mezcla.Si bién los cálculos no son complicados, es interesante notar que no se presentan en los artículos de Gilks o en el libro de Robert Casella. Estos resultados son necesarios para poder llevar a cabo la implementación del algoritmo.

Sean a_i, b_i el intercepto en y y la pendiente de la recta $\mathbb{L}_{i,i+1}(x)$ respectivamente, es decir $y = b_i x + a_i$. Sea c la constante normalizadora de la densidad presentada $g(x)$, sea $G^{-1}(u)$ la inversa de su cdf en donde $u \sim U(0,1)$ y sea $\mathbb{I}_A$ la función indicadora de A. Tenemos

(a) Para $x \leq x_1$ se tiene $g(x) = \exp\{\mathbb{L}_{1,2}(x)\}$ por lo que

$$c = \frac{e^{a_1}}{b_1}(e^{b_1 x_1} - 1)$$
$$g(x) = \frac{1}{c}\exp\{b_1 x + a_1\}\mathbb{I}_{0 \leq x \leq x_1}$$
$$G^{-1}(u) = \frac{1}{b_1}\ln\left(cb_1 e^{-a_1} u + 1\right),$$

en donde se seleccionan los puntos de tal forma que $b_1 > 0$.

(b) Para $x_1 \leq x \leq x_2$ se tiene $g(x) = \mathbb{L}_{2,3}$ por lo que

$$\begin{aligned} c &= \frac{e^{a_2}}{b_2}(e^{b_2 x_2} - e^{b_2 x_1}) \\ g(x) &= \frac{1}{c}\exp\{b_2 x + a_2\}\mathbb{I}_{x_1 \leq x \leq x_2} \\ G^{-1}(u) &= \frac{1}{b_2}\ln\left(cb_2 e^{-a_2}u + e^{b_2 x_1}\right), \end{aligned}$$

en donde se seleccionan los puntos de tal forma que $b_2 > 0$.

(c) Para $x_i \leq x \leq x_{i+1} 2 \leq i \leq M-2$ se tiene $g(x) = \exp\{\min\{\mathbb{L}_{i-1,i}(x), \mathbb{L}_{i+1,i+2}(x)\}$ por lo que $g(x)$ es una función seccionada en su dominio. $g(x)$ puede escribirse como

$$g(x) = \begin{cases} \frac{1}{c}\exp\{b_{i-1}x + a_{i-1}\} & x_i \leq x \leq k \\ \frac{1}{c}\exp\{b_{i+1}x + a_{i+1}\} & k \leq x \leq x_{i+1}, \end{cases}$$

en donde se tiene que

$$\begin{aligned} k &= \frac{a_{i-1} - a_{i+1}}{b_{i+1} - b_{i-1}} \\ c &= \frac{e^{a_{i-1}}}{b_{i-1}}\left(e^{b_{i-1}k} - e^{b_{i-1}x_i}\right) + \frac{e^{a_i+1}}{b_{i+1}}\left(e^{b_{i+1}x_{i+1}} - e^{b_{i+1}k}\right). \end{aligned}$$

Así como $g(x)$ es una función seccionada, también lo es $G^{-1}(u)$. Sea $u^* = G(k)$, entonces se tiene que

$$G^{-1}(u) = \begin{cases} \frac{1}{b_{i-1}}\ln\left(cb_{i-1}e^{-a_{i-1}}u + e^{b_{i-1}x_i}\right) & u \leq u^* \\ \frac{1}{b_{i+1}}\ln\left(cb_{i+1}e^{-a_{i+1}}u - \frac{b_{i+1}}{b_{i-1}}e^{a_{i-1}-a_{i+1}}\left[e^{b_{i-1}k} - e^{b_{i-1}x_i}\right] + e^{b_{i+1}k}\right) & u^* < u \leq 1, \end{cases}$$

en donde se seleccionan los puntos de tal forma que $b_{i-1} > 0$ y $b_{i-1} < 0$.

(d) Para $x_{M-1} \leq x \leq x_M$ se tiene $g(x) = \mathbb{L}_{M-2,M-1}$ por lo que

$$\begin{aligned} c &= \frac{e^{a_{M-2}}}{b_{M-2}}(e^{b_{M-2}x_M} - e^{b_{M-2}x_{M-1}}) \\ g(x) &= \frac{1}{c}\exp\{b_{M-2}x + a_{M-2}\}\mathbb{I}_{x_{M-1} \leq x \leq x_M} \\ G^{-1}(u) &= \frac{1}{b_{M-2}}\ln\left(cb_{M-2}e^{-a_{M-2}}u + e^{b_{M-2}x_{M-1}}\right), \end{aligned}$$

en donde se seleccionan los puntos de tal forma que $b_{M-2} < 0$.

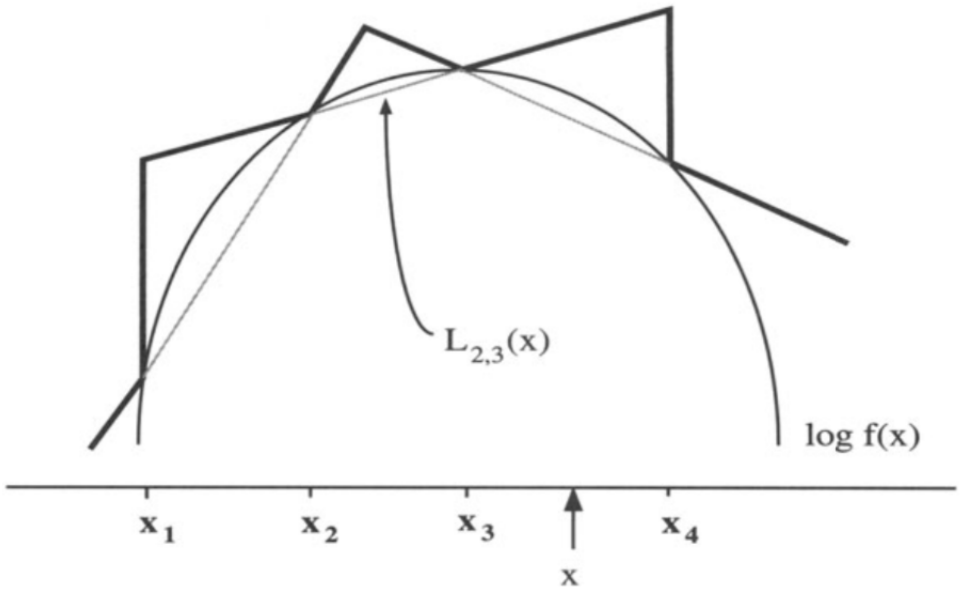
(e) Para $x \geq x_M$ se tiene $g(x) = \exp\{\mathbb{L}_{M-1,M}(x)\}$ por lo que

$$\begin{aligned} c &= \frac{-e^{a_{M-1}+b_{M-1}x_M}}{b_{m-1}} \\ g(x) &= \frac{1}{c}\exp\{b_{m-1}x + a_{m-1}\}\mathbb{I}_{x \geq x_M} \\ G^{-1}(u) &= \frac{1}{b_{M-1}}\log\left(cb_{M-1}e^{-a_{M-1}}u + e^{b_{M-1}x_M}\right), \end{aligned}$$

en donde se seleccionan los puntos de tal forma que $b_{M-1} < 0$.

Por los requerimientos anteriores se tiene que el algoritmo debe comenzar al menos con 4 puntos. Para poder asegurar los requerimientos de la selección de los puntos debemos seleccionar los primeros 2 puntos a la izquierda de la moda y los restantes a la derecha. El cálculo de la moda no es en realidad necesario, basta con asegurarnos del signo de las pendientes al seleccionar los puntos.

Gráficamente la envolvente toma la siguiente forma



Es por esta razón que debemos considerar los distintos casos. El caso c es en el que se forman los picos, las intersecciones de las rectas. Algo sumamente importante es que aquí solo estaremos considerando la envolvente superior, pues es lo que se discutió en el curso. Aun así, la envolvente inferior tiene un procedimiento similar.

Implementación:

Suponga que se tienen $M \geq 4$ puntos $x_1, .., x_M$ entonces tenemos $M-1$ intervalos en donde el primer intervalo corresponde al caso b, el ultimo intervalo corresponde al caso d y el resto de los intervalos corresponde al caso c. Teniendo esto en mente calculamos las constantes de normalización $c_i, i = 1, .., M-1$ y asignamos un peso w_i a cada intervalo (por tanto a su parte correspondiente de la envolvente). Antes de hacer esto debemos de tomar en cuenta que los puntos $0 \leq x \leq x_1$, $x_M \geq x$ también una parte de la envolvente y por tanto calculamos sus constantes de normalización. Es decir, que si bien tenemos $M-1$ intervalos, debemos considerar $M+1$ constantes de normalización así como densidades y por tanto pesos. Teniendo esto en cuenta tenemos $c_k, k = 1, ..., M+1$ y los pesos w_k dados por

$$w_k = \frac{c_k}{\sum_{k=1}^{M+1} c_k}, k = 1, ..., M+1.$$

Una vez que tenemos los pesos, podemos generar un valor de acuerdo a una distribución discreta que tiene como imagen $\{1, ..., M+1\}$ con pesos $\{w_k\}_{k=1}^{M+1}$. Para esto hicimos uso de la funcion de SciPy

`scipy.stats.rv_discrete`

Dado que $X \sim Gamma(2,1)$ se tiene que $X \geq 0$. Es por esto que el caso a tiene como limite inferior de la densidad el valor de 0. Esto podría modificarse para el caso de otra densidad. Si $X \sim Gamma(2,1)$ su densidad $f_X(x)$ viene dada por

$$f_X(x) = xe^{-x}\mathbb{I}_{x>0}.$$

f en efecto es log concava pues

$$\begin{aligned} \ln f &= \ln x - x \\ \frac{d \ln f}{dx} &= \frac{1}{x} - 1 \\ \frac{d^2 \ln f}{dx^2} &= -\frac{1}{x^2} < 0, \forall x > 0. \end{aligned}$$

Se considera que el código viene comentado y refleja lo discutido en este documento. El código ha sido dividido en secciones:

- *Funciones Auxiliares*. Aquí se considera específicamente quien es $f(x)$ asi como las funciones que ayudan a encontrar $\{a_i\}, \{b_i\}, \{c_i\}, \{w_i\}$. Prestar atención importante a la función *NC*, la cual se le alimenta X vector de puntos y retorna todos las constantes normalizadoras.
- *Funciones de variables aleatorias*. Estas simulan los casos y consideran las densidades correspondientes.
- *Funciones del Free Derivative Envelope*. Se construye aqui la envolvente superior mediante *DFEnvelope*. Esta regresa la envolvente y, luego se utiliza para poder extraer muestras mediante *dfARS*.

Resultados y Comentarios:

La selección de los puntos iniciales tiene un cierto nivel de importancia. Suponemos que los puntos iniciales son tales que permiten que el algoritmo corra, es decir se seleccionan al menos 2 a la izquierda de la moda y al menos 2 a su derecha. Si bien uno quisiera que la envolvente se volviese cada vez mas independiente de estos a medida acepta nuevos puntos, puede darse que los puntos nuevos que se aceptan tenga que ver con los puntos iniciales seleccionados. Es decir que los puntos iniciales puede ser restrictivos a largo plazo. Tras varios experimentos consideramos que lo mejor es seleccionar cuantiles en especifico, pues son informativos sobre la masa de la distribución. Para nuestro caso hemos elegido los cuantiles $q_{0.01}, q_{0.25}, q_{0.50}, q_{.75}$. Los cuales para la $Gamma(2,1)$ son $\{0.1484, 0.9612, 1.6783, 2.69\}$. Notemos que el cuantil $q_{0.50} = 0.9612$ esta cerca de la moda $x = 1$.

En la siguiente tabla resumimos los resultados de algunos experimentos. La envolvente comienza con los cuantiles mencionados anteriormente y llega a incluir M puntos. Se presenta el p valor de la prueba de Kolmogorov Smirnov, el valor mínimo de la envolvente $x_{(1)}$ y el valor máximo de los puntos de la envolvente x_M.

M	Kolmogorov Smirnov p value	min	max
30	0.864	0.1485	4.68
50	0.648	0.1262	5.78
250	0.7842	0.08768	7.386
500	0.8032	0.086	7.367

Note que en todos los casos Kolmogorov Smirnov no rechaza la hipótesis de que los puntos $x_1, .., x_M$ provienen de una distribución $Gamma(2,1)$. Así mismo, la envolvente logra explorar valores superiores al cuantil 0.99 ($q_{0.99} = 6.638$) de la $Gamma(2,1)$.