Hamiltonian Monte Carlo

- Santiago Eliges, Juan Quintero

1. Introducción

a. Motivación:

En estimación bayesiana es común buscar inferir una distribución objetivo de algún parámetro de algún modelo. En nuestro caso vamos a asumir que esta distribución es suave, y derivable, por lo que implica la existencia de una función de distribución de densidad de probabilidad, la cual caracteriza la distribución objetivo. La estimación se realiza cubriendo el espacio paramétrico, ponderando nuestras creencias con la verosimilitud de los datos observados bajo dicho modelo.

Nuestra distribución objetivo se comportará acorde a la siguiente proporcionalidad.

$$\pi(q) = \pi(q|X) \propto L(q|X)\pi(q)$$

Esto puede causar como interés la búsqueda de la constante de normalización.

$$C^{-1} = \int_Q L(q|X)\pi(q)dq$$

Esta constante coincide con otra cosa que suele ser de interés, la búsqueda de esperanzas de funciones f sobre espacios paramétricos Q

$$E_\pi(f) = \int_Q f(q)\pi(q)dq$$

donde $\pi(q)$ es la distribución de densidad del espacio paramétrico Q. Es importante notar que la esperanza es una integral que se mantendrá invariante por parametrización del espacio, es decir; Para cualquier parametrización Q' del espacio paramétrico se cumple lo siguiente.

$$E_\pi(f) = \int_Q f(q)\pi(q)dq = \int_{Q'} f(q')\pi(q')dq'$$

El cómputo de la esperanza en la práctica se realiza discretizando el espacio integrado computando una sumatoria sobre una grilla de valores del espacio evaluada en el término integrado. Esto conlleva un gran problema al aumentar la dimensión del espacio ya que el cómputo escalará exponencialmente con la dimensión. Por esta razón, una búsqueda de métodos eficientes para computar esperanzas puede recaer en intentar integrar únicamente

sobre los valores del espacio que representan valores significativos en el término integrado ($f(q)\pi(q)dq$), ignorando todos los términos despreciables de la integral.

b. Typical Set

En búsqueda de esta región relevante para nuestro cómputo, tiene sentido estudiar al término integrado $f(q)\pi(q)dq$ a fin de analizar para qué valores del espacio aporta información.

Vamos a analizar brevemente a los componentes del término por separado

- $f(q)$: afecta directamente al término integrado, sin embargo buscamos calcular la esperanza para distintas distribuciones $f(.)$ por lo que intentaremos encontrar un método independiente de este término.
- $\pi(q)$: Los valores q que tengan alta densidad serán relevantes al estudiar los términos informativos a la hora de calcular la esperanza. Notemos que al aislar $\pi(q)$ de la integral, la target distribution depende de la parametrización tomada. Esto será de especial importancia a la hora de entender la construcción del Hamiltonian Monte Carlo (HMC)
- dq: Este término referido al diferencial del volumen sobre el espacio integrado afecta directamente al valor de la integral. Este término puede interpretarse como la porción de volumen ponderada para ponderar el resto de términos. Debido a que el volumen incrementa exponencialmente al aumentar las dimensiones de nuestro espacio paramétrico

Notemos que mientras aumentan las dimensiones, el volumen toma más importancia en la integral a computar, por lo que suele tener un mayor efecto sobre la esperanza objetivo.

Esto crea una dinámica entre $\pi(q)$ y dq, ya que al buscar una región con valores altos de $\pi(q)$, tendremos que concentrar los valores q hacia la moda de π, achicando por consecuencia dq. Por el otro lado en consecuencia, el incremento del tamaño de la región para buscar volúmenes grandes causará que el promedio de $\pi(q)$ en dicha región baje (ver figura 1).

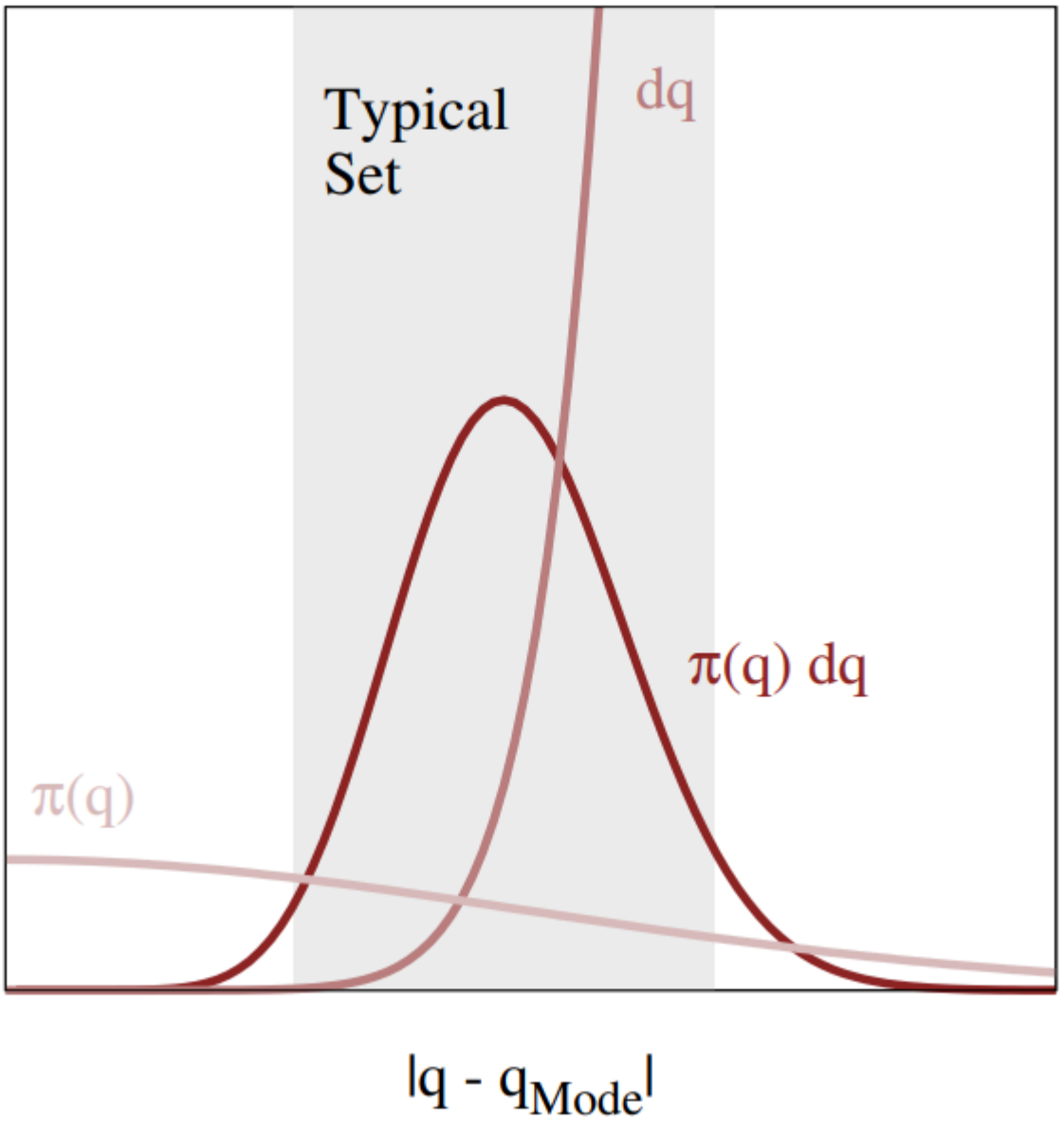


Figura 1. En rosa claro una ejemplificación de una target distrion $\pi(q)$, en rosa oscuro un ejemplo del crecimiento del volumen y en rojo el producto del volumen con la target distribution. Vemos que el producto de estas dos curvas genera una función con una moda que será informativa a la hora de calcular esperanzas. En gris se marca la zona relevante (el typical set)

Entonces, debe existir una zona de equilibrio, donde estas fuerzas contrarias contengan la mayoría de la información necesaria del cómputo. Dicha región del espacio de parámetros se llama el *typical set.* La diferencia entre el typical set y una región de alta densidad es complicada de visualizar, debido a que las visualizaciones se hacen en bajas dimensiones en las cuales el volumen del espacio no es lo suficientemente relevante, y por lo tanto, el typical set termina coincidiendo con una región de alta densidad.

c. Estimación con Cadenas De Markov

Sea $\{q_1, q_2, \ldots, q_N\}$ un sample generado por una cadena de markov representativo del espacio paramétrico Q , una forma de estimar la esperanza sobre el espacio es promediando la función objetivo sobre el sample de puntos.

$$\overline{f_N} = \frac{1}{N} \sum_{n=1}^{N} f(q_n).$$

Pero ya analizamos que no tiene sentido considerar puntos $\{q_1, q_2, \ldots, q_N\}$ que estén fuera del typical set, por lo que nuestro objetivo será samplear con cadenas de markov sobre puntos dentro del typical set ignorando los sectores relevantes del espacio paramétrico a fin de obtener buenas estimaciones de la esperanza en tiempos realistas.

Donde, sin detallar en la correctitud, sabemos que los estimadores de Monte Carlo -Cadenas de Markov (MCMC) convergen a la esperanza verdadera

$$\lim_{n \to \infty} \overline{f_N} = E_\pi(f)$$

Para realizar un *proceso de Markov*, es necesario definir una *transición de Markov.* Pensemos en la transición de Markov como una densidad condicional $T(q'|q)$ que define para un punto q, la probabilidad de transición a un punto q'.

En particular, es interesante tomar transiciones que preserven la target distribution, es decir, que si generamos un conjunto de muestras que se distribuyen como la función objetivo y aplicamos la transición, el conjunto transicionado se seguirá distribuyendo según la sample distribution.

$$\pi(q) = \int_Q \pi(q')T(q|q')dq'$$

Siempre que trabajemos con una transición de Markov que respete la preservación de la target distribution, sin importar el punto inicial que tomemos, la cadena de Markov se terminará concentrando alrededor del typical set. Detallaremos más el porqué esto resulta así, más adelante, cuando veamos una definición alternativa de preservar la distribución.

La elección de la transición $T(q|q')$ tendrá un rol fundamental en la correctitud del método para samplear el typical set y más aún, también será fundamental para determinar la velocidad de convergencia (y de exploración del método) obteniendo en menos tiempo buenos resultados.

d. Algoritmo de Metropolis-Hastings

Dada una transición de Markov que preserve la target distribution, MCMC asegura ser un método para obtener estimaciones asintóticas para la esperanza alrededor del typical set.

Sin embargo, la construcción de la transición de Markov no es para nada trivial y suele ser el fondo principal de estudio en los algoritmos de MCMC. El algoritmo MCMC más popular (e intuitivo) es el algoritmo de Metropolis-Hastings que consiste en dos pasos principales: la propuesta, y la aceptación.

En la propuesta, el algoritmo propone aleatoriamente cualquier valor cercano en probabilidad a un punto inicial para luego ser evaluado en la etapa de aceptación donde se determinará si se acepta la transición a dicho punto o se rechaza.

Dado un punto q fijo, nosotros buscamos un q' al cual transicionar. La probabilidad para cada punto del espacio paramétrico de ser propuesto como un nuevo punto al cual transicionar se llama *proposal distribution* $Q(q'|q)$. La proposal distribution suele ser una gaussiana centrada en q, con una varianza representando el tamaño del salto entre puntos.

$$Q(q'|q) = \mathrm{N}(q'|q, \Sigma)$$

Dado el mismo punto q, y la propuesta de un nuevo punto q', se define una variable binaria de aceptación de dicha propuesta, la cual está distribuido como una Bernoulli con parámetro $A(q'|q)$. La probabilidad de aceptación o *acceptance distribution* está definida en base a pesar la probabilidad de transición por la probabilidad del punto de llegada.

$$A(q'|q) = min(1, (\frac{\Theta(q|q')\pi(q')}{\Theta(q'|q)\pi(q)}))$$

Cómo se construye a la proposal density como una gaussiana que es simétrica bajo el cambio del punto inicial, $Q(q|q') = Q(q'|q)$.
Luego se puede simplificar la acceptance distribution a una función del ratio entre la distribución de densidad en el punto en el que estoy contra en el punto al que me quiero mover .

$$A(q'|q) = min(1, (\frac{\pi(q')}{\pi(q)})).$$

Con esto quedaría construido el ciclo básico del algoritmo de Metropolis Hastings, pero ahora nos encontramos en condiciones de definir una distribución más.

La *transition distribution* está definida como la distribución de los pasos conjuntos de propuesta de un nuevo punto, y de aceptación de dicho punto.

$$T(q'|q) = A(q'|q)Q(q|q')$$

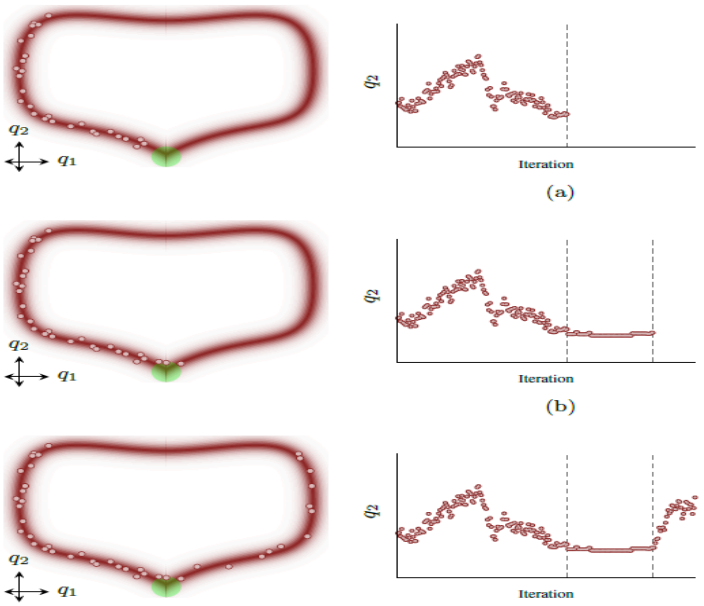
e. Problemas del Metropolis Hastings

Este método se presenta como una solución simple e intuitiva al problema original, sin embargo, el principal problema es que el tiempo de cómputo puede escalar muy mal al tener espacios paramétricos con grandes dimensiones. Esto se debe a que el aumento de dimensiones aumenta las direcciones posibles a las que ir a partir de un punto inicial dado

que, recordando que los puntos tienden a distribuirse cerca del typical set, hará cada vez menos probable la propuesta y aceptación de nuevos puntos.

Por último existen problemas cuando el step que se utiliza es muy grande o muy chico. En el caso de que el step sea muy chico, la autocorrelación de las cadenas sampleadas se vuelve muy alta, pero por otro lado se presentan problemas con muchos rechazos de nuevos puntos si el step es muy grande.

Por último, es difícil que este algoritmo se ajuste correctamente en regiones del typical set que presenten una fuerte curvatura. Aun así, debido a que la transición mantiene la target distribution, el algoritmo debe en el infinito aproximar la esperanza de la función. Lo que termina pasando es que la cadena termina pasando mucho tiempo resampleado de la región problemática, y termina compensando, generando fuertes oscilaciones que introducen un gran bias si se llega a cortar el algoritmo en cualquier tiempo finito. En rojo se puede apreciar el typical set, en verde la región que genera problemas



Por estos motivo, surge un especial interés en el estudio de la geometría del typical set a la hora de proponer rutas de transición a partir de un punto inicial que aproveche la topología del typical set con el fin evitar samples poco informativos y explorar el typical set de forma más eficiente.

2. Hamiltoniano y Propiedades

Ya se vio previamente que en altas dimensiones el algoritmo icónico de los métodos MCMC (Metropolis-Hasting) puede no ser tan bueno, eficiente y fidedigno a nuestra distribución. Mejorar el método recaerá esencialmente en la optimización de las direcciones propuestas.

Siguiendo esta línea de razonamiento, quisiéramos encontrar una forma de proponer direcciones que se encuentren dentro del typical set. Una buena idea para esto puede ser la

construcción de un espacio vectorial que para cada posición en el espacio paramétrico indique una dirección hacia el typical set.

Pero la construcción de este espacio vectorial no es para nada trivial, en principio se podría pensar en dirigirse en la dirección dada por el gradiente de la target distribution $\pi(q)$, pero esto nos alejará del typical set llevando los samples hacia la moda de la distribución. Además, la target distribution depende de la parametrización perdiendo la invarianza del typical set.

Para contrarrestar la fuerza que colapsa las direcciones hacia la moda de la target distribution, preservando la invarianza del typical set, resulta de utilidad la construcción de un modelo análogo a la **dinámica Hamiltoniana**.

a. Dinámica Hamiltoniana

El Hamiltoniano, es un funcional que busca describir un sistema físico. La definición formal del Hamiltoniano está basada en una transformación del Lagrangiano, otro funcional que describe sistemas. Cómo definir el Lagrangiano -y en consecuencia el Hamiltoniano- no es tan trivial, llegando a ser en algunos casos conceptos abstractos que simplemente cumplen las propiedades de describir el sistema. Afortunadamente, en la mayoría de los casos, incluyendo el que estamos interesados, el Hamiltoniano puede representarse como la suma de las energías del sistema.

Por lo tanto el Hamiltoniano es un funcional que actúa sobre la información de nuestro sistema, los momentos y coordenadas canónicas (por lo general coinciden con las nociones de momento y coordenadas intuitivas). En el tipo de sistema que estamos interesados toda la energía del sistema se reduce a la suma de la energía potencial, representada por U, y la cinética, K.

$$H(p,\ q)\ = K(p)\ +\ U(q)$$

Para especificar más todavía, el tipo de sistemas en el cual estamos interesados, son los sistemas conservativos, lo que significa que el Hamiltoniano se mantiene constante a lo largo de la evolución de nuestro sistema. Un ejemplo sencillo de un sistema conservativo es el de arrojar una canica en un bowl en el cual no haya fricción ni por el aire, ni por el bowl mismo.

Las ecuaciones que rigen nuestro sistema, derivadas del Hamiltoniano, se construyen sobre un vector de posición q n-dimensional y un vector de momentos p n-dimensional, por lo que se terminan produciendo 2n ecuaciones.

$$\frac{dq_i}{dt} = \frac{\delta H}{\delta p_i}$$
$$\frac{dp_i}{dt} = -\frac{\delta H}{\delta q_i}$$

Estas ecuaciones son para cada $i = 1,\ldots, n$. Luego, en todo intervalo temporal se define un mapeo de estados siguiendo el espacio de fases dado por las ecuaciones.

b. Invarianza del Hamiltoniano

Veamos rápidamente que las ecuaciones que rigen el movimiento en un sistema Hamiltoniano, condicen con la propiedad de invarianza previamente mencionada en los sistemas conservativos. En otras palabras, dado un punto en (p, q), y energías K y U definidas, la suma es constante.

$$\frac{dH}{dt} = \sum_{i}^{d} \frac{\delta H}{\delta q_i} \cdot \frac{dq_i}{dt} + \frac{\delta H}{\delta p_i} \cdot \frac{dq_i}{dt}$$

$$\sum_{i=1}^{d} \frac{\delta H}{\delta q_i} \cdot \frac{\delta H}{\delta p_i} - \frac{\delta H}{\delta p_i} \cdot \frac{\delta H}{\delta q_i}$$

$$\sum_{i=1}^{d} (0)$$

$$\frac{dH}{dt} = 0$$

$$H = Cte$$

Esta propiedad se usará para definir la construcción de la probabilidad de transición.

c. Consecuencias del Teorema de Liouville

Como consecuencia del Teorema de Liouville, el volumen en el espacio $(p,\ q)$ se conserva, es decir, no hay compresiones ni divergencias en el espacio. En este caso alcanza con ver que la divergencia del hamiltoniano es nula.

$$Div(H(p,q)) = \sum_{i}^{d} \frac{\delta}{\delta q_i} \frac{dq_i}{dt} + \frac{\delta}{\delta p_i} \frac{dp_i}{dt}$$

$$\sum_{i}^{d} \frac{\delta}{\delta q_i} \frac{\delta H}{\delta p_i} - \frac{\delta}{\delta p_i} \frac{\delta H}{\delta q_i}$$

$$\sum_{i}^{d} \frac{\delta^2 H}{\delta q_i \delta p_i} - \frac{\delta^2 H}{\delta p_i \delta q_i}$$

$$\sum_{i=1}^{d} (0)$$

$$Div(H(p,q)) = 0$$

Esta propiedad se utiliza en HMC para no tener que usar el jacobiano a la hora de definir el volumen de la región de aceptación.

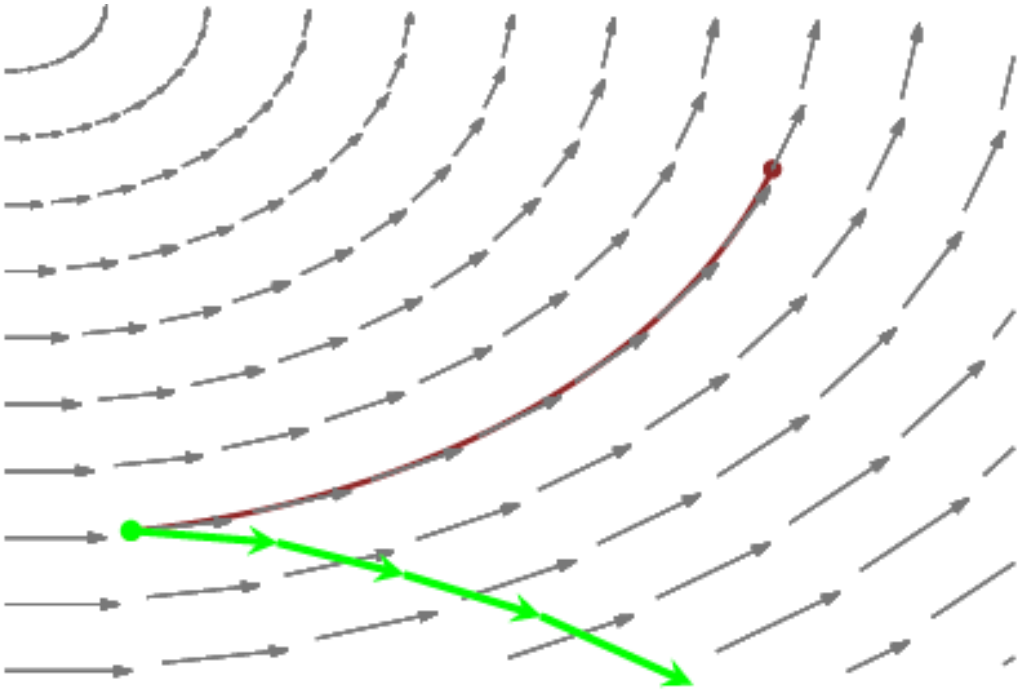
d. Reversibilidad

Las dinámicas Hamiltonianas son fácilmente reversibles negando las derivadas temporales de las ecuaciones diferenciales, o simplemente negando el momento. Esto se debe al aspecto conservativo del sistema

Esta propiedad puede utilizarse para ver que los valores de MCMC actualizados no modifican la distribución deseada, construyendo una Q proposal distribution que sea reversible.

e. Métodos Numéricos

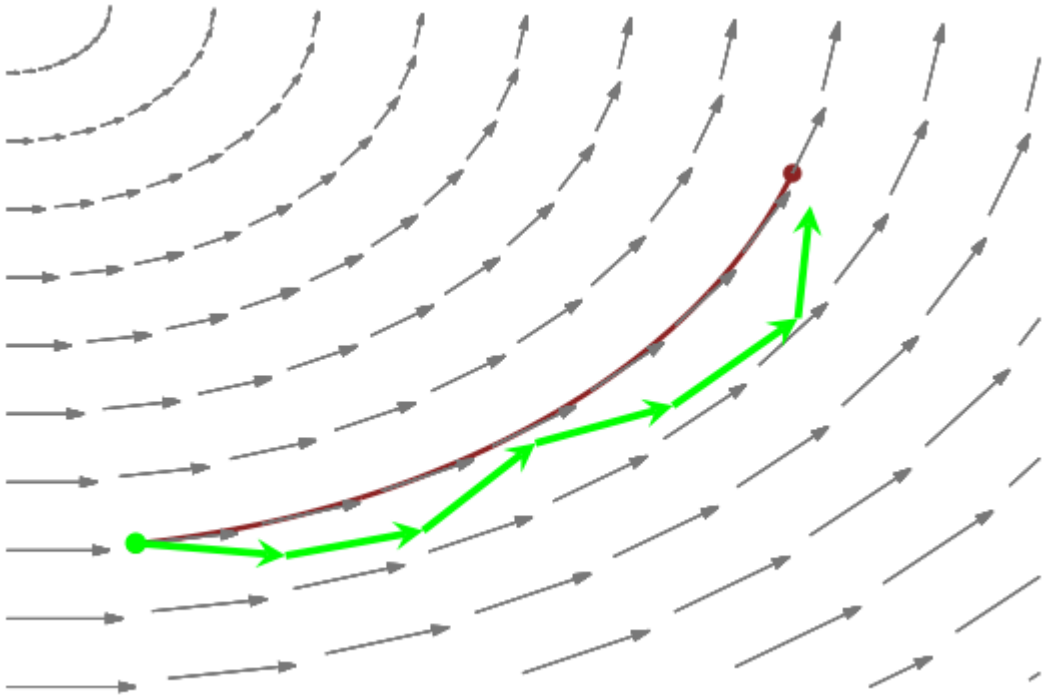
Vamos a necesitar la capacidad de evolucionar un sistema Hamiltoniano. Como es común con las ecuaciones diferenciales, no tienen necesariamente soluciones analíticas, y para lo que nos interesa nos alcanza con integración numérica. Por lo general la integración numérica sufre de artefactos de deriva, pero en el caso de sistemas Hamiltonianos existen familias especiales de integradores numéricos llamados **integradores simpleticos** que se aprovechan de la simetría del sistema para ser más estables.



Los integradores simpleticos son muy potentes por que las trayectorias numéricas que generan preservan el volumen en el espacio de fases, exactamente igual que las trayectorias Hamiltonianas que se busca estimar.

Esta preservación del volumen genera una restricción a que tanto error respecto de las trayectorias exactas en energía puede existir, generando que las trayectorias estimadas no

se alejen de las exactas en cada iteración(ver figura…), sino que estas oscilan alrededor de la verdadera trayectoria sin desviarse incluso tras varias iteraciones (ver figura …).



Por fortuna, los integradores simpleticos no sólo permiten estimar trayectorias acotando el error, sino que además suelen ser fáciles de implementar. Por ejemplo, un integrador simpletico muy usado es el *Leapfrog Integrator*.

Supongamos que la distribución de probabilidad elegida para los momentos es independiente de la posición o coordenada, entonces podemos emplear al integrador leapfrog de la siguiente manera. Dada una grilla de tiempos hasta el tiempo T con un salto temporal ϵ, se simula a la trayectoria exacta como

$q0 \leftarrow q,\ p0 \leftarrow p$

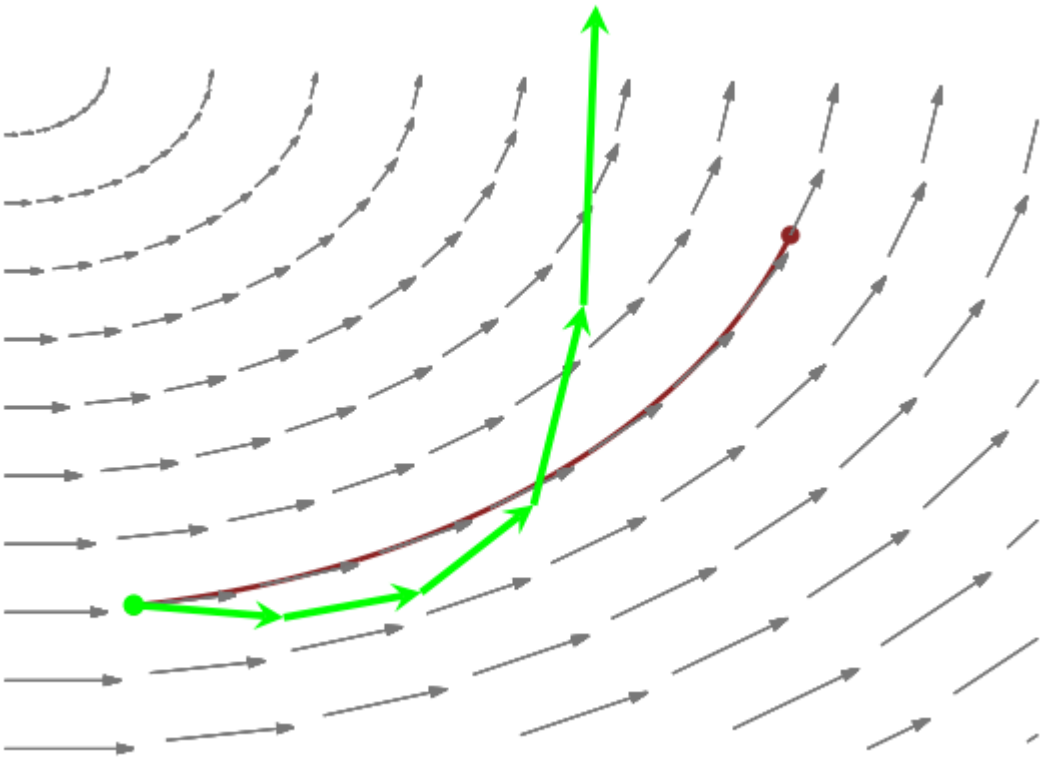
$$p_{n+1/2} \leftarrow p_n - \frac{\epsilon}{2}\frac{\delta V}{\delta q}(q_n)$$

$$q_{n+1} \leftarrow q_n + \epsilon p_{n+1/2}$$

$$p_{n+1} \leftarrow p_{n+1/2} - \frac{\epsilon}{2}\frac{\delta V}{\delta q}(q_{n+1})$$

$endfor$

Hay que tener en cuenta que los integradores simpliciales pueden funcionar mal en trayectorias de curvaturas muy pronunciadas, cuando la cantidad de pasos y la longitud de los mismos no son óptimas (elección de $\epsilon\ y\ T$) producto del error en las trayectorias. Sin embargo, esta anomalía es fácil de identificar ya que se observa que las trayectorias atraviesan muchas energías dirigiéndose a niveles infinitos (ver figura …).



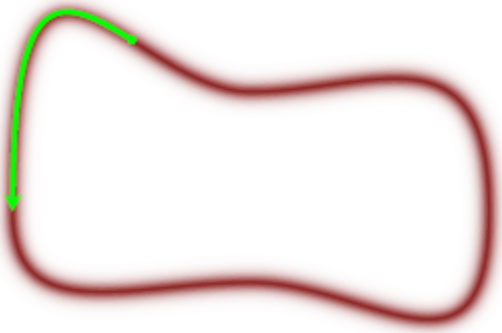
La aproximación de trayectorias mediante integradores, si bien permite llevar a la práctica los fundamentos teóricos del Hamiltonian Monte Carlo, introduce un error que por más pequeño, desvía las trayectorias Hamiltonianas. Además, debemos ser capaces de conseguir un integrador bien formado con la target distribution.

Para arreglar el error en las trayectorias por los integradores, es posible tratar a las transiciones como el proposal en métodos de Metropolis-Hastings en el espacio de fases. Si logramos construir analíticamente una transition distribution la cual logre preservar la target distribution, entonces no importará que el integrador tenga un pequeño error de deriva.

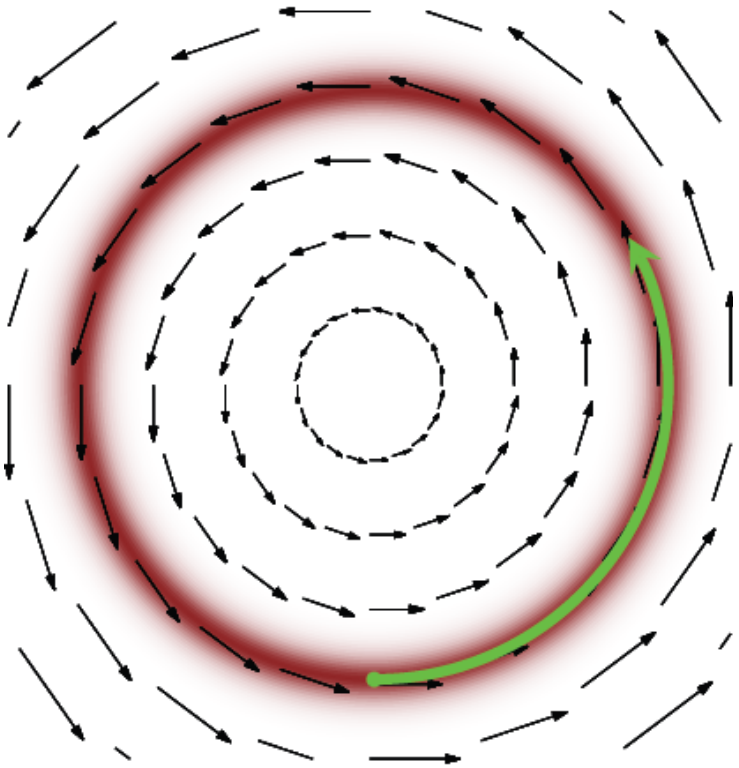
3. Transformación Canónica

a. Motivación de una Transformación

La idea del algoritmo de HMC, nace de la necesidad de explorar el typical set de manera más eficiente. Se busca una exploración que esté alineada a este.



Una idea natural que surge es la de un campo vectorial alineado con el typical set, que dado un punto q, nos indique cómo continuar la trayectoria de nuestra cadena de Markov. Esto generaría una exploración del espacio paramétrico eficiente.



El problema ahora se transforma en la construcción de un campo vectorial, alineado con el typical set, con la información proporcionada con la target distribution. Al ver la estructura diferencial de la target distribution, conseguiremos el campo gradiente. Lamentablemente este campo nos guiará directamente a la moda, lo cual no es lo que estamos buscando en

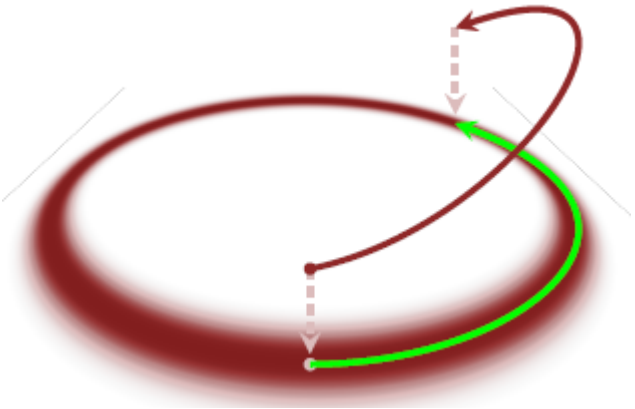
este caso. Lo que buscamos es alguna transformación de la información disponible, el gradiente, que fuerce aquellas propiedades que queremos.

b. Transformación al espacio canónico

La transformación propuesta tiene ciertas instituciones atrás, pero la presentaremos primero. A una posición en el target space, se le asigna un momento. Esto genera una nueva distribución de probabilidad conjunta, para $(p,\ q)$, el cual ahora es llamado el espacio canónico. Aun así, esta transformación se realiza de manera tal que sea fácil de marginalizar nuevamente la distribución objetivo. Generalmente pueden verse como distribuciones independientes y en los pocos casos que no se trata con mucho cuidado.

$$(q) \to (p,\ q)$$
$$\pi(p,\ q) \ = \ \pi(p|q)\pi(q) \ = \ \pi(p)\pi(q)$$

Construyendo la nueva distribución conjunta de este modo, podemos asegurarnos que el nuevo typical set, correspondiente al espacio $(p,\ q)$ el cual llamaremos el canonical typical set, es fácilmente proyectable al target typical set, correspondiente al espacio de q.



Propongamos entonces la transformación de sistemas de mecánica estadística. Siendo E la energía del sistema, T la temperatura, y cte la constante de normalización. En este caso la energía del sistema es equivalente al Hamiltoniano , si asumimos que la temperatura del sistema es 1, y desarrollamos, nos quedamos con una igualdad la cual expresa el Hamiltoniano en función de de nuestras distribuciones.

$$\pi(p,\ q) \ = \ cte.\, e^{-E(p,q)/T}$$
$$\pi(p,\ q) \ = \ cte.\, e^{-H(p,q)}$$
$$ln(\pi(p,\ q)) \ = \ ln(cte) \ - \ H(p,q)$$
$$-\ ln(\pi(p|q)\pi(q)) \ = \ H(p,q) \ - \ ln(cte)$$
$$-\ ln(\pi(p|q)) \ - \ ln(\pi(q)) \ = \ H(p,q) \ - \ ln(cte)$$

Ignorando la constante que se le resta al hamiltoniano, la cual en este caso es 0, comparando dicha ecuación para el Hamiltoniano con la ecuación tradicional para el sistema tenemos que hay en cada una un término dependiente de p, y otro de q. Esto nos indica cuales serían las distribuciones de la energía potencial y de cinética.

$$K(p) =- ln(\pi(p|q))$$
$$U(q) =- ln(\pi(q))$$

Esto tiene como consecuencia que al igual que queríamos hacer con la instauración de un momento, sea fácil recuperar la distribución objetivo en base al momento. Esto deja a la energía potencial de nuestro sistema completamente fijada por nuestra distribución. Una intuición para ella es que a menor potencial, o por consecuencia mayor la densidad de la distribución objetivo, más profundo el bowl en ese lugar del espacio paramétrico.

Veamos que en el desarrollo de las ecuaciones diferenciales se puede apreciar que en el cambio de la posición, se utiliza $\frac{\delta U}{\delta q_i}$ que vendría a representar nuestro gradiente, a través del momento y no de forma directa.

$$\frac{dq_i}{dt} = \frac{\delta H}{\delta p_i} = \frac{\delta K}{\delta p_i}$$
$$\frac{dp_i}{dt} =- \frac{\delta H}{\delta q_i} =- \left(\frac{\delta K}{\delta q_i} + \frac{\delta U}{\delta q_i}\right)$$

o asumiendo independencia entre $\pi(p|q)$ y $\pi(q)$.

$$\frac{dq_i}{dt} = \frac{\delta H}{\delta p_i} = \frac{\delta K}{\delta p_i}$$
$$\frac{dp_i}{dt} =- \frac{\delta H}{\delta q_i} =- \frac{\delta U}{\delta q_i}$$

De esta forma, aumentamos la dimensión del espacio paramétrico con las variables de momento elegidas de forma tal de contrarrestar a la energía potencial que será la fuerza que hace tender el sistema a la moda de la distribución paramétrica. Esto se obtiene forzando que el hamiltoniano se mantenga constante con el sistema de ecuaciones diferenciales logrando que el typical set se mantenga invariante en el espacio formado por las las variables paramétricas y las de momento y además, que el espacio de fases esté alineado con el typical set ya que este se encuentra en el “equilibrio” entre la expansión volumétrica y la compresión a la moda de la distribución paramétrica.
Encontrar trayectorias dentro del typical set en el espacio formado con las nuevas variables de momento nos permitirá encontrar samples en el espacio original gracias a la marginalización de la conjunta.

c. Ciclo del HMC

Dada una distribucion de momentos y su energía cinética acorde, y una energía potencial definida por nuestra target distribution, el ciclo del HMC consiste en un primer paso en el cual se toma un punto inicial de nuestro parameter space q_0 , el cual será el punto de partida. Proponer un momento p_0, produciendo en consecuencia un punto $a = (p_0, q_0)$.

El segundo paso del ciclo consiste en evolucionar a acorde a las dinámicas Hamiltonianas con un integrador numérico por L pasos de integración, con un paso de tamaño ε, siendo estos hiper parámetros a definir. Luego de la evolución del sistema Hamiltoniano se llega a un nuevo punto (p_1, q_1) el cual no necesariamente preservó el Hamiltoniano por errores de integración.

Luego para volver este punto simétrico con respecto a la evolución del sistema, se niega el momento del nuevo punto, consiguiendo por consecuente, $b = (-p_1, q_1)$. Con esto se consigue que la proposal distribution sea igual para $a|b$ y $b|a$, de lo contrario el valor de Q para una de las dos combinaciones. Esto nos permite definir la acceptance distribution.

$$A(b|a) = min\left(1, \frac{Q(a|b)\pi(b)}{Q(b|a)\pi(a)}\right)$$
$$A(b|a) = min\left(1, \frac{\pi(b)}{\pi(a)}\right)$$
$$\frac{\pi(b)}{\pi(a)} = e^{-H(b)+H(a)}$$

En base a $A(b|a)$, es que se elige y se acepta el nuevo punto en el espacio canónico, y para conseguir nuestro punto en el espacio q, es tan fácil como simplemente descartar el momento. Este proceso luego se repite re sampleando un momento nuevo, y manteniendo la última q a la cual se transiciono .

d. Preservación de la distribución canónica

Veamos ahora qué realizar nuestro ciclo del HMC, preserva la distribución canónica, lo cual implica que los puntos de nuestra cadena de Markov se concentran cerca del canonical typical set, y por la construcción de esta distribución, la proyección sobre el espacio objetivo, el descarte del momento de nuestros puntos, nos produce una cadena de Markov que se concentra cerca de la proyección del canonical typical set, el cual es nuestro typical set deseado.

En el primero paso, el sample de un momento aleatorio es acorde a la distribución condicional, la cual es la misma que la marginal por independencia, por lo que al no moverse, se preserva la invarianza de la distribución canónica.

En la segunda etapa en el caso de lograr una transición, podemos ver que se preserva la invarianza, bajo una pequeña reformulación de esta. Supongamos que se particiona el espacio canónico en (A_k) conjuntos, cada uno con el mismo volumen V_A, consideremos la imagen de (A_k) bajo el Leapfrog Integrator bajo cierta cantidad de pasos y con un tamaño de paso fijos, a dicha imagen la llamaremos (B_k). Como el Leapfrog Integrator es reversible, es fácil ver que (B_k) también establece una partición en el espacio canónico, por último como el Leapfrog Integrator también preserva volúmenes, entonces (B_k) son del mismo volumen que (A_k). La invarianza se mantendrá si se cumple lo siguiente para todo par de conjuntos.

$$P(A_i)T(B_j|A_i) = P(B_j)T(A_i|B_j)$$

En estas igualdades, P representa la probabilidad bajo la distribución canónica, y recordemos a T(a|b), la transition distribution, como la probabilidad condicional de primero alcanzar el punto b desde el punto a, y posteriormente aceptarlo.

$$T(a\ |b) = A(b|a)Q(a|b)$$

En este caso T siempre y cuando i sea distinto que j valdrá 0 en ambos sentidos, esto causa que para todo i distinto de j la igualdad se satisfaga trivialmente. Para los casos en los que el índice es el mismo, al ser el hamiltoniano continuo para casi todo punto, siempre se puede achicar el volumen de los conjuntos a un punto en el cual el valor del hamiltoniano en todos los puntos es el mismo, llamémoslos H_{A_i} y H_{B_i} para cada i. Ahora veamos cómo queda la ecuación para cada i, ignorando el índice ya que estas son iguales en estructura.

$$Q(A|B) = 1$$
$$P(A)T(B|A) = P(B)T(B|A)$$
$$V.exp(-H_A)min\Big(1, exp(-H_B + H_A)\Big) = V.exp(-H_B)min\Big(1, exp(-H_A + H_B)\Big)$$
$$exp(-H_A + H_B)min\Big(1, exp(-H_B + H_A)\Big) = min\Big(1, exp(-H_A + H_B)\Big)$$

En el caso de que $min\Big(1, exp(-H_B + H_A)\Big)$ sea menor a uno, el término contrario será mayor, por lo que se divide en dos casos y viceversa.

$$\text{i) } exp(-H_A + H_B) = exp(-H_A + H_B)$$
$$\text{ii) } exp(-H_A + H_B)exp(-H_B + H_A) = 1$$
$$exp(0) = 1$$

e. Significado de la preservación de la distribución

Tomemos R como la probabilidad de que se rechace una nueva región en función de qué región estamos

$$R(B_k) = 1 - \sum_i T(A_i|B_k)$$

Entonces estudiemos la probabilidad de que un nuevo estado del HMC caiga en algún B_k llamemos a esto $T(B_k)$, sin ningún estado actual dado. Esto puede darse debido a que ya estábamos en B_k y se rechazó un nuevo estado, más la probabilidad de que caigamos en $B_k$desde otra región, y se acepte. El desarrollo quedaría del siguiente modo.

$$T(B_k) = P(B_k)R(B_K) + \sum_i P(A_i)T(B_k|A_i)$$

$$T(B_k) = P(B_k)R(B_K) + \sum_i P(B_k)T(A_i|B_k)$$

$$T(B_k) = P(B_k)R(B_K) + P(B_k)(1 - R(B_K))$$
$$T(B_k) = P(B_k)$$

Habiendo visto que, preservar la distribución significa que el próximo estado de nuestra cadena de Markov se distribuye de la misma manera que nuestra distribución, se vuelve intuitivo entender porque nuestra cadena se asemeja al typical set de la distribución.

d. Conversaciones varias

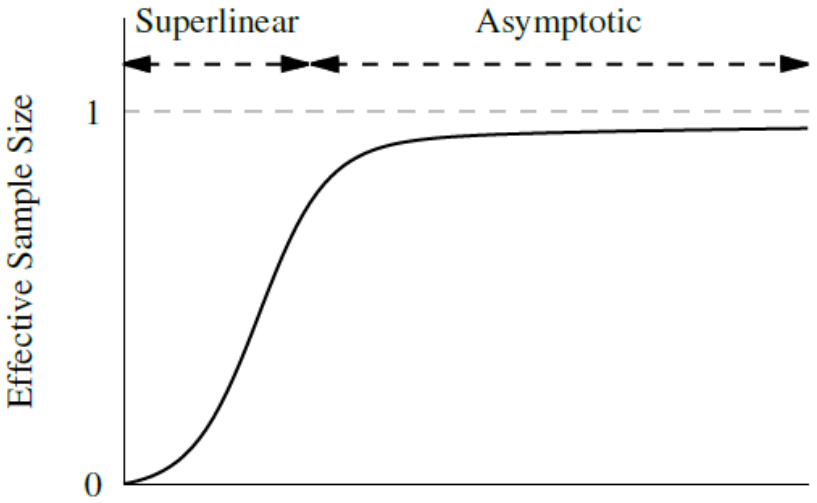
Primero queríamos decir que lo propuesto hasta ahora deja de manera agnóstica la elección de la energía cinética. Hay muchas implementaciones y maneras de ajustar las energías cinéticas, permitiendo espacio para ajustar el modelo acordemente. La energía cinética más común es una normal multivariada con covarianza 0 y media 0. Esta puede ser fácilmente vista como la distribución a la cual tras tomarle el operador p (-log) se consigue algo similar a la energía cinética tradicional de física. Otra cosa interesante es la existencia de energías cinéticas no independientes de la distribución objetivo, las cuales requieren herramientas especiales para demostrar invarianza de distribución pero parecen proveer energías adaptables.

La bibliografía también recomienda no proponer sólo el último punto de la trayectoria, sino que en cambio se podría proponer a un sampleo dentro de los puntos dentro de la trayectoria numérica, obteniendo un proposal que sigue siendo reversible y se define la probabilidad de aceptación como:

$$A(q_L, p_L|q_0, p_0) = min(1, e^{-H(q_i,-p_i)+H(q_0,-p_0)}) \; para \; 0 < i < L \, .$$

Este método no optimiza la elección de los parámetros ya que se verá forzado a analizar estados con poca probabilidad y rechazarlos cuando lo óptimo sería que directamente analice a los estados con alta probabilidad (Estados con alta probabilidad son estados con un error pequeño).

Por otro lado hay otra problemática que se genera en el HMC el cual es el decrecimiento de la información ganada con un tiempo de integración lo suficientemente grandes, esto no es difícil de imaginar, debido a que el sistema es conservativo, muy probablemente sea periodico, y en el momento en el que se pasa por el vecindario nuevamente, la información ganada no es tan relevante. En este gráfico se ve como crece el effective sample size comparativamente al tiempo de integración con el mismo Hamiltonian. Existen optimizadores del tiempo de integración automáticos, un ejemplo es el criterio No-U-Turn.



Pareciera que el HMC también se desvirtúa cuando el typical set es inconexo, no tenemos mucha información al respecto más que en STAN estaban trabajando en una solución, aunque seguía siendo muy primitiva.

4. Bibliografía y recomendaciones

La bibliografía sobre la cual más aprendimos y mas nos basamos es la siguiente:

- A Conceptual Introduction to Hamiltonian Monte Carlo - Michael Betancourt
- Chapter 5 of the Handbook of Markov Chain Monte Carlo - Steve Brooks, Andrew Gelman, et all

Lamentablemente no pudimos cubrir el concepto de pensar el sistema en función a sus niveles de energía, y explorar estos, entre otras cosas que nos parecen interesantes también, por lo que sí también le resulta interesante recomendamos la lectura de estos papers, y esperamos que con lo explicado acá se vuelva más amena y disfrutable la lectura.

Esta Bibliografía la consideramos útil para un primer acercamiento a la definición formal de un Hamiltoniano:

- https://profoundphysics.com/hamiltonian-mechanics-for-dummies
- https://profoundphysics.com/lagrangian-mechanics-for-beginners/
- https://en.m.wikipedia.org/wiki/Calculus_of_variations
- https://en.m.wikipedia.org/wiki/Fundamental_lemma_of_the_calculus_of_variations

Afortunadamente la definición formal del Hamiltoniano no es necesaria, y nos basta con la más simple de las proporcionadas en el primer link.

Lecturas Interesantes pero un poco más complejas:

- The Geometry of Hamiltonian Monte Carlo - Michael Betancourt

- The Geometrical Foundations of Hamiltonian Monte Carlo - Michael Betancourt, Simon Byrne, et all

No fuimos capaces de digerir la información de estos papers por completo, no esperamos que el lector pueda tampoco, pero esta bueno saber donde poder buscar definiciones incluso más técnicas para el momento en el cual tengamos las herramientas. Estos conceptos geométricos son las bases para mejorar la elección de energías cinéticas.