



UNAM
POSGRADO



Programa
Universitario
de Estudios
del Desarrollo
UNAM

Algoritmos de muestreo y Monte Carlo: GIBBS y HMC

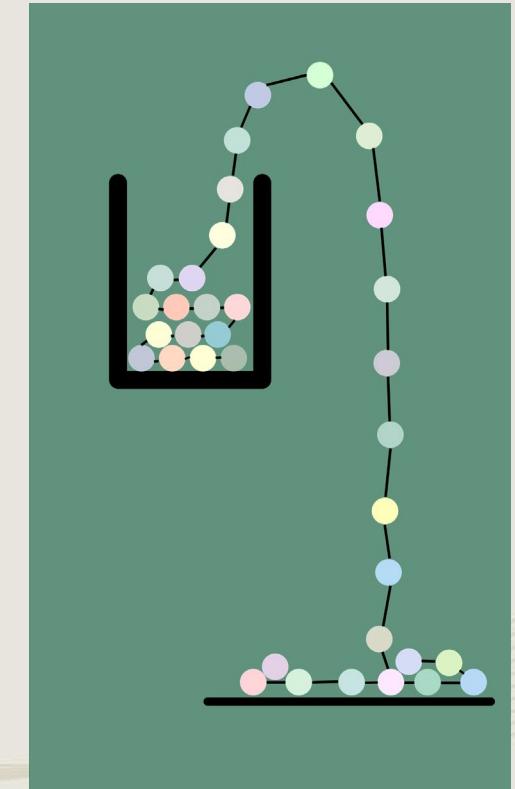
Dr. Héctor Nájera
Dr. Curtis Huffman

¿Cómo se muestrea a partir de una distribución de probabilidad?

- Usualmente, en las aplicaciones al “mundo real”, no es nada fácil muestrear a partir de distribuciones (CDFs a posteriori) complicadas.

Los Métodos Montecarlo aplicados a Cadenas de Markov permiten (tarde o temprano) recuperar las distribuciones objetivo.

- Metropolis (1953)
- Metropolis-Hastings (1970)
- Gibbs (1984)
- Hamiltoniano (o híbrido; 1987)



Algoritmo Metropolis

- Este algoritmo muestrea de una función objetivo complicada (la posteriori) usando una distribución simple como propuesta de transición a partir de la última posición.
 - Se arranca de una posición aleatoria cualquiera (válida) en el espacio de parámetros.
 - Con base en la ubicación actual, se extrae de manera aleatoria, de una función de distribución simple y simétrica, una propuesta del siguiente paso (en la vecindad) a dar.
 - Si el paso conduce a un punto con mayor densidad de probabilidad (evaluando la posteriori en el punto), se acepta la propuesta y se avanza en la dirección planteada.
 - Si el paso conduce a un punto con menor densidad, se acepta la propuesta con probabilidad igual a la proporción de densidad que la propuesta representa de la posición actual. Si la propuesta se rechaza la posición actual se cuenta de nuevo.

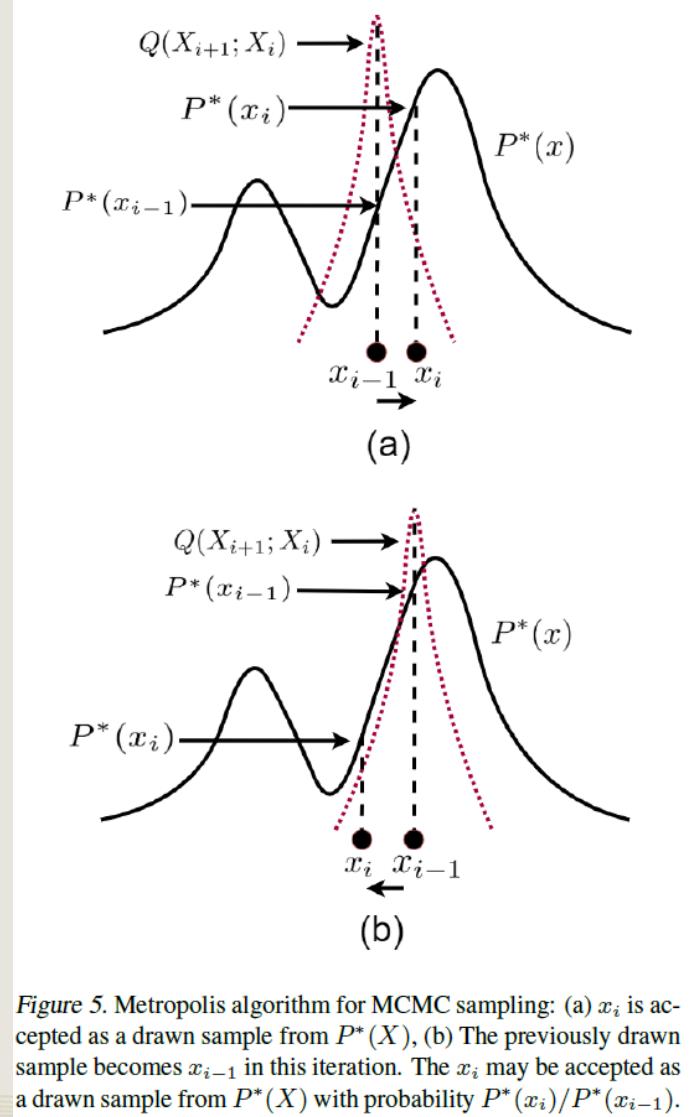
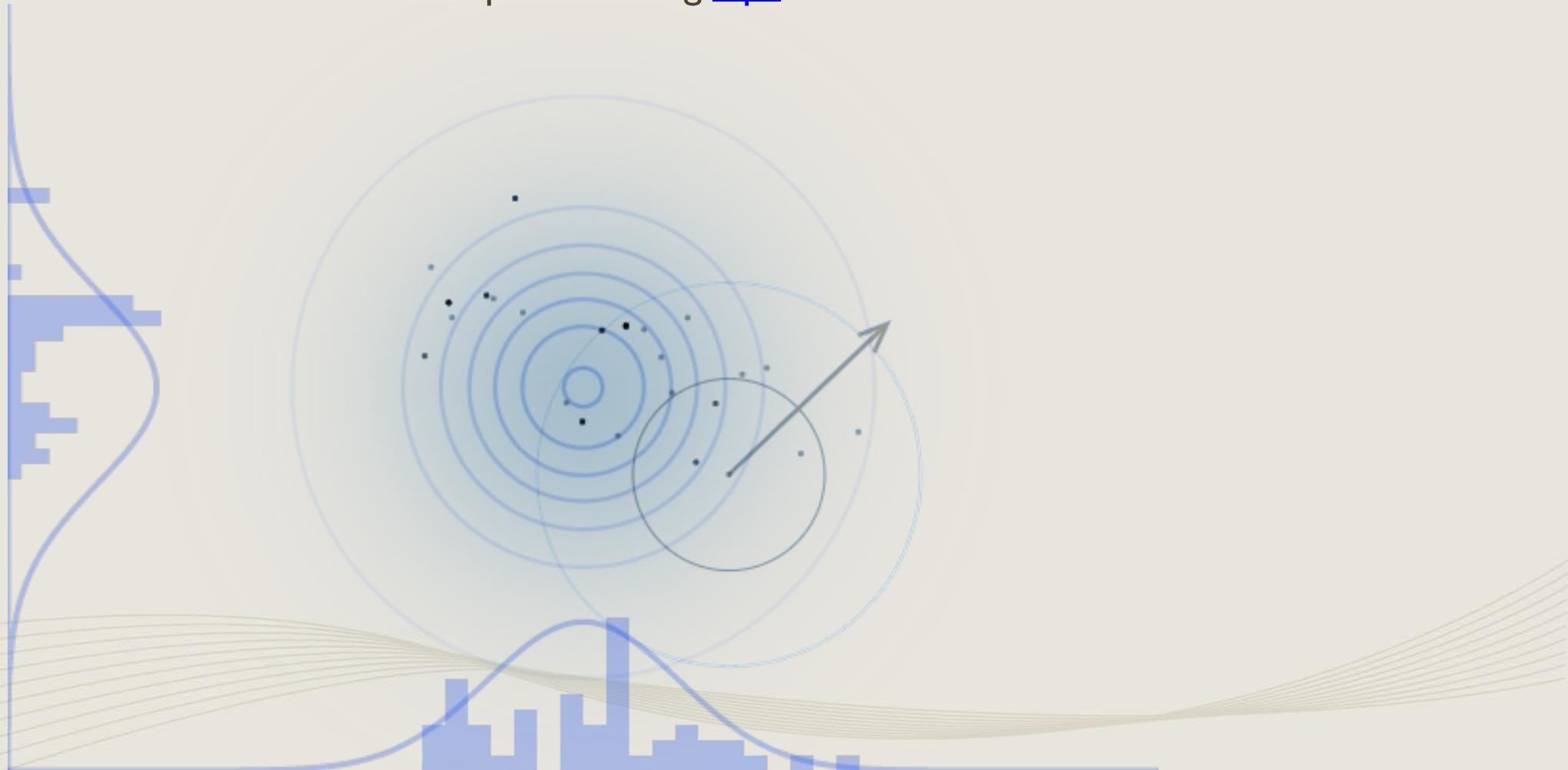


Figure 5. Metropolis algorithm for MCMC sampling: (a) x_i is accepted as a drawn sample from $P^*(X)$, (b) The previously drawn sample becomes x_{i-1} in this iteration. The x_i may be accepted as a drawn sample from $P^*(X)$ with probability $P^*(x_i)/P^*(x_{i-1})$.

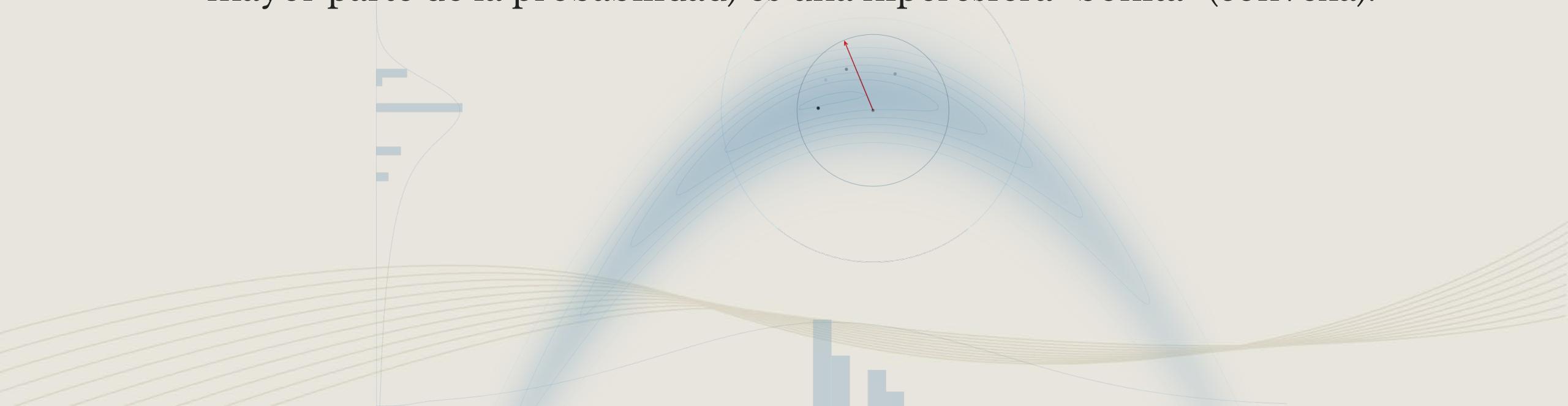
Simulaciones

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).



Algoritmo Metropolis

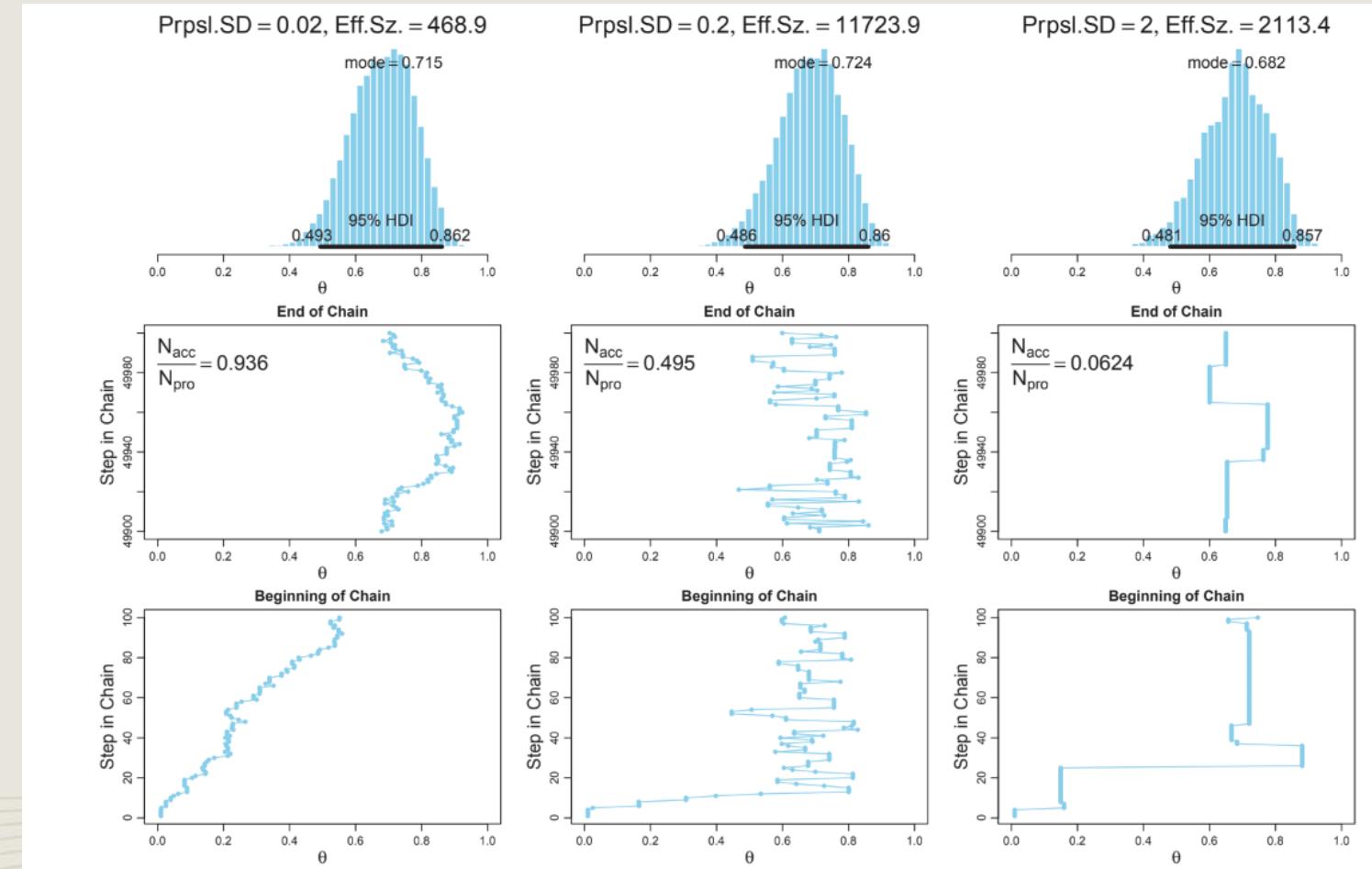
- A pesar de ir “dando tumbos” en una *senda aleatoria*, el algoritmo de Marshall y Arianna Rosenbluth (y su generalización Metropolis-Hastings) funciona!, pero ese es justo su problema: es demasiado aleatorio.
 - Gasta muchísimo tiempo reexplorando las mismas partes de la distribución objetivo desperdiциando tiempo de cómputo valioso.
 - Prácticamente sólo es eficiente en el caso normal/gausiano de bajas dimensiones donde el *conjunto típico* (la parte del domino que concentra la mayor parte de la probabilidad) es una hiperesfera “bonita” (convexa).





¿Cómo saber que el muestreo es pobre?

Veremos esto con calma y la evaluación formal –próxima clase- pero de entrada piensen en los siguientes casos.



Los problemas en la exploración con MCMC

Por qué es un problema?

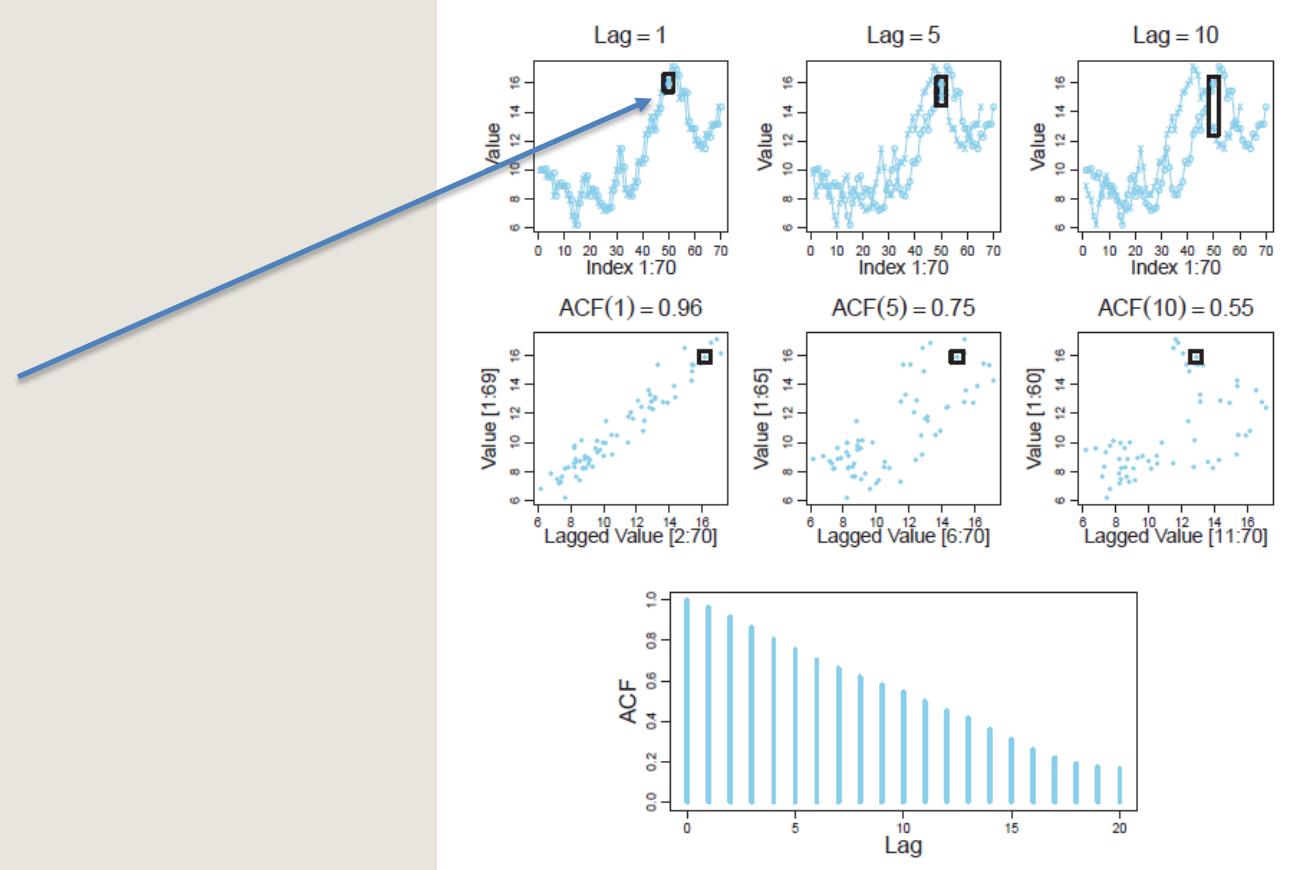


Figure 7.12: Autocorrelation of a chain. Upper panels show examples of lagged chains. Middle panels show scatter plots of chain values against lagged chain values, with their correlation annotated. Lowest panel shows the autocorrelation function (ACF). Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.



Algoritmo Gibbs (MH)

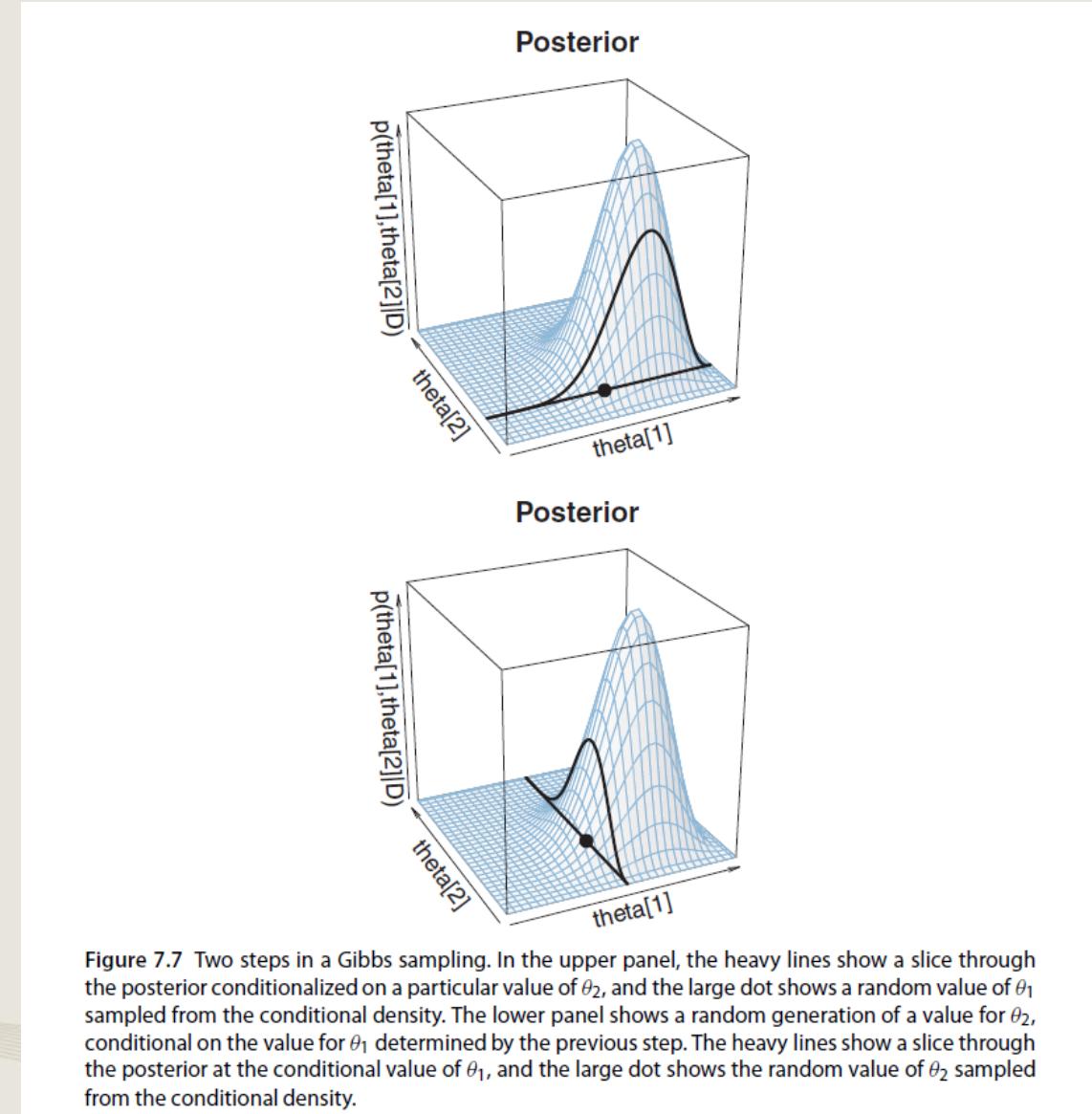


De izquierda a derecha, Stuart y Donald Geman, París 1983

Algoritmo Gibbs (MH)

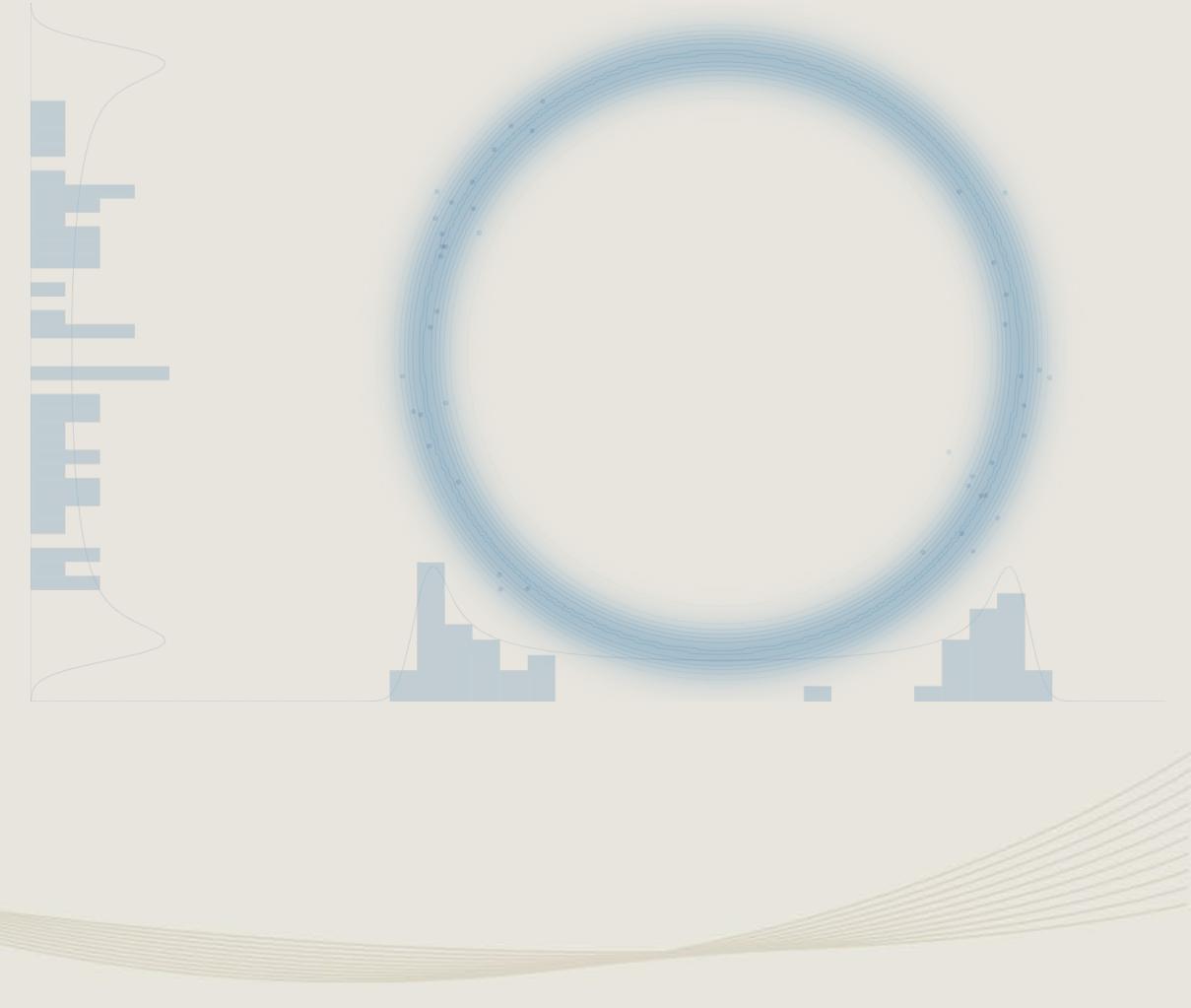
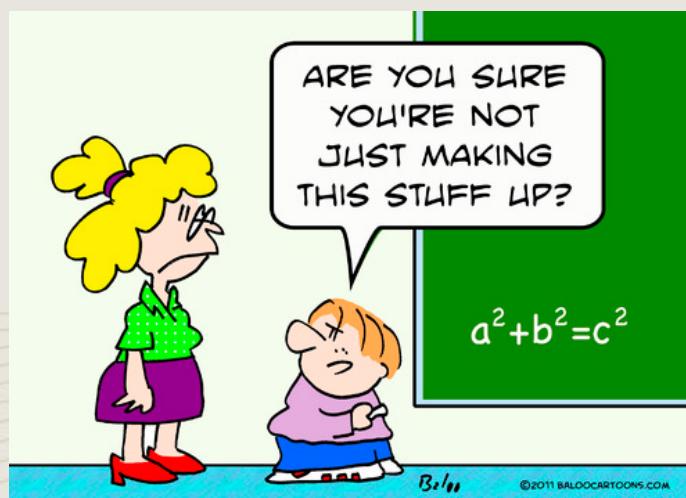
- El muestreo Gibbs es un caso particular de MH que modifica la distribución de donde se extraen las propuestas de siguiente paso en la senda aleatoria.
 - A cada paso de la senda, sólo una de las dimensiones del espacio de parámetros es seleccionada, y la propuesta de paso siguiente se extrae de manera aleatoria de la **distribución posterior condicional** (univariada) de ese parámetro, $p(\theta_1 | \{\theta_{j \neq 1}\}, D)$.

Noten que este algoritmo requiere que se pueda muestrear **directamente** de todas las distribuciones condicionales.



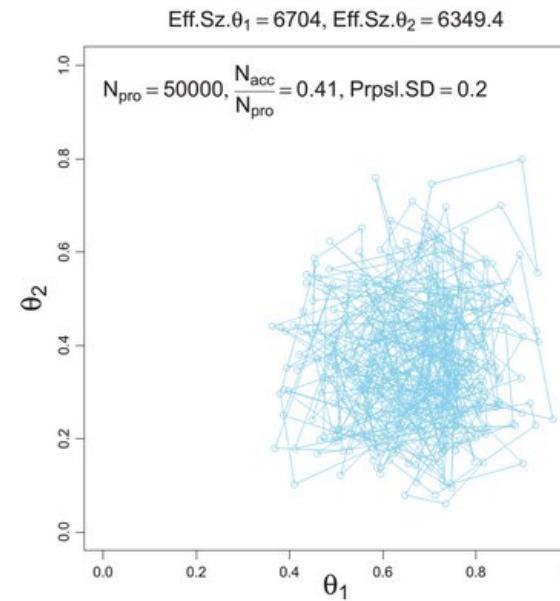
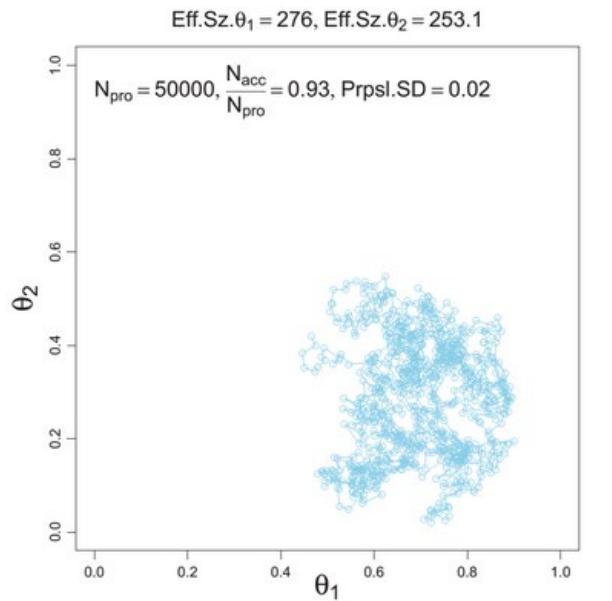
Algoritmo Gibbs (MH)

- Debido a que el muestreo à la Gibbs sólo cambia un parámetro a la vez, éste puede “atascarse” (dar pasos muy pequeños) con parámetros altamente correlacionados (conjuntos típicos esbeltos).



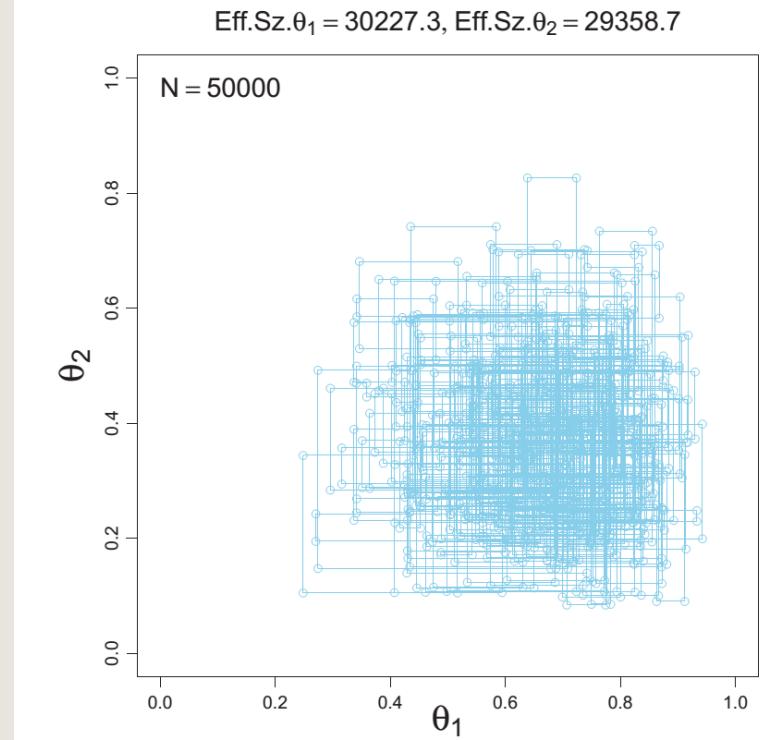
MH vs Gibbs

Metropolis-Hastings



Metropolis-Hastings is great, simple, and general. BUT ... sensitive to step size. AND ... can be too slow, because it can end up rejecting a great many steps.

Gibbs sampling



Lo que ustedes quieren son más muestras de calidad (efectivas) en menos tiempo!

De qué depende que sean más eficientes?



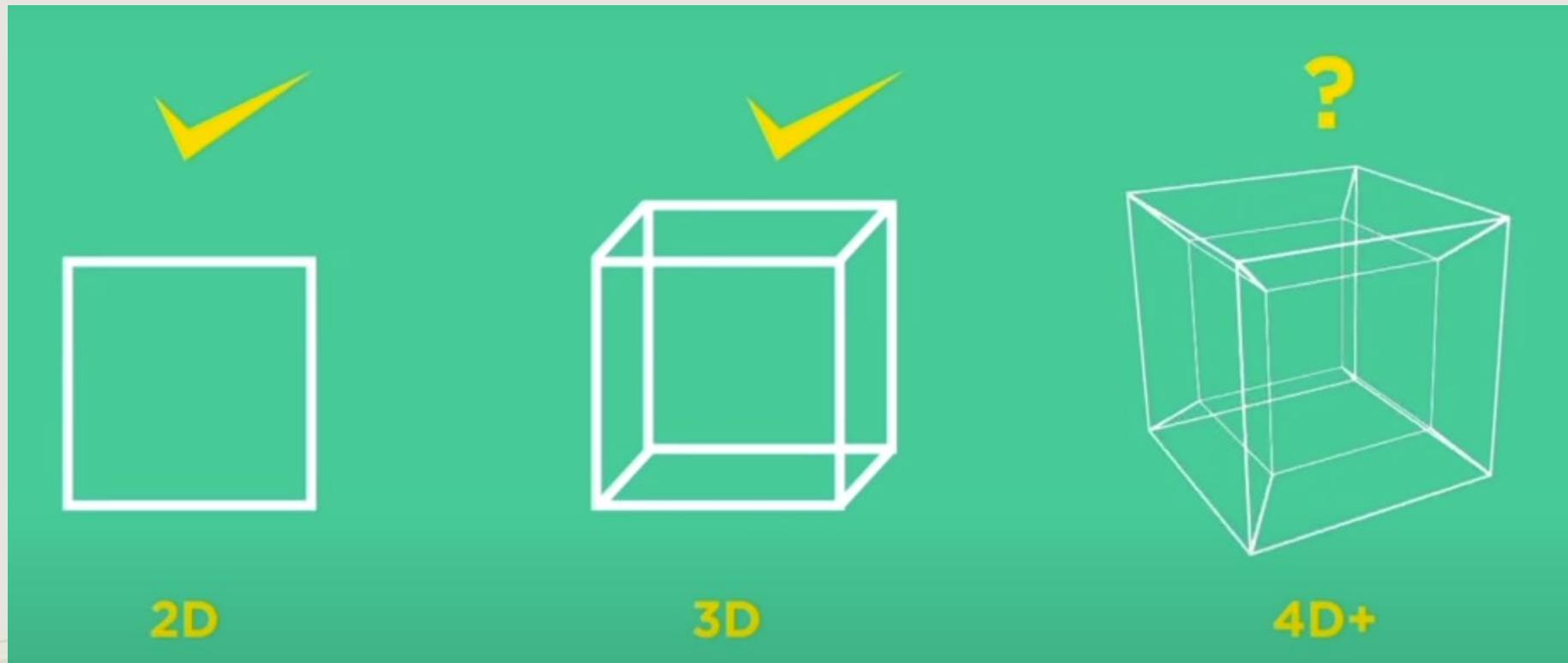
Mirando el Gibbs

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).



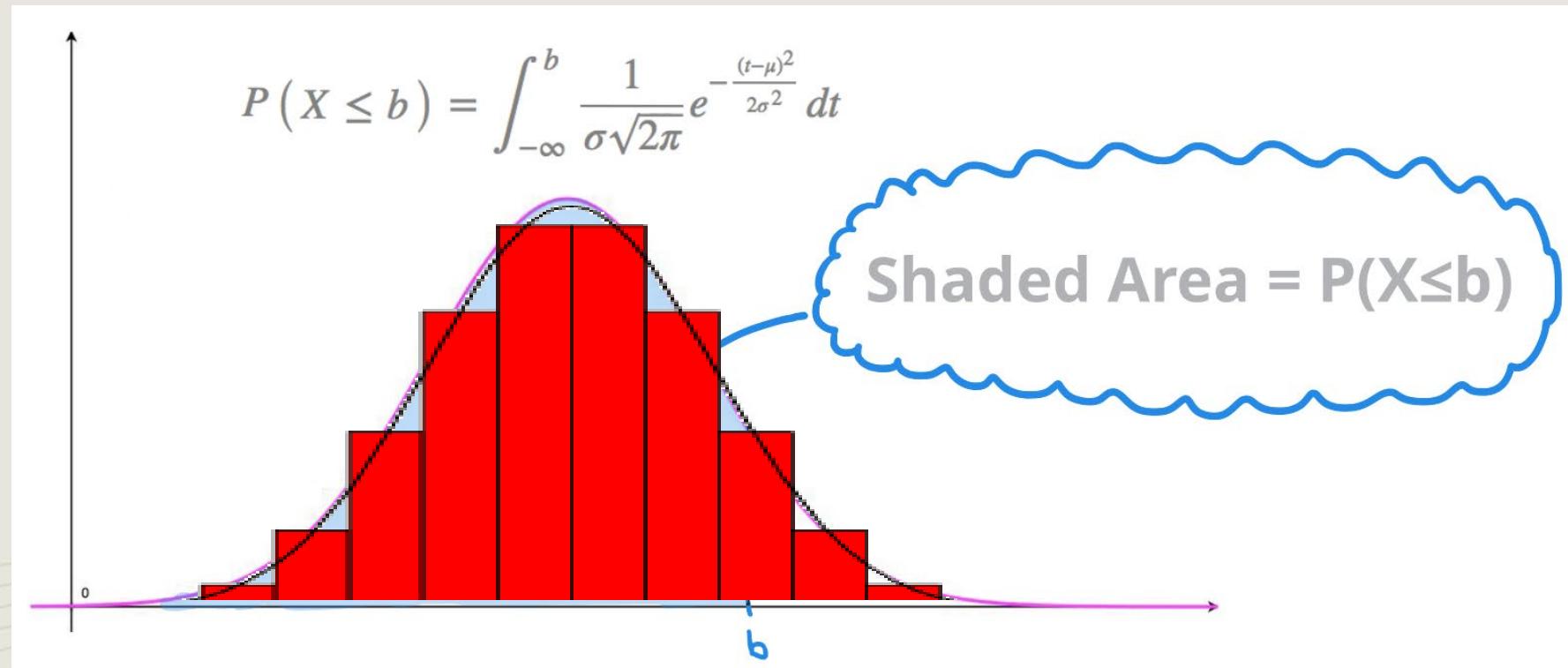
¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales



¿Conjuntos típicos esbeltos?

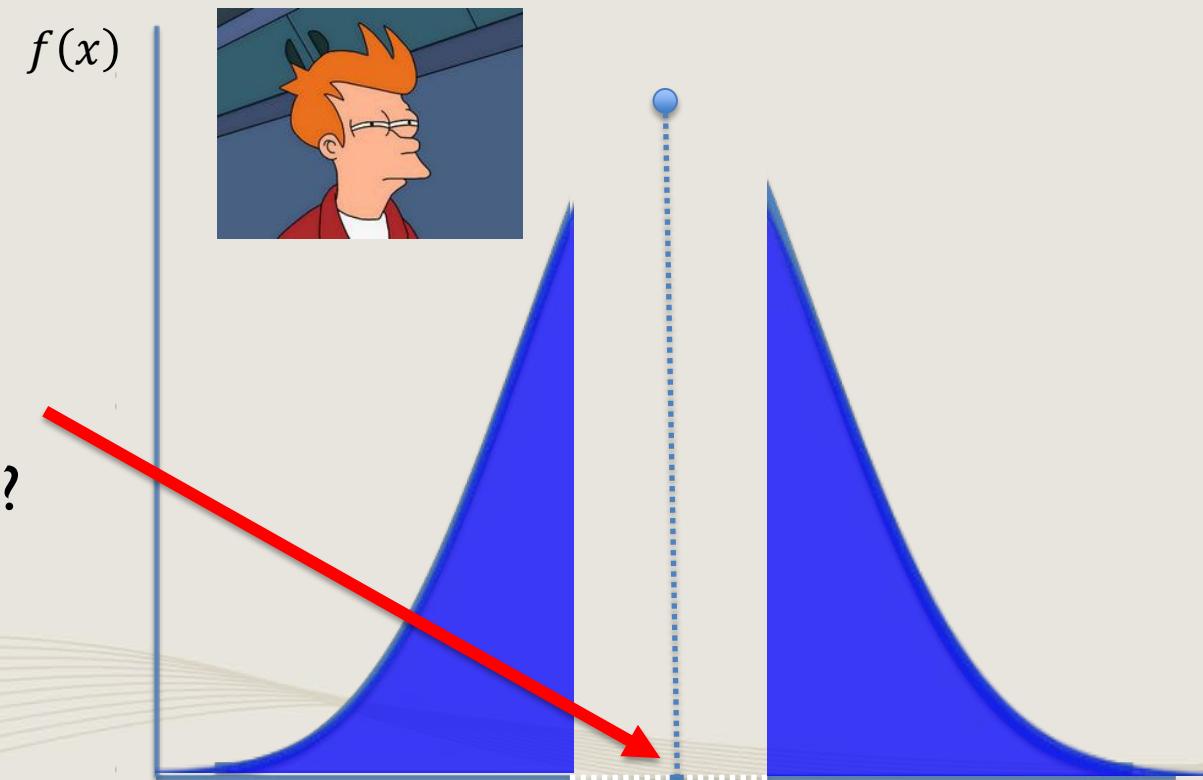
- Recuerden que, en el caso continuo, la masa de probabilidad está dada no meramente por la densidad (la altura de la función), sino por el **área** (base x altura) bajo la curva: el límite del área de los rectángulos rojos cuando la base tiende a cero (la integral)



¿Conjuntos típicos esbeltos?

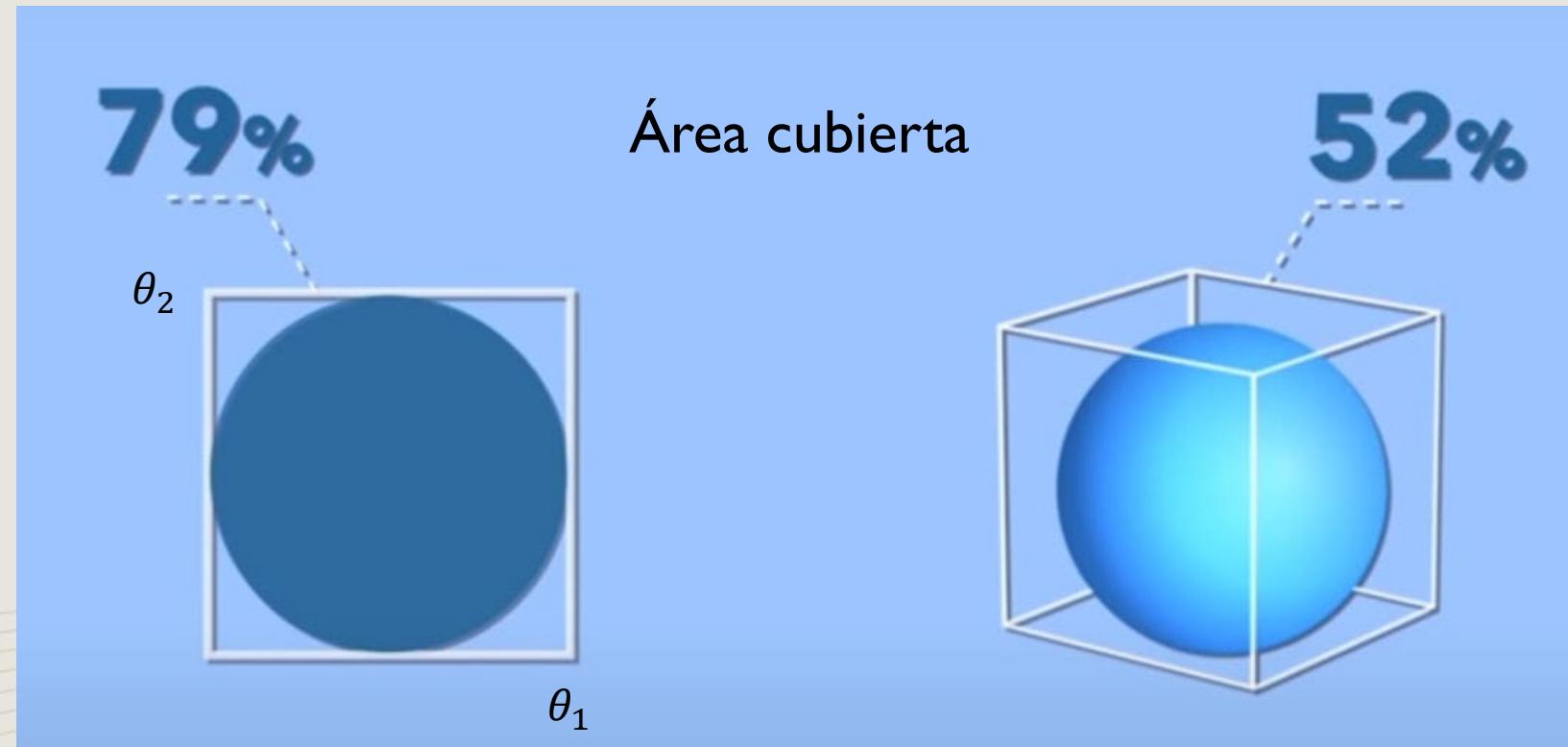
- La densidad (la altura), $f(x)$, puede ser mucha (la moda), pero si el dominio (la base) recorre poca distancia (área/volumen, dt), la masa de probabilidad que acumula (la integral) es despreciable.

¿Cómo puede recorrer
poca distancia
(área/volumen) el dominio?



¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).





¿Conjuntos típicos esbeltos?

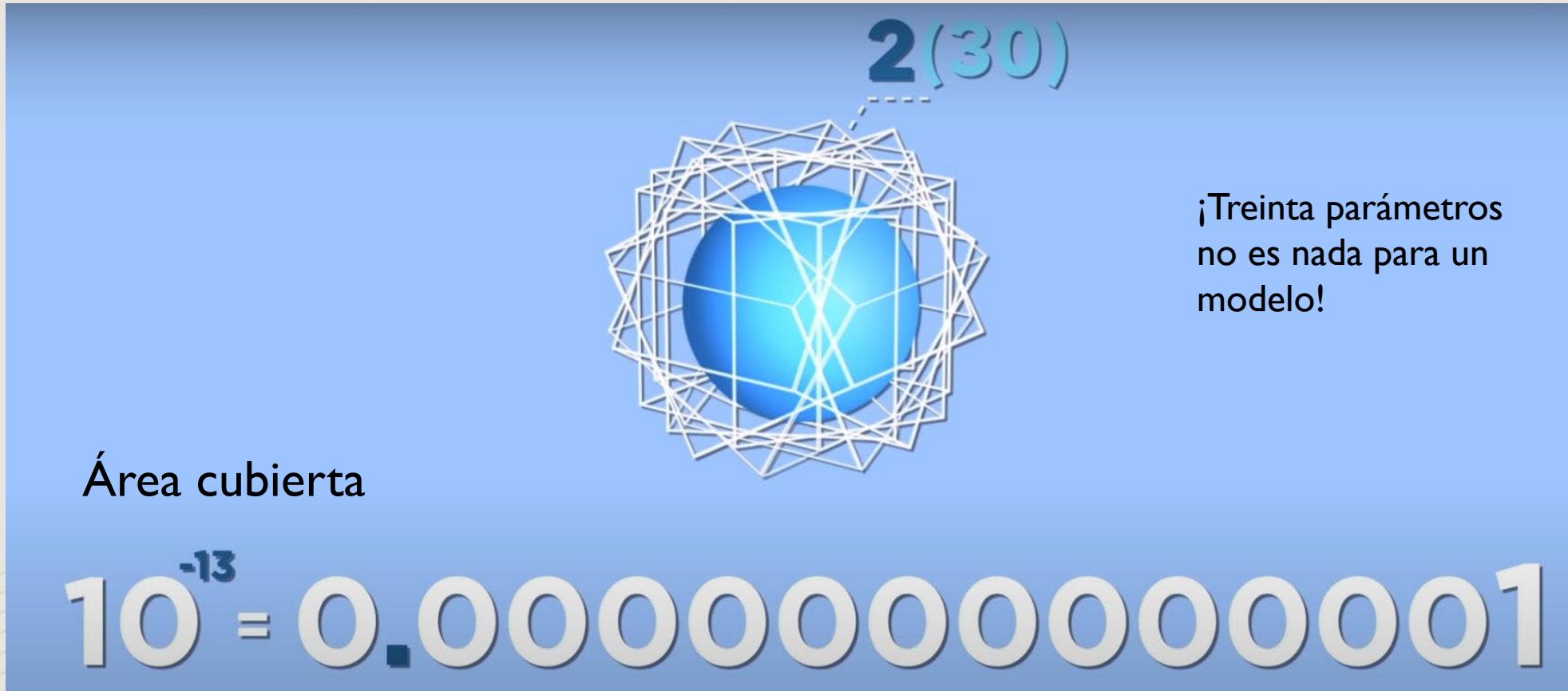
- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).





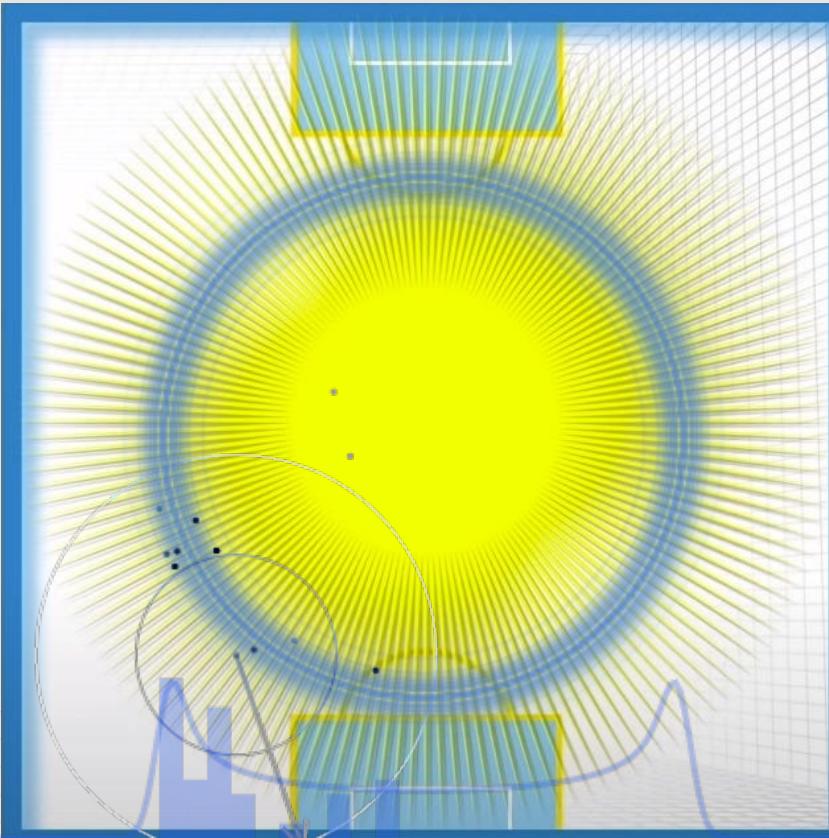
¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



¿Conjuntos típicos esbeltos?

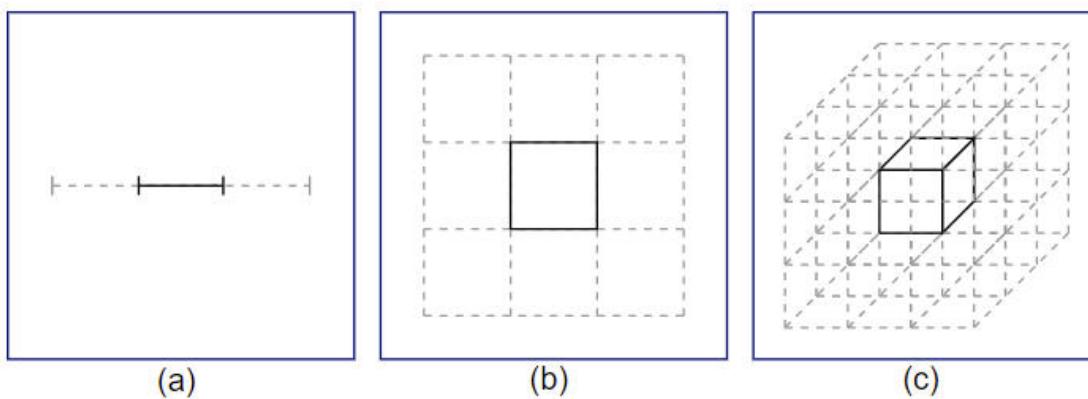
- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



- Con sólo 30 dimensiones (parámetros), el recorrido/área/volumen cubierto por una hiperesfera circunscrita en ese espacio es proporcional a la de un grano de arroz en una cancha de futbol.
- La densidad puede ser mucha cerca de la moda (el centro de la cancha), pero el recorrido del dominio es despreciable.
- Hay muchísima más área lejos de la moda.

¿Conjuntos típicos esbeltos?

- Cosas inesperadas y constraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).

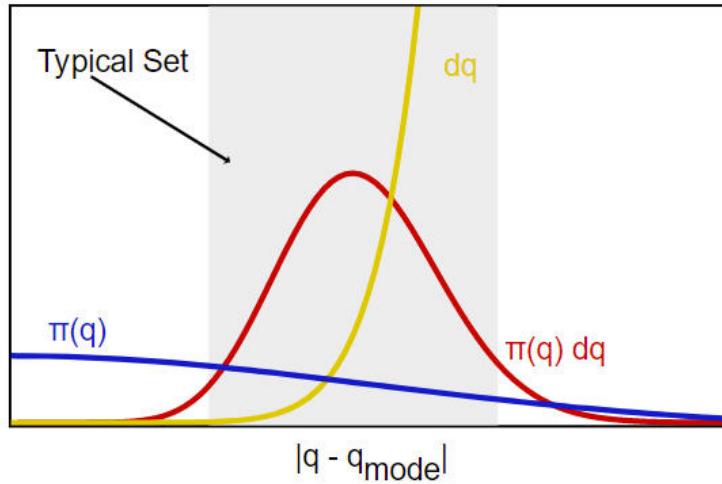


To illustrate the distribution of volume in increasing dimensions, consider a rectangular partitioning centered around a distinguished point such as the mode. The relative weight of the center partition is (a) $1/3$ in one dimension, (b) $1/9$ in two dimensions, (c) and only $1/27$ in three dimensions. In d dimensions there are 3^{d-1} neighboring partitions. Very quickly the volume in the center partition becomes negligible compared to the neighboring volume. This effect only amplifies if we consider larger regions around the mode, i.e. partitions beyond the nearest neighbors. For instance, for next-to-nearest neighbors, the base would be 5^{d-1} .

- Con sólo 30 dimensiones (parámetros), el recorrido/área/volumen cubierto por una hiperesfera circunscrita en ese espacio es proporcional a la de un grano de arroz en una cancha de futbol.
- La densidad puede ser mucha cerca de la moda (el centro de la cancha), pero el recorrido del dominio es despreciable.
- Hay muchísima más área lejos de la moda.

¿Conjuntos típicos esbeltos?

- Cosas inesperadas y contraintuitivas ocurren en espacios multidimensionales.
Imaginen un círculo (hiperesfera) inscrito(a) en un cuadrado (hipercubo).



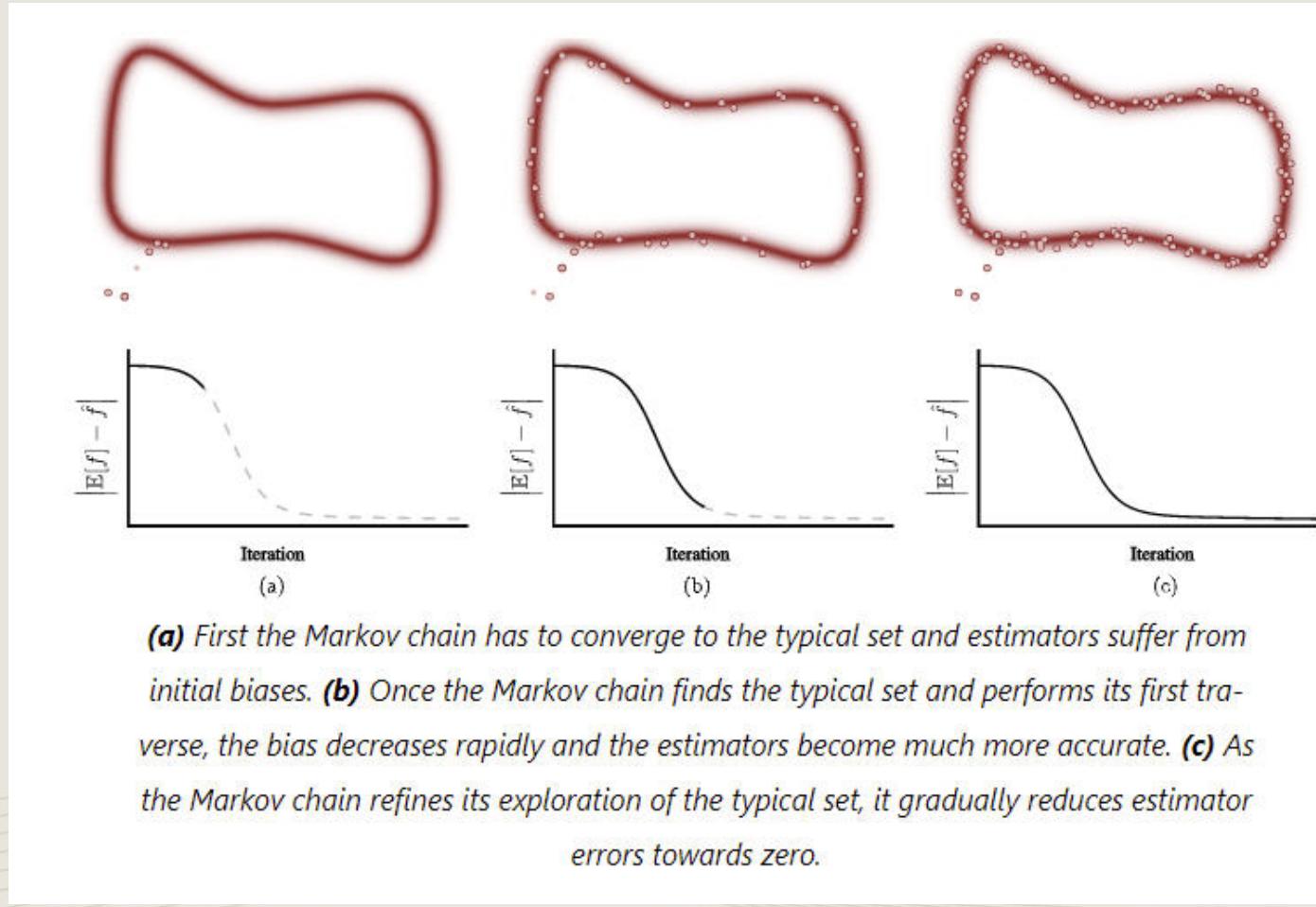
In high dimensions a probability density $\pi(\mathbf{q})$ will concentrate around its mode, but the volume $d\mathbf{q}$ over which we integrate that density is much larger away from the mode. Contributions to expectations are determined by the product $\pi(\mathbf{q}) d\mathbf{q}$. In sufficiently high dimensions, this condition is satisfied only in a nearly-singular region of \mathcal{Q} called the typical set. (This plot only shows a 10-dimensional independent, identically-distributed unit Gaussian. Hence the finite width of the typical set.. \mathbf{q} is the full 10-dimensional vector in parameter space and q in the above figure denotes its radial component.)

- Como aumenta el número de dimensiones (parámetros en el modelo), el conjunto típico (que acumula el grueso de la masa de probabilidad) se vuelve súper esbelto y prácticamente imposible de explorar dando tumbos (MH) o à la Gibbs.



¿Conjuntos típicos esbeltos?

En condiciones ideales, las cadenas de Markov exploran la distribución posterior en tres fases.



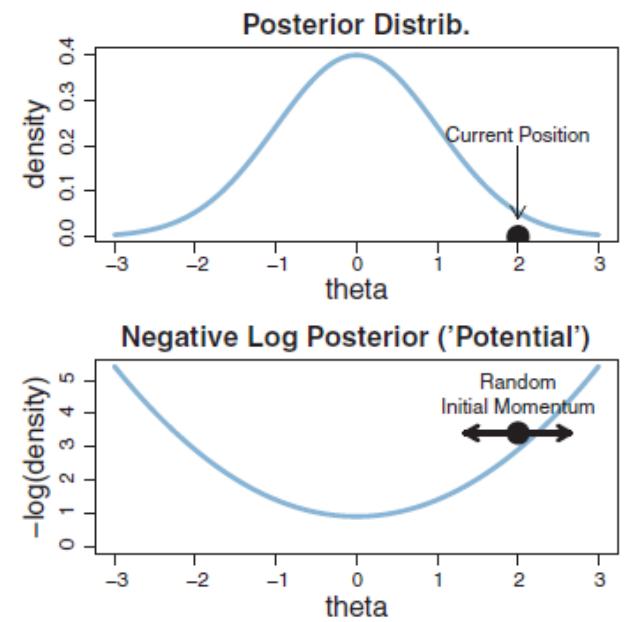
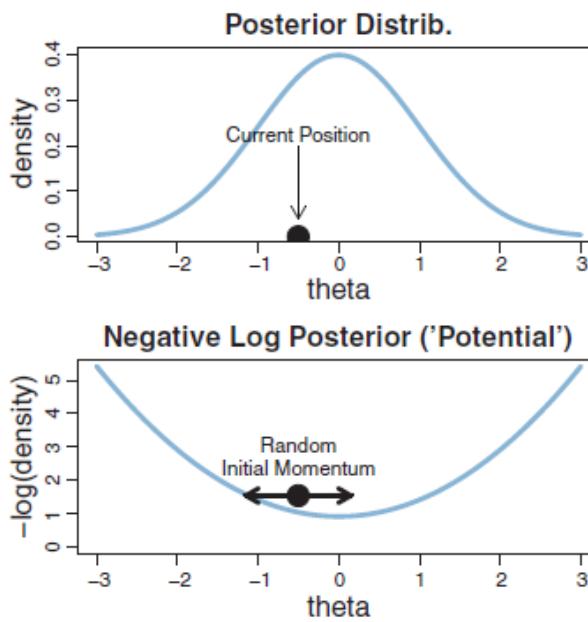


Mirando el Gibbs

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).

Monte Carlo Hamiltoniano (HMC)

- Variación ingeniosa del algoritmo Metrópolis (véase Kruschke 2018, cap. 14) en la que la distribución de propuestas cambia de acuerdo con la posición del más reciente eslabón de la cadena en la dirección del conjunto típico.
 - HMC genera propuestas, por analogía con la exploración de un sistema físico, explotando la geometría del conjunto típico como se rodaría una canica sobre la distribución posterior vuelta de cabeza.



Monte Carlo Hamiltoniano (HMC)

- Toda vez que el gradiente (la derivada multivariada) de la distribución posterior apunta en la dirección de máximo crecimiento de la función, es necesario complementar/contrarestar esta “gravedad” (energía potencial) con otra “fuerza” (energía cinética) que le permita “rodar”/“orbitar” (eternamente) por el conjunto típico (tomando muestras): espacio de fases.

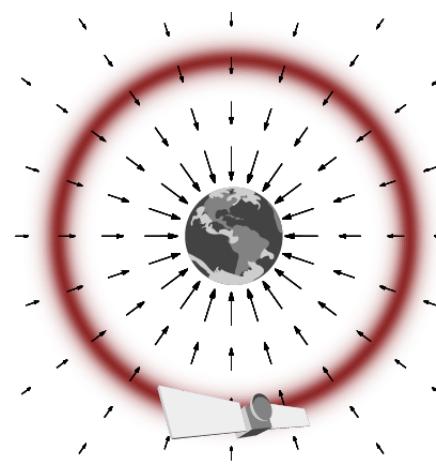
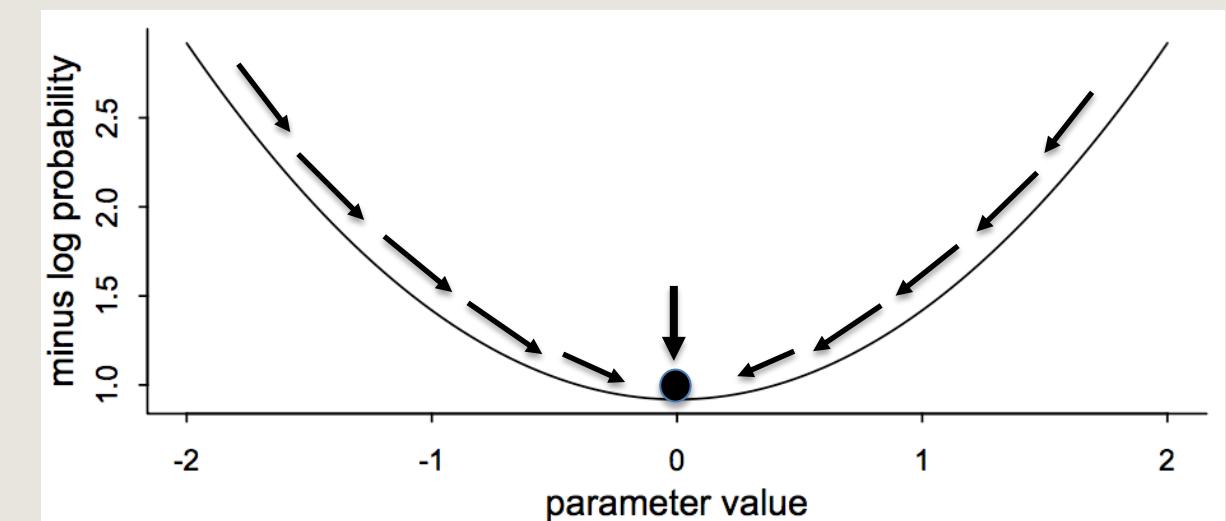


FIG 14. *The exploration of a probabilistic system is mathematically equivalent to the exploration of a physical system. For example, we can interpret the mode of the target density as a massive planet and the gradient of the target density as that planet's gravitational field. The typical set becomes the space around the planet through which we want a test object, such as a satellite, to orbit.*



Monte Carlo Hamiltoniano (HMC)

- Toda vez que el gradiente (la derivada multivariada) de la distribución posterior apunta en la dirección de máximo crecimiento de la función, es necesario complementar/contrarestar esta “gravedad” (energía potencial) con otra “fuerza” (energía cinética) que le permita “rodar”/“orbitar” (eternamente) por el conjunto típico (tomando muestras).

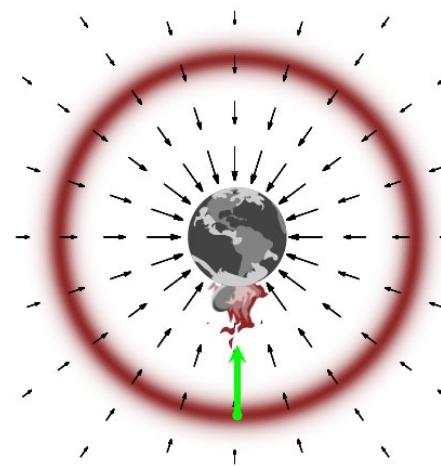
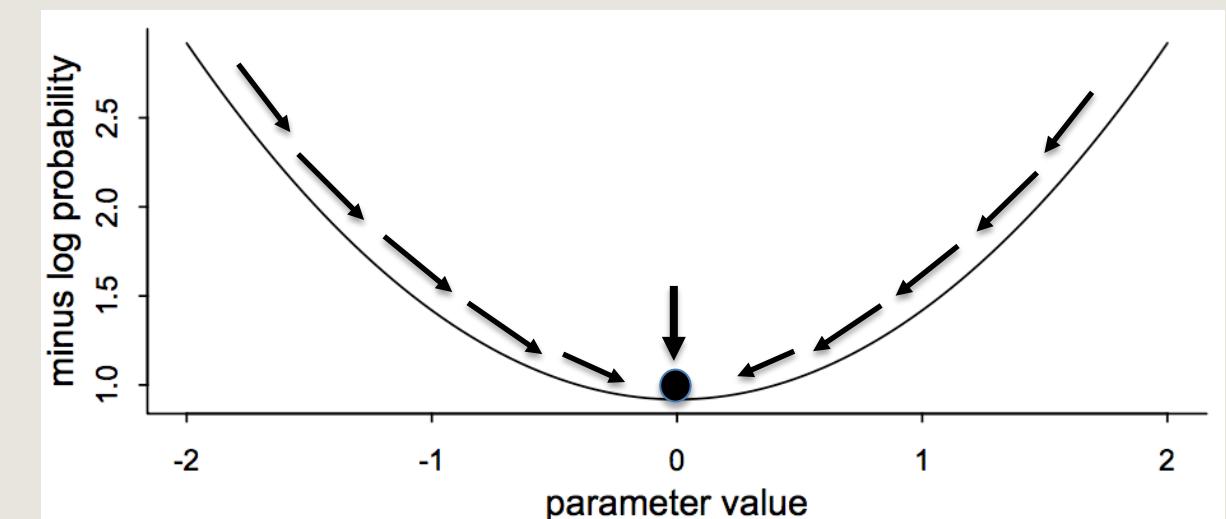


FIG 14. *The exploration of a probabilistic system is mathematically equivalent to the exploration of a physical system. For example, we can interpret the mode of the target density as a massive planet and the gradient of the target density as that planet's gravitational field. The typical set becomes the space around the planet through which we want a test object, such as a satellite, to orbit.*



Monte Carlo Hamiltoniano (HMC)

- Las propuestas se generan tomando aleatoriamente de la distribución de momento (fuerza con la que echamos a andar la canica, dimensión agregada al sistema) dejando “fluir” la “partícula” (rodar la canica) determinísticamente por un tiempo determinado (No U-turn Turn Sampler). La nueva posición es la propuesta a evaluar como siguiente paso en el algoritmo Metropolis.

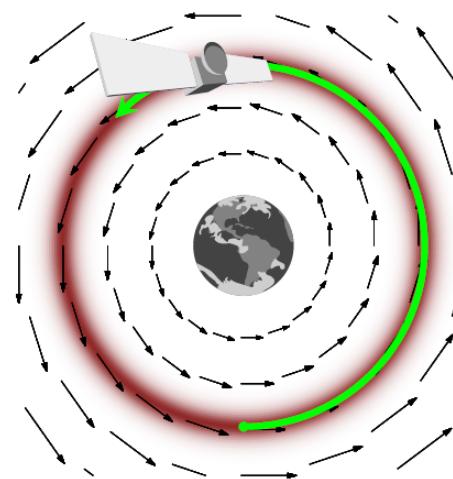
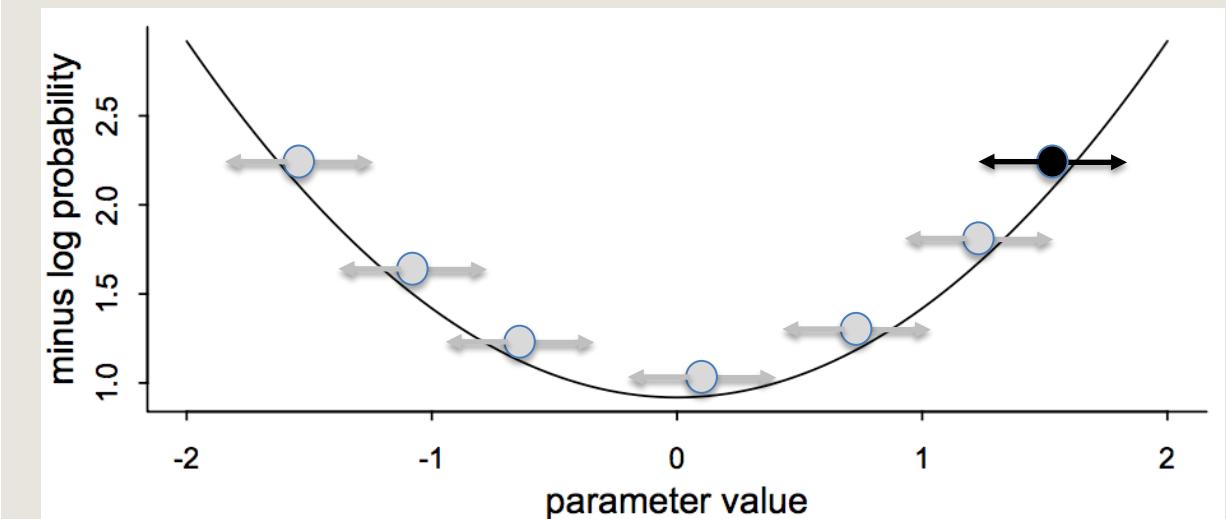
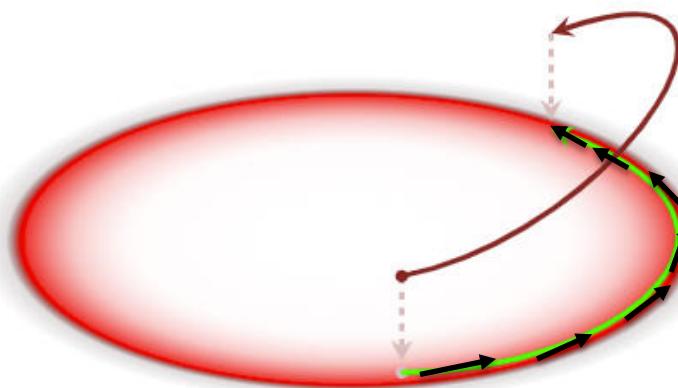


FIG 17. When we introduce exactly the right amount of momentum to the physical system, the equations describing the evolution of the satellite define a vector field aligned with the orbit. The subsequent evolution of the system will then trace out orbital trajectories.



Monte Carlo Hamiltoniano (HMC)

- Las propuestas se generan tomando aleatoriamente de la distribución de momento (fuerza con la que echamos a andar la canica, dimensión agregada al sistema) dejando “fluir” la “partícula” (rodar la canica) determinísticamente por un tiempo determinado (No U-turn Turn Sampler). La nueva posición es la propuesta a evaluar como siguiente paso en el algoritmo Metropolis.



Trajectories exploring the typical set of a probability distribution in phase space. That phase space distribution marginalizes to the target distribution, thus projecting us from the phase space to the target distribution's typical set.

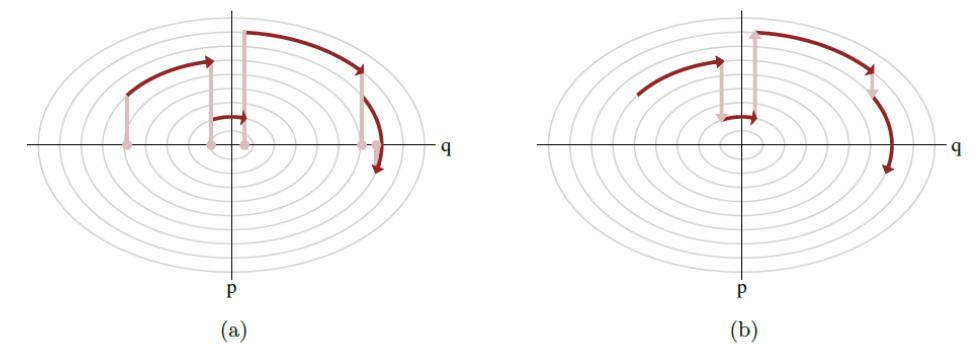


FIG 22. (a) Each Hamiltonian Markov transition lifts the initial state onto a random level set of the Hamiltonian, which can then be explored with a Hamiltonian trajectory before projecting back down to the target parameter space. (b) If we consider the projection and random lift steps as a single momentum resampling step, then the Hamiltonian Markov chain alternates between deterministic trajectories along these level sets (dark red) and a random walk across the level sets (light red).

Monte Carlo Hamiltoniano (HMC)

- Las propuestas se generan tomando aleatoriamente de la distribución de momento (fuerza con la que echamos a andar la canica, dimensión agregada al sistema) dejando “fluir” la “partícula” (rodar la canica) determinísticamente por un tiempo determinado (No U-turn Turn Sampler). La nueva posición es la propuesta a evaluar como siguiente paso en el algoritmo Metropolis.

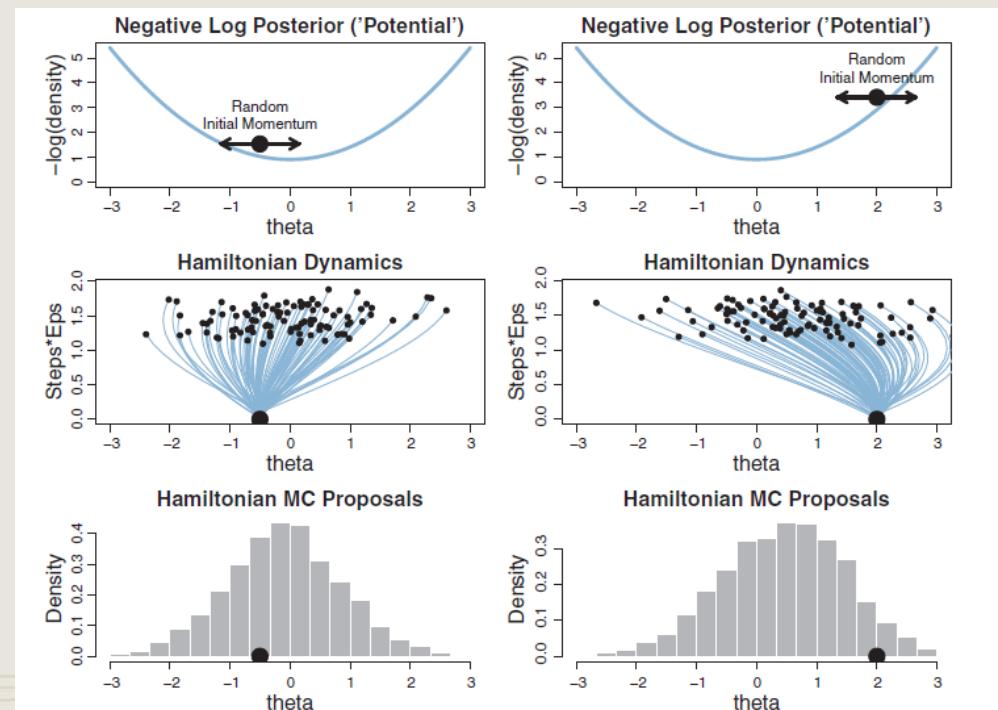


Figure 14.1 Examples of a Hamiltonian Monte Carlo proposal distributions. Two columns show two different current parameter values, marked by the large dots. First row shows posterior distribution. Second row shows the potential energy, with a random impulse given to the dot. Third row shows trajectories, which are the theta value (x-axis) as a function of time (y-axis marked Steps*Eps). Fourth row shows histograms of the proposals.



Monte Carlo Hamiltoniano (HMC)

- Aunque atractiva, HMC conduce en principio a una infinidad de posibles transiciones según la distribución de momento que se agregue al sistema, el tamaño y número de pasos (la fuerza con la se le pegue y el tiempo que se deje rodar la canica).
- Afortunadamente los desarrolladores de Stan (Stan Development Team. YEAR. Stan Modeling Language Users Guide and Reference Manual, VERSION. <https://mc-stan.org>), especialistas en geometría diferencial, trabajan constantemente en automatizar la combinación óptima de estas opciones para que los usuarios se puedan concentrar en la formulación de sus modelos estadísticos.



Stan interfaces with the most popular data analysis languages (R, Python, shell, MATLAB, Julia, Stata) and runs on all major platforms (Linux, Mac, Windows).

<https://mc-stan.org/>



Mirando el HMC

- La manera más fácil de entender cómo funcionan estos algoritmos es verlos trabajar. Se pueden probar diferentes simulaciones de MCMC escritas por Chi Feng [aquí](#).



OK Pero en la práctica qué pasa entonces?

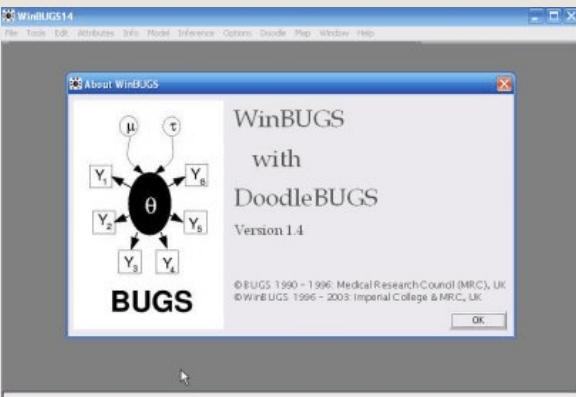
([Bayesian inference](#)
Using [Gibbs Sampling](#))

1987

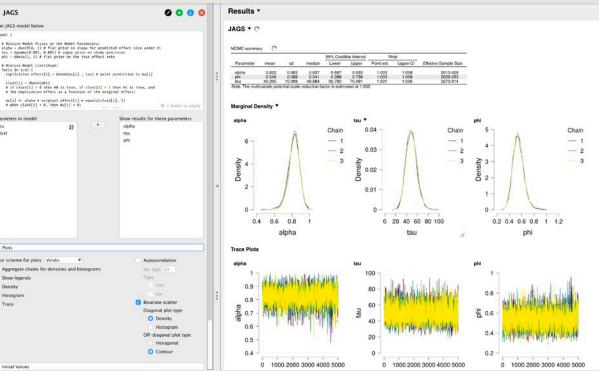
1997

2003

2010



JAGS is Just Another Gibbs Sampler



2010

Rjags
MCMCglmm
RbugsRwinbugs
...

Gibbs sigue dominando pero... años más tarde



Un
nuevo
orden?

¿Dónde se estima?

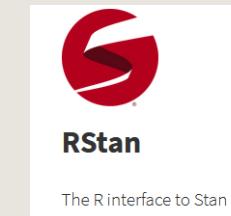
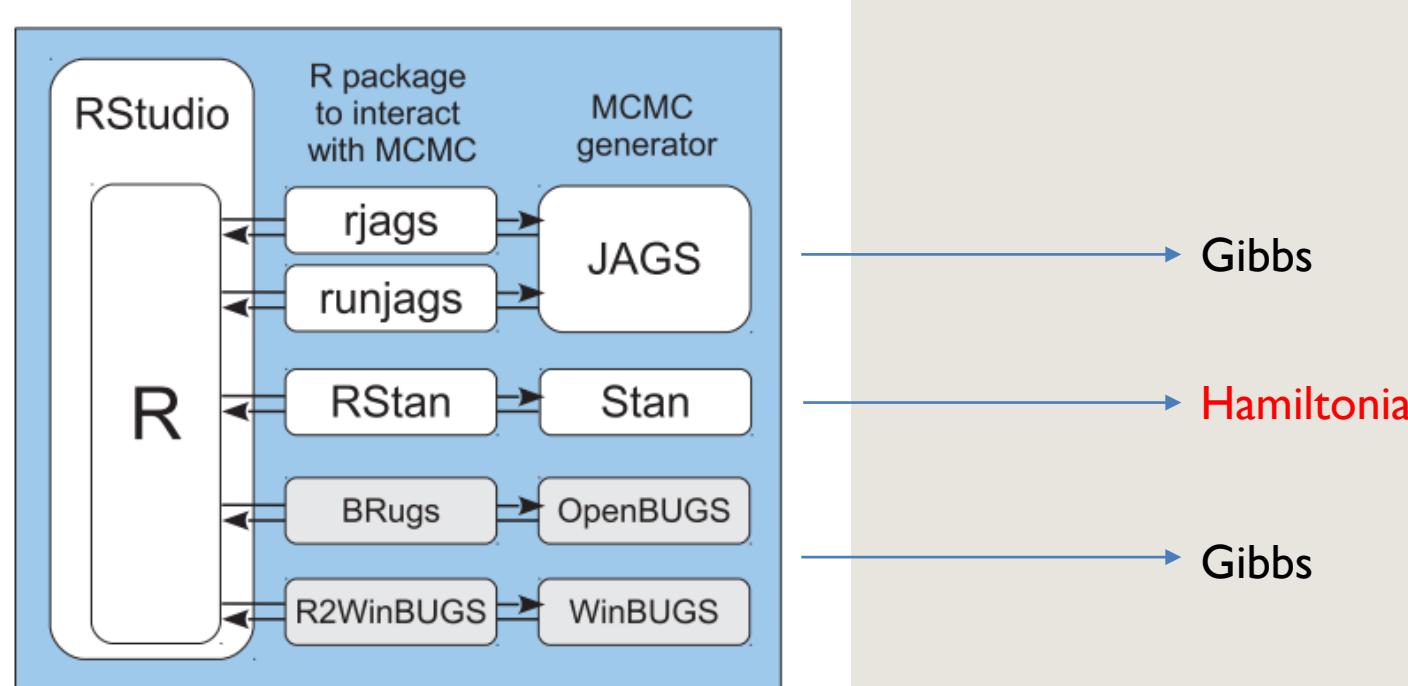


Figure 8.1: Relation of R programming language to other software tools. On the left, RStudio is an editor for interacting with R. The items on the right are various programs for generating MCMC samples of posterior distributions. The items in the middle are packages in R that interact with the MCMC generators. Copyright © Kruschke, J. K. (2014). *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. 2nd Edition. Academic Press / Elsevier.



Stan: Diferentes plataformas



Stan está escrito
en C++

Programas

rstan
PyStan
Stan.jl
CmdStan

Rstan (crudo)
rstan
(interfaces
para rstan)

rstanarm
rethinking
brms
cmdstanr

Avanzado: mayor potencia y flexibilidad

Intermedio: menor flexibilidad



Referencias

Geman, Stuart and Geman, Donald. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, PAMI-6(6):721–741, 1984.

Ghojogh, B., Nekoei, H., Ghojogh, A., Karray, F., & Crowley, M. (2020). Sampling algorithms, from survey sampling to Monte Carlo methods: Tutorial and literature review. arXiv preprint arXiv:2011.00901.

Hastings, W Keith. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1): 97–109, 1970.

Metropolis, Nicholas, Rosenbluth, Arianna W, Rosenbluth, Marshall N, Teller, Augusta H, and Teller, Edward. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Sitios de interés

<http://albertolumbreras.net/posts/maldicion-dimensionalidad.html>

<https://janosh.dev/blog/hmc-intro>

<https://chi-feng.github.io/mcmc-demo/>

Series y conferencias de Youtube

<https://youtu.be/ciM6wigZK0w>

<https://youtu.be/U561HGMVWjcw>