Introducción a Markov Chain Monte Carlo

Procesos Estocásticos

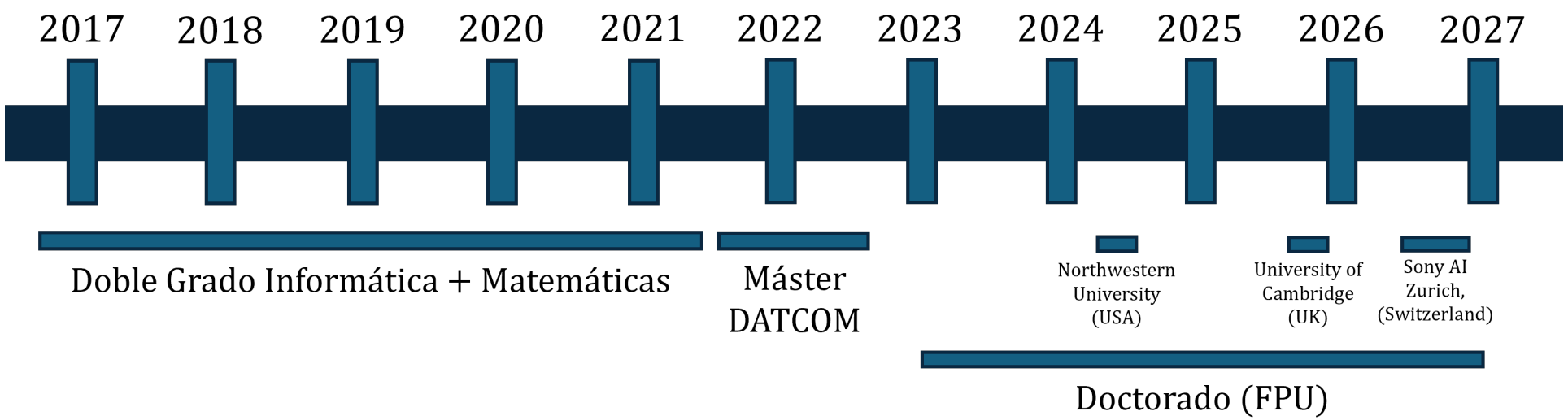
Universidad de Granada

Francisco Miguel Castro Macías

fcastro@ugr.es

Un poco de información sobre mí

Tesis doctoral: «Probabilistic machine learning methods for weakly supervised and inverse problems. Applications in medicine.»



Resumen

1. Muestreo de Distribuciones
2. Markov Chain Monte Carlo (MCMC)
3. Metropolis-Hastings (MH)
4. Comentarios Finales

Muestreo de Distribuciones

Muestreo de Distribuciones

$\mathcal{X} \equiv$ Espacio de estados (discreto, $\mathbb{R}^D$, un grafo, ...)

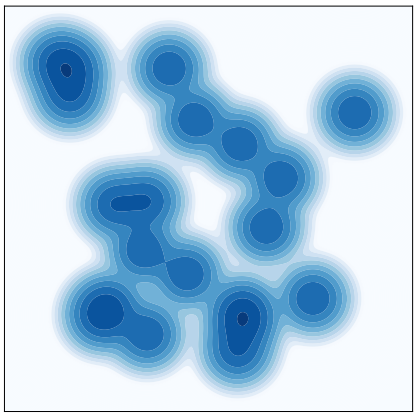
Distribución de probabilidad con función de densidad $\nu : \mathcal{X} \to \mathbb{R}^+$.

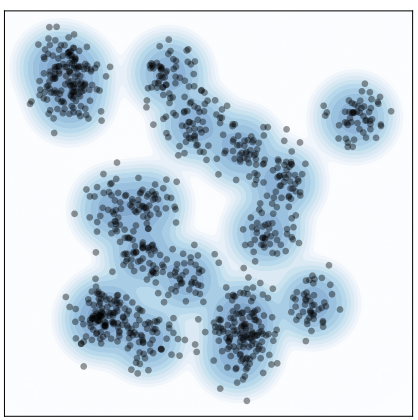
$$\nu(x) = Z^{-1}\widetilde{\nu}(x), \quad Z = \int_{\mathcal{X}} \widetilde{\nu}(x)\mathrm{d}x$$

Suponemos: podemos evaluar $\widetilde{\nu} : \mathcal{X} \to \mathbb{R}^+$ *eficientemente*, pero no podemos evaluar Z.

Objetivo 1 (muestrear): obtener muestras de ν: $X \sim \nu$.

Objetivo 2 (calcular esperanzas): dada una función $f : \mathcal{X} \to \mathbb{R}$, calcular $\mathbb{E}_{X\sim\nu}[f(X)]$.





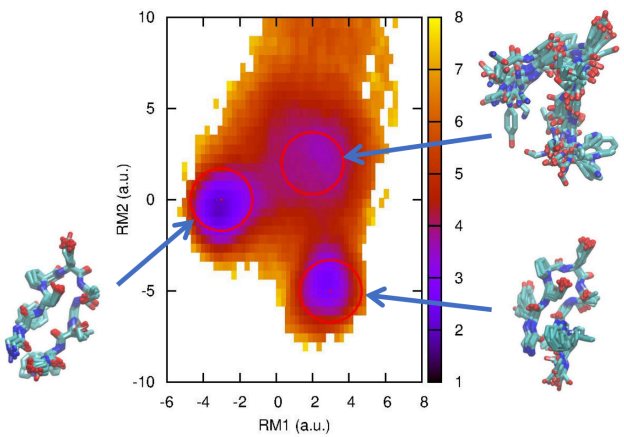
Muestreo de Distribuciones: ¿para qué?

Formulación física. $U : \mathcal{X} \to \mathbb{R}$ energía potencial; codifica la física del sistema.

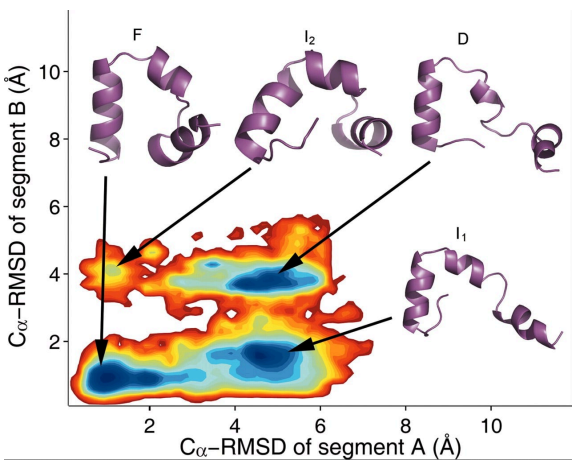
$$\nu(x) = \frac{\widetilde{\nu}(x)}{Z}, \quad \widetilde{\nu}(x) = \exp(-U(x)).$$

Molecular Dynamics.

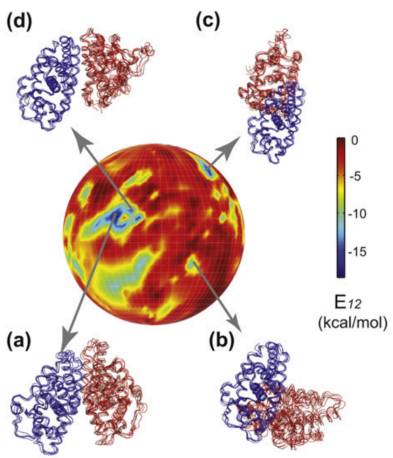
- $\mathcal{X}$ contiene configuraciones de una molécula: $\mathcal{X} = \mathbb{R}^{3\times N}$, $N \equiv$ número de átomos.
- $U(x)$ modela interacciones entre distintos átomos.
- $\mathbb{E}[f(X)] \equiv$ Velocidad de reacción, afinidad de unión, propiedades de materiales, ...



Fuente



Fuente



Fuente

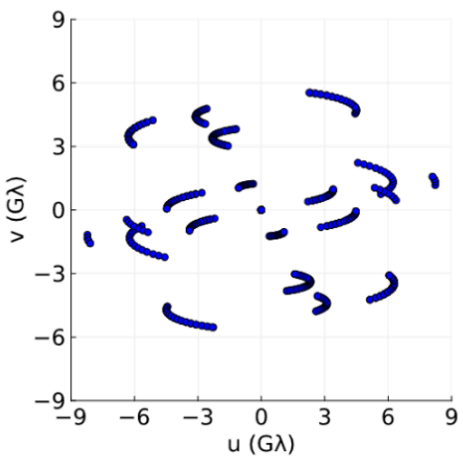
Muestreo de Distribuciones: ¿para qué?

Bayes. ν es una distribución *a posteriori*:

$$\nu(x) = p(x \mid y) = \frac{p(y \mid x)p(x)}{p(y)}.$$

Imaging Inverse Problems.

- $\mathcal{X}$ contiene imágenes: $\mathcal{X} = \mathbb{R}^{3\times H\times W}$.
- y es una señal *degradada* observada.
- x es la imagen limpia.
- $y = G(x) \to p(y \mid x)$.



Inference

M87, 2019

$\sim p(x \mid y)$

Imagen tomada de estas diapositivas. Vídeo explicativo.

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC)

Idea.

1. Construir una Cadena de Markov (MC) $\{X_t\}$ que sea *ν-invariante*.
2. Considerar el estimador Monte Carlo (MC):

$$S_T(f) = \frac{1}{T}\sum_{t=1}^{T} f(X_t).$$

3. Aproximar $\mathbb{E}_{X\sim\nu}[f(X)] \approx S_T(f)$ para T *muy grande*.

¿Por qué? Bajo *ciertas hipótesis* sobre $\{X_t\}$, se cumple:

$$\lim_{t\to\infty} \mathrm{Dist}(X_t) = \nu,$$

$$\lim_{T\to\infty} S_T(f) = \mathbb{E}_{X\sim\nu}[f(X)].$$

MCMC: Kernels de Markov

$(\mathcal{X}, \mathcal{F})$ espacio medible. $\mathcal{P}(\mathcal{X})$ conjunto de distribuciones de probabilidad sobre $\mathcal{X}$.

Un Kernel de Markov $K : \mathcal{X} \times \mathcal{X} \to [0, 1]$ es una función tal que:

1. Para cada $y \in \mathcal{X}$, $x \mapsto K(y, x)$ es medible.
2. Para cada $x \in \mathcal{X}$, $y \mapsto K(y, x) \in \mathcal{P}(\mathcal{X})$.

Probabilidad condicional.

> $K(y, x)$ es la densidad de la probabilidad condicional de Y dado $X = x$.

> *Instrucciones* para obtener $Y \sim K(\cdot, x)$ desde el estado $X = x$.

Ejemplo. $\mathcal{X} = \{x_1, ..., x_N\}$ discreto, podemos identificar:

$$K \equiv \left[K\left(x_i, x_j\right)\right]_{1 \leq i,j \leq N}.$$

¡Es la matriz de transición de una Cadena de Markov!

MCMC: Kernels de Markov

Los kernels **modifican distribuciones**: dada $\mu \in \mathcal{P}(\mathcal{X})$, definimos $K\mu \in \mathcal{P}(\mathcal{X})$ como:

$$(K\mu)(y) = \int_{\mathcal{X}} K(y, x)\mu(x)\mathrm{d}x.$$

Ejemplo. $\mathcal{X} = \{x_1, ..., x_N\}$ discreto:

$$K\mu = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & ... & K(x_1, x_N) \\ K(x_2, x_1) & K(x_2, x_2) & ... & K(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_N, x_1) & K(x_n, x_2) & ... & K(x_n, x_N) \end{bmatrix} \begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_N) \end{bmatrix}$$

Medidas invariantes. K es μ-invariante si $K\mu = \mu$.

MCMC: Kernels de Markov

Podemos **componer kernels**: para $t \geq 1$:

$$K^0\mu := \mu,$$
$$K^t\mu := K\big(K^{t-1}\mu\big), \quad t \geq 1$$

Ejemplo. $\mathcal{X} = \{x_1, ..., x_N\}$ discreto:

$$K^t\mu = \begin{bmatrix} K(x_1,x_1) & K(x_1,x_2) & ... & K(x_1,x_N) \\ K(x_2,x_1) & K(x_2,x_2) & ... & K(x_2,x_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_N,x_1) & K(x_n,x_2) & ... & K(x_n,x_N) \end{bmatrix}^t \begin{bmatrix} \mu(x_1) \\ \vdots \\ \mu(x_N) \end{bmatrix}$$

Medidas invariantes. Si K es μ-invariante, entonces $K^t\mu = \mu$.

MCMC: Kernels $\longleftrightarrow$ Cadenas de Markov

Dado un kernel K, construimos una Cadena de Markov $(X_0, X_1, ...)$ *aplicando* K:

$$X_0 \sim \mu_{\text{ref}}, \quad X_t \sim K(\cdot, X_{t-1}).$$

¿Y al revés?

> Si la cadena es homogénea, las transiciones definen el kernel.

> Si no, el kernel depende del tiempo: $K_t(y, x)$.

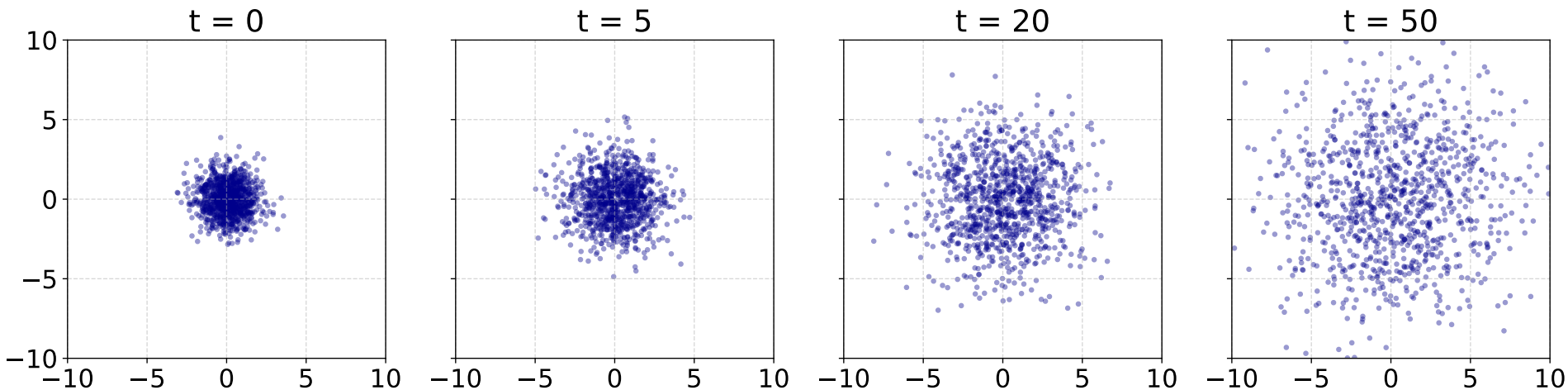
Evolución de la distribución.

$$\text{Dist}(X_t) = K^t \mu_{\text{ref}}$$

MCMC: Random Walk (RW)

$$\mathcal{X} = \mathbb{R}^2, \quad K(y, x) = \mathcal{N}(y \mid x, \sigma^2 I),$$

$$X_t \sim K(\cdot, X_{t-1}) \Leftrightarrow X_t = X_{t-1} + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I).$$



$$\mathrm{Dist}(X_t) = K^t \mu_{\mathrm{ref}} = \mathcal{N}(0, t\sigma^2 I)$$

MCMC: Kernels invariantes

Recuerda. Queremos muestrear de $\nu \in \mathcal{P}(\mathcal{X})$.

Estrategia.

1. Diseñamos un kernel K que sea ν-invariante.
2. Tomamos $X_0 \sim \mu_{\text{ref}}$.
3. Aplicamos K en cada paso: $X_t \sim K(\cdot, X_{t-1})$.

¿Cómo diseñar kernels invariantes? Si K cumple la siguiente condición:

$$K(y,x)\nu(x) = K(x,y)\nu(y), \quad \forall x, y \in \mathcal{X}, \qquad \text{(Balance Detallado)}$$

entonces K es ν-invariante. Prueba:

$$(K\nu)(y) = \int_{\mathcal{X}} K(y,x)\nu(x)\mathrm{d}x = \int_{\mathcal{X}} K(x,y)\nu(y)\mathrm{d}x =$$

$$= \nu(y) \int_{\mathcal{X}} K(x,y)\mathrm{d}x = \nu(y).$$

MCMC: Convergencia

Recuerda. Queremos muestrear de $\nu \in \mathcal{P}(\mathcal{X})$.

Estrategia.

1. Diseñamos un kernel K que sea ν-invariante.
2. Tomamos $X_0 \sim \mu_{\text{ref}}$.
3. Aplicamos K en cada paso: $X_t \sim K(\cdot, X_{t-1})$.

¿Llegaremos en algún momento a ν? Es decir, ¿la distribución de X_t será ν para algún t?

No se aleja. Para cualquier $\mu \in \mathcal{P}(\mathcal{X})$:

$$\text{Distancia}(\nu, K\mu) \leq \text{Distancia}(\nu, \mu).$$

Convergencia. Bajo *ciertas condiciones*:

$$\lim_{t\to\infty} \text{Distancia}(\nu, K^t\mu) = 0, \qquad \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} f(X_t) = \mathbb{E}_{X\sim\nu}[f(X)].$$

Metropolis-Hastings (MH)

Metropolis-Hastings

THE JOURNAL OF CHEMICAL PHYSICS VOLUME 21, NUMBER 6 JUNE, 1953

Equation of State Calculations by Fast Computing Machines

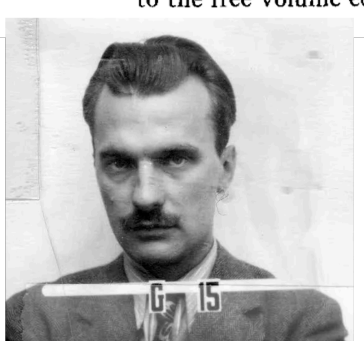
NICHOLAS METROPOLIS, ARIANNA W. ROSENBLUTH, MARSHALL N. ROSENBLUTH, AND AUGUSTA H. TELLER,
Los Alamos Scientific Laboratory, Los Alamos, New Mexico

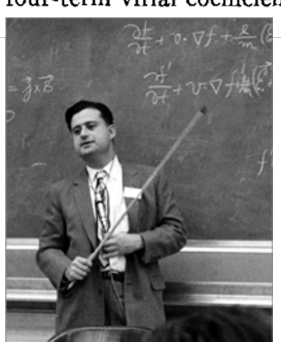
AND

EDWARD TELLER,* *Department of Physics, University of Chicago, Chicago, Illinois*

(Received March 6, 1953)

A general method, suitable for fast computing machines, for investigating such properties as equations of state for substances consisting of interacting individual molecules is described. The method consists of a modified Monte Carlo integration over configuration space. Results for the two-dimensional rigid-sphere system have been obtained on the Los Alamos MANIAC and are presented here. These results are compared to the free volume equation of state and to a four-term virial coefficient expansion.




Fuente

Metropolis-Hastings (MH)

Recuerda. Queremos muestrear de $\nu \in \mathcal{P}(\mathcal{X})$.

Idea. Proponer movimientos. Aceptarlos / Rechazarlos para preservar ν.
> $Q(y, x)$ es un kernel que podemos evaluar y muestrear.

Kernel MH. Dada la posición actual X_{t-1}:
1. Proponer: $Y_t \sim Q(\cdot, X_{t-1})$.
2. Calcular la «probabilidad de aceptación»:

$$A(X_{t-1}, Y_t) = \frac{\nu(Y_t)Q(X_{t-1}, Y_t)}{\nu(X_{t-1})Q(Y_t, X_{t-1})} = \frac{\widetilde{\nu}(Y_t)Q(X_{t-1}, Y_t)}{\widetilde{\nu}(X_{t-1})Q(Y_t, X_{t-1})}$$

3. Aceptar / Rechazar:

$$X_t = \begin{cases} Y_t & \text{con probabilidad } \min(1, A(X_{t-1}, Y_t)) \\ X_{t-1} & \text{en otro caso} \end{cases}$$

Random Walk Metropolis-Hastings (RWMH)

$$\mathcal{X} = \mathbb{R}^D, \quad Q(y, x) = \mathcal{N}(y \mid x, \Sigma).$$

La probabilidad de aceptación queda:

$$A(X_{t-1}, Y_t) = \frac{\nu(Y_t)Q(X_{t-1}, Y_t)}{\nu(X_{t-1})Q(Y_t, X_{t-1})} = \frac{\widetilde{\nu}(Y_t)}{\widetilde{\nu}(X_{t-1})},$$

Ejemplo. https://www.saifsyed.com/sampling-demo/app.html

Metropolis Adjusted Langevin Algorithm (MALA)

Las propuestas de RWMH: $Y = X + \varepsilon$.

> La mayoría de los movimientos propuestos son rechazados.

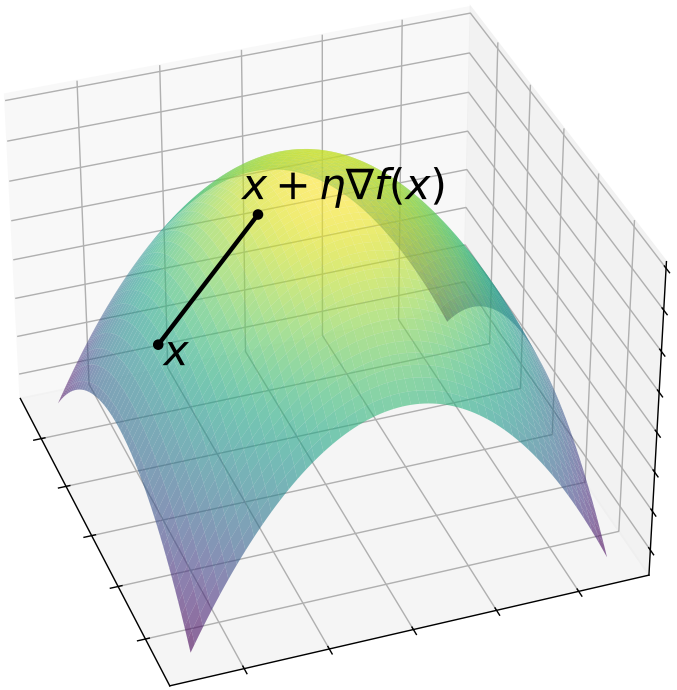
> No usa información sobre ν.

Langevin Proposal. Proponer movimientos hacia zonas de alta probabilidad:

$$Y = X + \frac{\eta}{2}\nabla \log \nu(X) + \sqrt{\eta}\varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I)$$

$$Q(y, x) = \mathcal{N}\left(y \mid x + \frac{\eta}{2}\nabla \log \nu(x), \eta I\right)$$

Ejemplo. https://www.saifsyed.com/sampling--demo/app.html



Comentarios Finales

Comentarios Finales

- Hemos visto la *punta del iceberg*.
- Queda mucho más...
  - Multimodality and mode trapping
  - Mixing time
  - Annealing-based methods
  - Convergence diagnostics
  - Adaptative MCMC
  - ...
- Si queréis más info: `fcastro@ugr.es`

Recursos.

- Estas diapositivas están basadas en la introducción al curso «Scalable Sampling» (STAT 547E) de la University of British Columbia (Link).
- Unas notas formalizando todo lo anterior y mucho más (Link).