Table of Contents

Semana 4

Métodos de muestreo

Muestreo simple

Estimación basada en el muestreo

$D = \{x[1], \ldots, x[M]\}$ muestras identicamente distribuidas de P

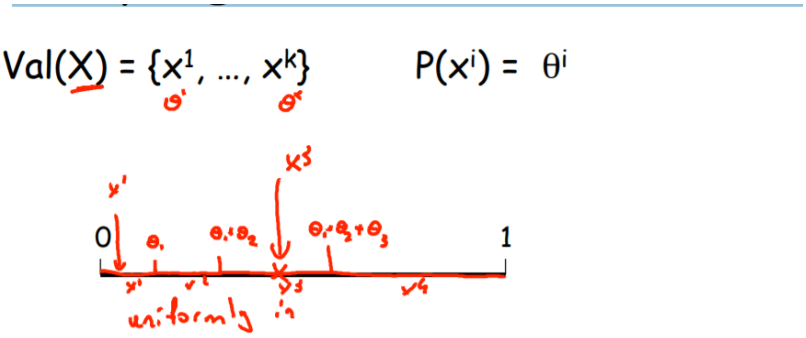
si $P(X = 1) = p$, la estipacion para p es

$$T_{\mathcal{D}} = \frac{1}{M} \sum_{m=1}^{M} x[m]$$

Más generalmente, para cualquier distribución P, función F:

$$\boldsymbol{E}_P[f] \approx \frac{1}{M}\sum_{m=1}^{M} f(x[m])$$

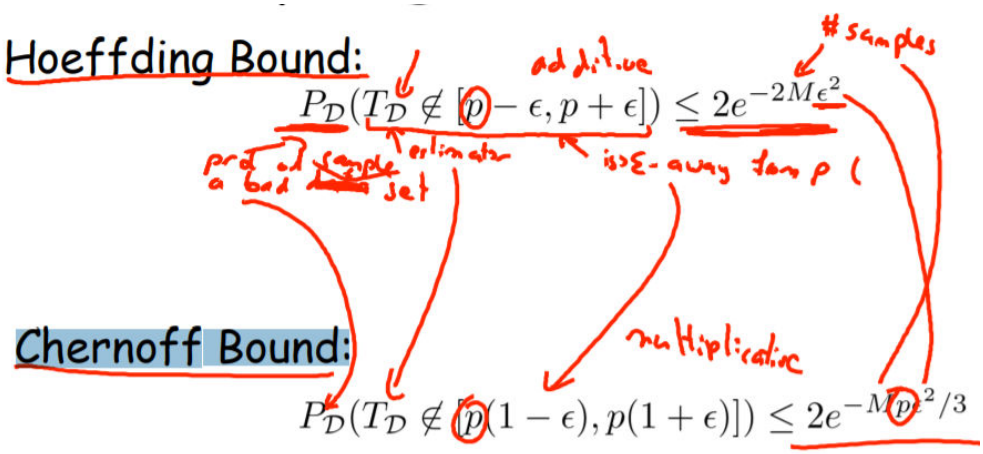
Muestreo de distribución discreta.



es transformada inverza para distribuciónes discetas

En teoría de la probabilidad, la **desigualdad de Hoeffding** proporciona una cota superior a la probabilidad de que la suma de variables aleatorias se desvíe una cierta cantidad de su valor esperado

En la teoría de probabilidad, las **Cotas de Chernoff** fueron nombradas luego de su presentación por Herman Chernoff y, gracias a Herman Rubin,1 se dieron cotas exponencialmente decrecientes para las distribuciones de sumas de variables aleatorias independientes.



Hoeffding Bound

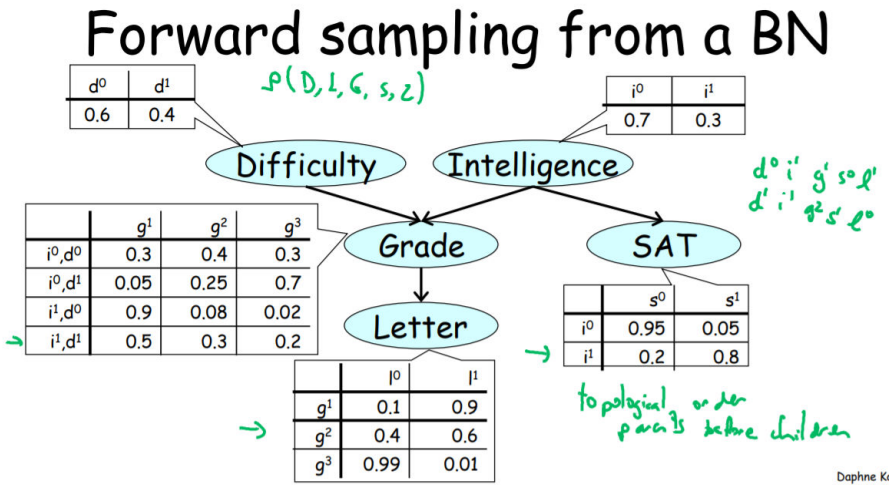
Para un limte aditivo ε en un error con probabilidad $> 1 - \delta$

$$M \geq \frac{\ln(2/\delta)}{2\epsilon^2}$$

Chernoff Bound:

Para un limte multiplicativo ε en un error con probabilidad $> 1 - \delta$

$$M \geq 3\frac{\ln(2/\delta)}{p\epsilon^2}$$



Muestreo hacia adelante para consultas

• Meta: Estimar P (y = y)

- Generar muestras de BN

- Compute fracción donde $Y = y$

Consultas con evidencia

• Meta: Estime P (Y = y | E = e)

• Algoritmo de **muestreo de rechazo**

- Generar muestras de BN

- tirar todos los que E ≠ e

- Compute fracción donde Y = y

Fracción esperada de muestras mantenidas ~ P (e)

muestras necesarias filas exponencialmente con # de variables observadas

Resumen

• Generar muestras de un BN es fácil

• (ϵ, δ) existen, pero la utilidad es limitada:

- límites aditivos: inútiles para eventos de baja probabilidad

- límites multiplicativos: # de muestras crecen como $\frac{1}{p(y)}$

• Con evidencia, # de muestras requeridas crece exponencialmente con # de variables observadas

• Muestreo hacia adelante generalmente inviable para MNS

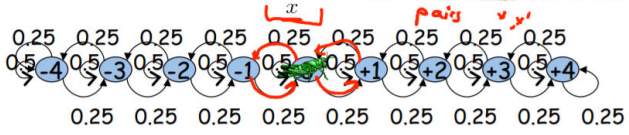
Cadena de markov monte carlo

Cadena de Markov

Una cadena de Markov define un modelo de transición probabilístico $T(x \rightarrow x')$ sobre todo los estados x

- para todo x $\sum_{x'} T(x \rightarrow x') = 1$

$$P^{(t+1)}(X^{(t+1)} = x') = \sum_{x} P^{(t)}(X^{(t)} = x) T(x \rightarrow x')$$

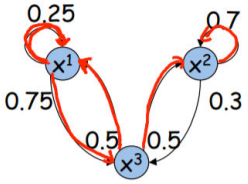


	-2	-1	0	+1	+2
$P^{(0)}$	0	0	1	0	0
$P^{(1)}$	0	.25	.5	.25	0
$P^{(2)}$	$.25^2$= .0625	2×(.5×.25) = .25	$.5^2$+2×$.25^2$ = .375	2×(.5×.25) = .25	$.25^2$= .0625

Distribución estacionaria (Probabilidad de estado estable)

$$P^{(t)}(x') \approx P^{(t+1)}(x') = \sum_x P^{(t)}(x)T(x \to x')$$

$$\pi(x') = \sum_x \pi(x)T(x \to x')$$



$$\begin{aligned}
\pi(x^1) &= 0.25\pi(x^1) + 0.5\pi(x^3) \\
\pi(x^2) &= 0.7\pi(x^2) + 0.5\pi(x^3) \\
\pi(x^3) &= 0.75\pi(x^1) + 0.3\pi(x^2)
\end{aligned}$$

$$\pi(x^1) + \pi(x^2) + \pi(x^3) = 1$$

$$\begin{aligned}
\pi(x^1) &= 0.2 \\
\pi(x^2) &= 0.5 \\
\pi(x^3) &= 0.3
\end{aligned}$$

Daphne Koller

Cadenas de Markov regulares

Una cadena de Markov es regular si existe k de tal manera que, para cada x, x ', la probabilidad de pasar de x a x' en exactamente k pasos es> 0

Teorema: una cadena regular de Markov converge a una distribución estacionaria única, independientemente del estado inicial

Condiciones suficientes para la regularidad:

- Cada dos estados están conectados.
- Para cada estado, hay una auto-transición.

Usando una cadena de markov

meta: calcular $P(x \in S)$

- pero P es demasiado difícil de muestrear directamente

Construir una cadena de Markov T cuya distribución estacionaria única es P

Muestrear $X^{(0)}$ de algun $P^{(0)}$

Para T = 0, 1, 2, ...

- Generar $x^{(x+1)}$ de $T(x^t \to x')$

Solo queremos usar muestras que se muestrean de una distribución similar a P original

En iteraciones tempranas, $P^{(t)}$ suele estar lejos de P

Comience a recolectar muestras solo después de que la cadena se haya ejecutado lo suficiente para "mezclar"

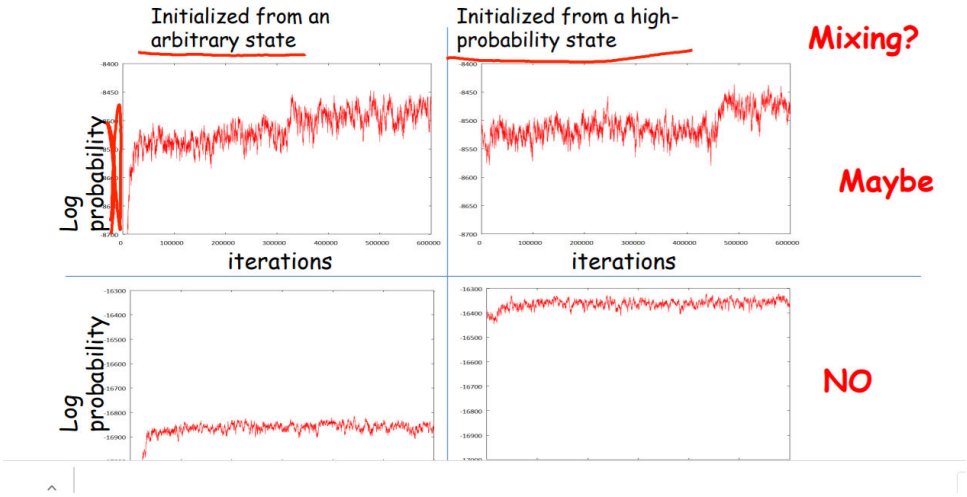
Mezcla

¿Cómo sabes si una cadena ha mezclado o no?

- En general, nunca puedes "probar" 2ue una cadena ha mezclado
- Pero en muchos casos pueden mostrar que no ha mezclado

¿Cómo sabes que una cadena no ha mezclado?

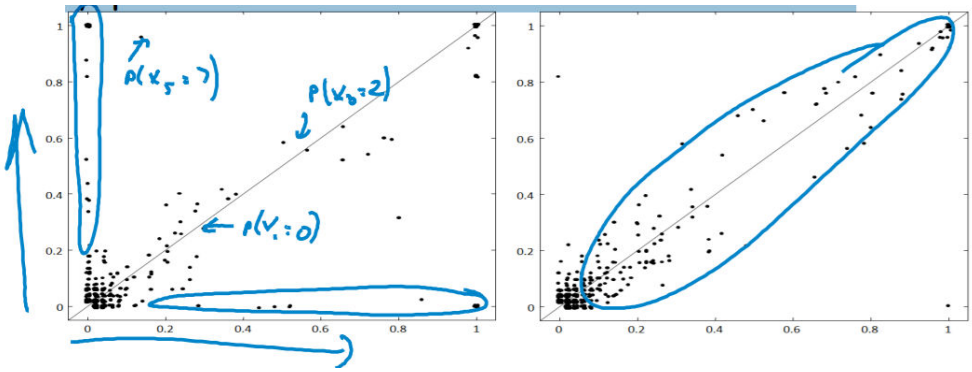
- Compare estadísticas de cadena en diferentes ventanas dentro de una sola corrida de la cadena
- y a través de diferentes corridas que se inicializaron de manera diferente



Cada punto es una estadística (por ejemplo, p (x ∈ s))

• La posición X es su valor estimado de la cadena 1

• La posición Y es su valor estimado de la cadena 2



Usando las muestras

Una vez que la cadena se mezcla, todas las muestras $P^{(t)}$ son de la distribución estacionaria π

- asi que podemos (y deberíamos) usar todos los x (t) para $t > t_{\text{mix}}$

Sin embargo, las muestras cercanas están correlacionadas!

- - Entonces, no debemos sobreestimar la calidad de nuestra estimación simplemente contando muestras

Cuanto más rápido se mezcla una cadena, menos correlacionada (más útiles) son las muestras

Algoritmo P1

Para c=1,...,C

- Muestrea $x^{(c,0)}$ de $P^{(0)}$

Repite hasta el muestreo

- Para c=1,...,C
- • Generar $x^{(c,t+1)}$ de $T(x^{(c,t)} \rightarrow x')$
- • • Compare estadísticas de la ventana en diferentes cadenas para determinar la mezcla.
- • • $t := t + 1$

Algoritmo P2

Repita hasta tener suficientes muestras

- $D := \varnothing$
- Para c=1,...,C
- • Generar $x^{(c,t+1)}$ de $T(x^{(c,t)} \rightarrow x')$
- • $D := D \cup \{x^{(c,t+1)}\}$
- • $t := t + 1$

Sea $D = \{x[1], \ldots, x[M]\}$, estimar

$$\boldsymbol{E}_P[f] \approx \frac{1}{M} \sum_{m=1}^{M} f(x[m])$$

Resumen

Pros:

- propósito muy general
- a menudo fácil de implementar
- buenas garantías teóricas como t → ∞

Contras:

- Muchos parámetros sintonizables / opciones de diseño.
- Puede ser bastante lento para converger
- Difícil decir si está funcionando.

Muestreo de gibbs

Cadena de gibbs

Objetivo de distribución $P_\Phi(X_1, \ldots, X_n)$

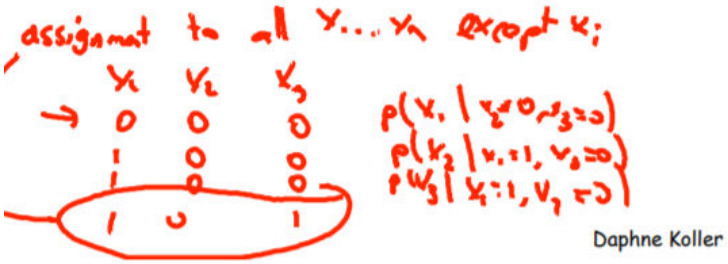
Espacio de estado de la cadena de Markov: asignaciones completas x a $X = \{X_1, \ldots, X_n\}$

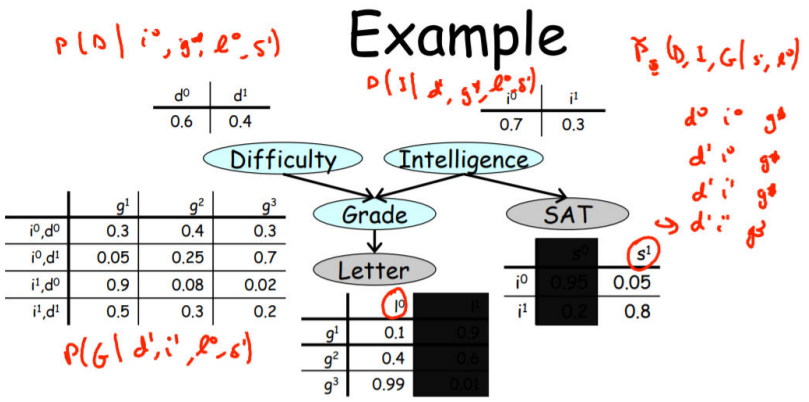
Modelo de transición dado el estado inicial x:

Para i=1,...,n

- muestrear $x_i \sim P_\Phi(X_i|x_{-i})$

Establecer x'=x





Costo computacional

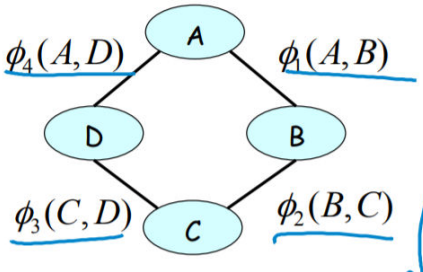
Para i=1,...,n

- muestrear $x_i \sim P_\Phi(X_i|x_{-i})$

$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

tomamos la regla de bayes

$$P(A_i|B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(B \cap A_i))}{\sum_{k=1}^{n} P(B|A_k)P(A_k)}$$



$$P_\Phi(A \mid b,c,d) = \frac{\tilde{P}_\Phi(a,b,c,d)}{\sum_{A'} \tilde{P}_\Phi(A',b,c,d)}$$

$$\frac{\phi_1(A,b)\phi_2(b,c)\phi_3(c,d)\phi_4(A,d)}{\sum_{A'} \phi_1(A',b)\phi_2(b,c)\phi_3(c,d)\phi_4(A',d)}$$

normalizing constant

$\propto \phi_1(A,b)\,\phi_4(A,d)$

factors that involve A

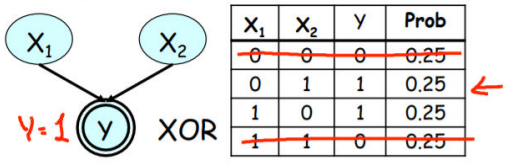
$$P_\Phi(X_i \mid \boldsymbol{x}_{-i}) = \frac{P_\Phi(X_i, \boldsymbol{x}_{-i})}{P_\Phi(\boldsymbol{x}_{-i})} = \frac{\tilde{P}_\Phi(X_i, \boldsymbol{x}_{-i})}{\tilde{P}_\Phi(\boldsymbol{x}_{-i})}$$

only X_i and its neighbors

$$\propto \prod_{j: X_i \in Scope[C_j]} \phi_j(X_i, \boldsymbol{x}_{j,-i})$$

factors that involve X_i

Gibbs cadena y regularidad



X_1	X_2	Y	Prob
~~0~~	~~0~~	~~0~~	~~0.25~~
0	1	1	0.25
1	0	1	0.25
~~1~~	~~1~~	~~0~~	~~0.25~~

- Si todos los factores son positivos, la cadena Gibbs es regular
- Sin embargo, la mezcla puede ser muy lenta.

Resumen

Convierte el problema difícil de la inferencia a una secuencia de pasos de muestreo "fáciles"

• Pros:

- Probablemente la cadena de Markov más simple para PGMS.
- Computacionalmente eficiente para probar.

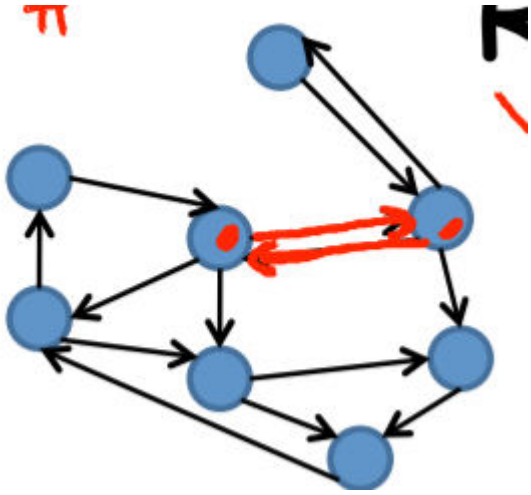
• Contras:

- a menudo lento para mezclar, especialmente Cuando las probabilidades están alcanzadas en su punto máximo
- Solo se aplica si podemos probar el producto de los factores.

Algoritmo de Hastings Metropolis

Cadenas reversibles

Una cadena es reversible si $\pi\ (x)T(x \to x') = \pi\ (x')T(x' \to x)$ (balance detallado)



Teorema: si el balance detallado se mantiene, y T es regular, entonces T tiene una distribución estacionaria única π

$$\sum_x \pi\ (x)T(x \to x') = \sum_x \pi\ (x')T(x' \to x)$$

$$\sum_x \pi\ (x)T(x \to x') = \pi(x')$$

y esto es la definicion de la distribución estacionaria

Cadena de Hastings Metropolis

distribución propuesta $Q(x \to x')$

Probabilidad de aceptación $A(x \to x')$

• En cada estado x, muestrear x' de $Q(x \to x')$

• Aceptar la propuesta con probabilidad $A(x \to x')$

- Si se aceptan propuestas, muévete a x'
- De lo contrario, quédate en x

Entonces tenemos lo siguiente

$$T(x \to x') = Q(x \to x')A(x \to x') \quad \text{si } x \neq x'$$
$$T(x \to x') = Q(x \to x') + \sum_{x' \neq x} Q(x \to x')(1 - A(x \to x')) \quad \text{si } x = x'$$

Probabilidad de aceptación

$$\pi(x)T(x \to x') = \pi(x')T(x' \to x)$$
$$\pi(x)Q(x \to x')A(x \to x') = \pi(x')Q(x' \to x)A(x' \to x)$$

$$\frac{A(x \to x')}{A(x' \to x)} = \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')} = p < 1$$

$$A(x \to x') = \min\left[1, \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')}\right]$$

Elección de Q

$$A(x \to x') = \min\left[1, \frac{\pi(x')Q(x' \to x)}{\pi(x)Q(x \to x')}\right]$$

Q debe ser reversible:

- $Q(x \to x') > 0 \Rightarrow Q(x' \to x) > 0$

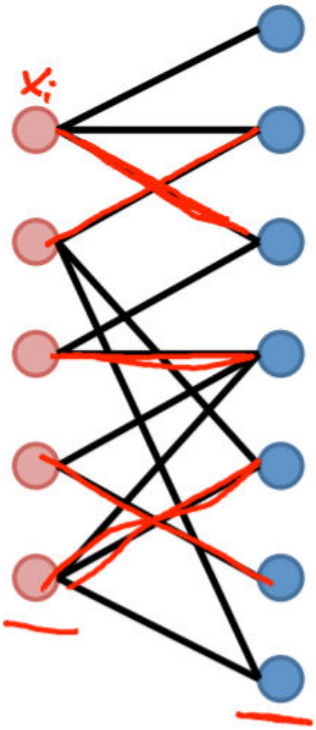
• Fuerzas opositoras

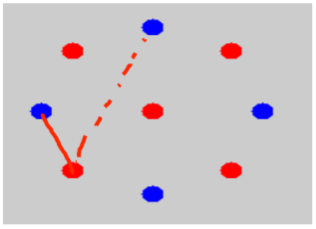
- Q debería intentar extenderse, para mejorar la mezcla
- Pero entonces la probabilidad de aceptación a menudo baja

MCMC para emparejamiento

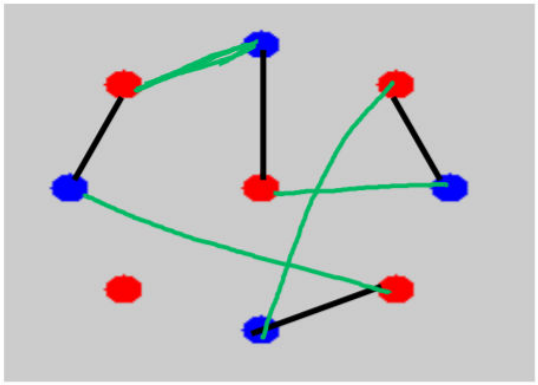
X_i = j si i coincide con j

$$P(X_1 = \nu_1, \ldots, X_4 = \nu_4) \propto \begin{cases} \exp\left(-\sum \text{dist}(i, \nu_i)\right) & \text{Si cada } X_i \text{ tiene un valor diferente} \\ 0 & \text{cuualquier otro caso} \end{cases}$$





Ruta de aumento

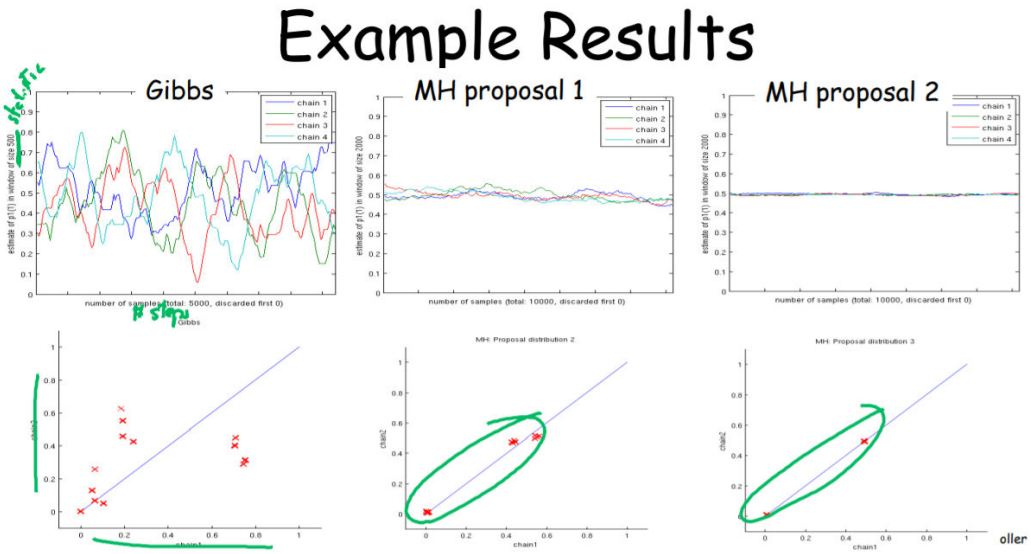


1) Elige aleatoriamente una variable X_i

2) muestrea X_i, fingiendo que todos los valores están disponibles

3) Elija la variable cuya asignación fue tomada (conflicto) y volver al paso 2

• Cuando el paso 2 no crea ningún conflicto, modifique la asignación a la ruta de aumento de flip



Resumen

MH es un marco general para construir cadenas de Markov con una distribución estacionaria particular

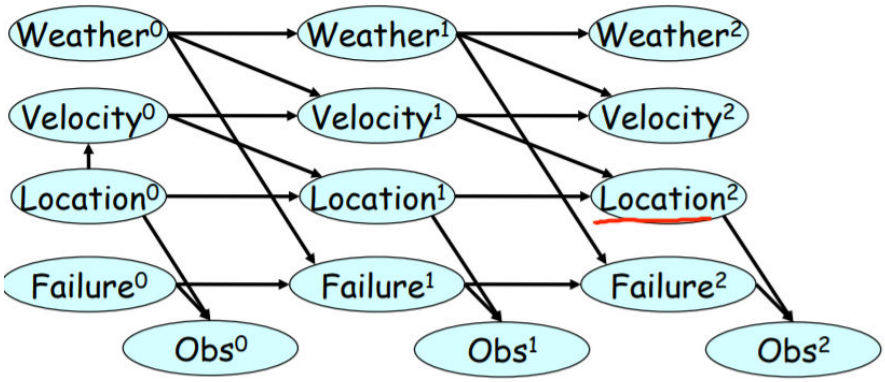
- Requiere una distribución de propuestas.
- Aceptación calculada a través de balance detallado.

• Flexibilidad tremenda en el diseño de distribuciones de propuestas que exploran el espacio rápidamente.

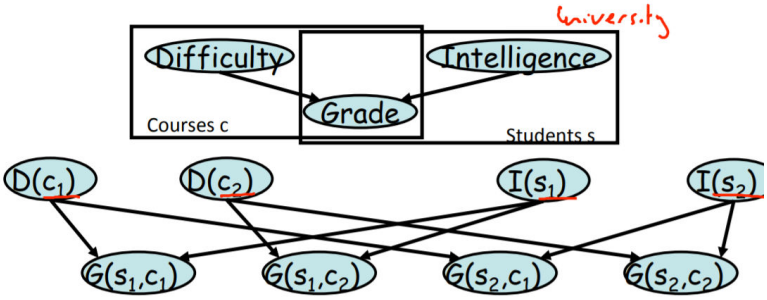
- Pero la distribución de la propuesta hace una gran diferencia.
- y encontrar uno bueno no siempre es fácil

Inferencia en modelos de plantilla

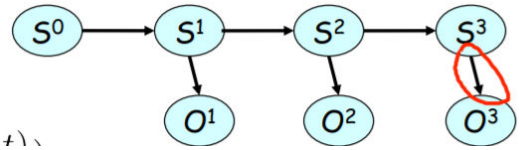
Puede "desenrollar" DBN para una trayectoria dada y ejecutar inferencia sobre la red de "tierra"



Lo mismo aplica para los modelos placa



Seguimiento del estado de creencias



$$\sigma^{(t)}(S^{(t)}) = P(S^{(t)}|o^{(1:t)})$$

$$\begin{aligned}\sigma^{(t+1)}(S^{(t+1)}) &= P(S^{(t+1)}|o^{(1:t)}) \\ &= \sum_{S^{(t)}} P\left(S^{(t+1)}|S^{(t)}, o^{(1:t)}\right) P(S^{(t)}|o^{(1:t)}) \\ &= \sum_{S^{(t)}} P\left(S^{(t+1)}|S^{(t)}, o^{(1:t)}\right) \sigma^{(t)}(S^{(t)})\end{aligned}$$

$$\begin{aligned}\sigma^{(t+1)}(S^{(t+1)}) &= P(S^{(t+1)}|o^{(1:t)}, o^{(t+1)}) \\ &= \frac{P(o^{(t+1)}|S^{(t+1)}, o^{(1:t)}) P(S^{(t+1)}|o^{(1:t)})}{P(o^{(t+1)}|, o^{(1:t)})} \\ &= \frac{P(o^{(t+1)}|S^{(t+1)}) \sigma^{(t+1)}(S^{(t+1)})}{P(o^{(t+1)}|, o^{(1:t)})}\end{aligned}$$

Resumen

- La inferencia en la plantilla y los modelos temporales se pueden realizar desenrollando la red de tierra y utilizando métodos estándar.
- Los modelos temporales también plantean nuevas tareas de inferencia, como seguimiento en tiempo real, que requieren que adaptemos nuestros métodos.
- Además, la red de tierra a menudo es grande y densamente conectada, lo que requiere un diseño y uso cuidadoso del algoritmo de métodos aproximados

Resumen total

Map vs marginals

marginals:

- menos frágil
- confianza en las respuestas
- apoya la toma de decisiones

- los errores a menudo son atenuados (disminuidos), obtienes respuestas más solidas en inferencia aproximada

Map:

- si tratamos de calcular una asignacion conjunta coherente
- Tiene una gama de modelos más manejables (más eficientes)
- Proporciona algunas garantias teoricas
- capacidad para medir si el algorimo esta trabajando

Algoritmos para marginales

- inferencia exacta (para problema pequeño)
- Algorimo de propagacion de creencias
- metodo de muestreo

algoritmos para Map

- inferencia exacta
- metodos de optimizacion
- --- aproximados o exactos
- Métodos basados en la búsqueda(incluidos meustreo)

Factores en la inferencia aproximada.

- Estructura de conectividad (muestroes tienen este problema)
- Fuerza de influencia(complican tanto los algoritmos de paso de mensaje como los de muestreo)
- Influencias opuestas
- Múltiples picos en funcion de verosimilitud (paso de mensaje son malos aqui)

¿Y ahora que?

. Identificar "regiones problemáticas" en la red

. Tratar de hacer la inferencia en estas regiones más exactas

- Clusters más grandes en gráfico de clústeres
- La propuesta se mueve sobre múltiples variables.
- más grande "esclavo" en dual descomposición