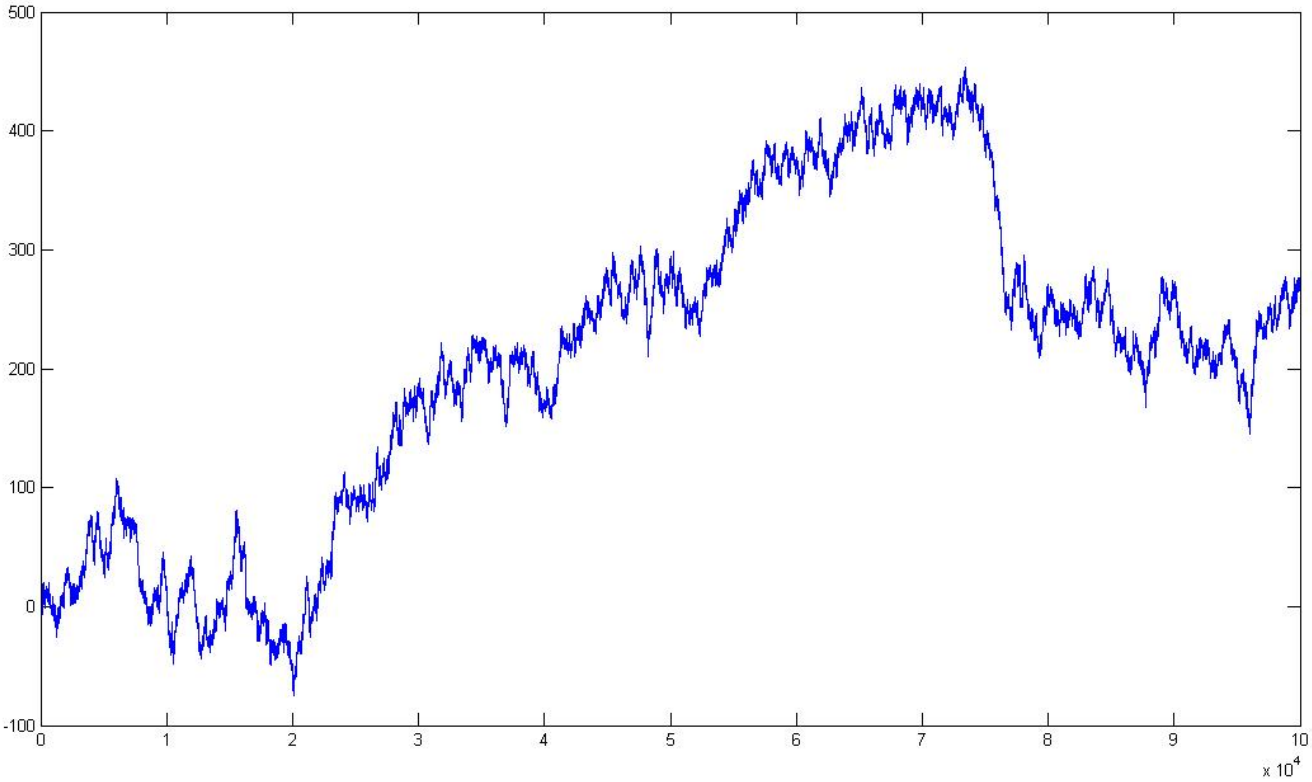
# Montecarlo methods

Course of Machine Learning
Master Degree in Computer Science
University of Rome “Tor Vergata”

Giorgio Gambosi

a.a. 2017-2018

# The problem




Integrate a“high dimensional” function . . .

# Monte Carlo Integration

# Idea

See the integral as an expectation!

## Approach

Assume we have a function $f(x)$ and a density $p(x)$ in $[a, b]$ such that $g(x) = f(x)p(x)$, we may write

$$\int_a^b g(x)dx = \int_a^b g(x)p(x)dx = E_{p(x)}[f(x)]$$

and approximate this value through the mean of $n$ values $f(x_1), \ldots, f(x_n))$ sampled from $p(x)$:

$$E[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

## Note

Let $\mathrm{Var}[f(X)] = \sigma^2$, then $\frac{1}{n}\sum_{i=1}^{n} f(x_i)$ has standard deviation $\frac{\sigma}{n}$

We may apply **Monte Carlo** to posterior distributions:

$$\begin{aligned} I(y) &= \int f(y \mid x) p(x) dx \mapsto \\ &\mapsto \hat{I}(y) = {}^{1}\!/\!_{\mathrm{n}} \sum_{i=1}^{n} f(y \mid x_i) \end{aligned}$$

Si ha la seguente stima

Monte Carlo standard error

$$SE^2[\hat{I}(y)] = {}^{1}\!/\!_{\mathrm{n}} \left( {}^{1}\!/\!_{\mathrm{n-1}} \sum_{i=1}^{n} \left( f(y \mid x_i) - \hat{I}(y) \right)^2 \right)$$

## Importance sampling

Supponiamo di poter approssimare la densità di interesse $q(x)$ con una più maneggevole $p(x)$, allora

$$\int f(x)q(x)dx = \int f(x)\frac{q(x)}{p(x)}p(x)dx = E_{p(x)}[f(x)\frac{q(x)}{p(x)}]$$

$$\text{pertanto ora } \int f(x)q(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)\frac{q(x_i)}{p(x_i)}$$

...dove gli $x_i$ sono campionati non da $q(x)$, ma da $p(x)$!

## Esempio bayesiano

Se siamo interessati a una densità marginale $J(y) = \int f(y \mid x)q(x)dx$ possiamo approssimarla

$$J(y) \approx \frac{1}{n}\sum_{i=1}^{n} f(y \mid x_i)\frac{q(x_i)}{p(x_i)}$$

# Importance sampling(hacked)

Una formulazione alternativa è la seguente, ponendo $w_i = \frac{q(x_i)}{p(x_i)}$

$$\int f(x)q(x)dx \approx \hat{I} = \sum_{i=1}^{n} \frac{w_i f(x_i)}{\sum_{i=1}^{n} w_i}$$

dove gli $x_i$ sono campionati dalla densità $p(x)$.
In questa formulazione si sta **pesando** $f(x)$ secondo quanto $p(x)$ approssima $q(x)$.

# Markov Chains

## L'equazione di Chapman-Kolmogorov

Per le catene di Markov vale

$$\begin{aligned}\pi_i(t+1) &= \Pr(X_{t+1} = s_i) \\ &= \sum_k \Pr(X_{t+1} = s_i | X_t = s_k) \cdot \Pr(X_t = s_k) \\ &= \sum_k \Pr(k \to i) \pi_k(t) \\ &= \sum_k P(k,i) \pi_k(t)\end{aligned}$$

### Il bilancio dettagliato

Condizione sufficiente per l'unicità della distribuzione stazionaria:

$$\forall i, j \quad P(j \to k)\pi_j^* = P(k \to j)\pi_k^*$$

In tal caso la catena si dice **reversibile**.

### Estensione al continuo

L'equazione di Chapman-Kolmogorov diviene

$$\pi_t(y) = \int \pi_{t-1}(x)P(x, y)dy$$

per cui la distribuzione stazionaria soddisferà

$$\pi^*(y) = \int \pi^*(x)P(x, y)dy$$

# L’algoritmo di Metropolis-Hasting

# L’algoritmo

## Il problema

Supponiamo di voler campionare da una distribuzione $p(\theta) = f(\theta)/K$ con $K$ ignota e difficile da calcolare.

Problema assai frequente in statistica bayesiana, dove spesso si ragiona secondo la relazione $\propto$.

## L'algoritmo di Metropolis

In questo processo $\Pr(\theta_t|\theta_{t-1}, \theta_{t-2}, \ldots, \theta_0) = \Pr(\theta_t|\theta_{t-1})$... **genera una catena di Markov!**

## L'output dell'algoritmo

Dopo un certo numero $k$ di passi il processo è vicino alla propria *distribuzione stazionaria* (**mixing time** o **burn-in period**), per cui campionare da $(\theta_k, \theta_{k+1}, \ldots, \theta_{k+n})$ "equivale" a campionare da $p(x)$.

## L'algoritmo di Metropolis-Hastings

- $q(\theta_1, \theta_2) = \Pr(\theta_1 \to \theta_2)$ non è necessariamente simmetrica!

- $\alpha = \min\left(\frac{f(\theta^*) q(\theta^*, \theta_{t-1})}{f(\theta_{t-1}) q(\theta_{t-1}, \theta^*)}, 1\right)$

Si tratta di una generalizzazione: se $q(\theta_1, \theta_2)$ è simmetrica si riottiene l'algoritmo di Metropolis.

Per dimostrare che l'algoritmo di Metropolis-Hasting converge alla densità candidata $p(x)$, è sufficiente mostrare che l'**equazione del bilancio dettagliato è soddisfatta**.

### Osservazione

La probabilità di transire da $x$ a $y$ è uguale alla probabilità di saltare da $x$ a $y$ per la probabilità che $y$ venga accettato essendovi saltati da $x$:

$$\Pr(x \to y) = q(x,y)\alpha(x,y) = q(x,y) \cdot \min\left\{ {}^{p(y)q(y,x)}/_{p(x)q(x,y),1} \right\}$$

ricordando che $\alpha = \min\left( {}^{f(y)q(y,x)}/_{f(x)q(x,y),1} \right)$,

l'eq. del bilancio $P(x \to y)p(x) = P(y \to x)p(y)$ diviene

$$q(x,y)\alpha(x,y)p(x) = q(y,x)\alpha(y,x)p(y)$$

# Dimostrazione

Abbiamo tre possibili casi:

1. $q(x,y)p(x) = q(y,x)p(y)$ nel cui caso $\alpha(x,y) = \alpha(y,z) = 1$ da cui $P(x \to y)p(x) = q(x,y)$ e $P(y \to x)p(y) = q(y,x)p(y)$, pertanto $P(y \to x)p(y) = P(x \to y)p(x)$
2. $q(x,y)p(x) > q(y,x)p(y)$ nel cui caso $\alpha(x,y) = {}^{p(y)q(y,x)}/_{p(x)q(x,y)}$ e $\alpha(x,y) = 1$ pertanto $P(x,y)p(x) = q(x,y)\alpha(x,y)p(x)q(x,y)\alpha(x,y)p(x) = {}^{p(y)q(y,x)}/_{p(x)q(x,y)}$
3. $q(x,y)p(x) < q(y,x)p(y)$ nel cui caso $\alpha = 1$ e $\alpha(y,x) = {}^{p(x)q(x,y)}/_{p(y)q(y,x)}$ pertanto $P(y,x)p(y) = q(y,x)\alpha(y,x)p(y)q(y,x)\alpha(y,x)p(y) = {}^{p(x)q(x,y)}/_{p(y)q(y,x)}$

QED :-)

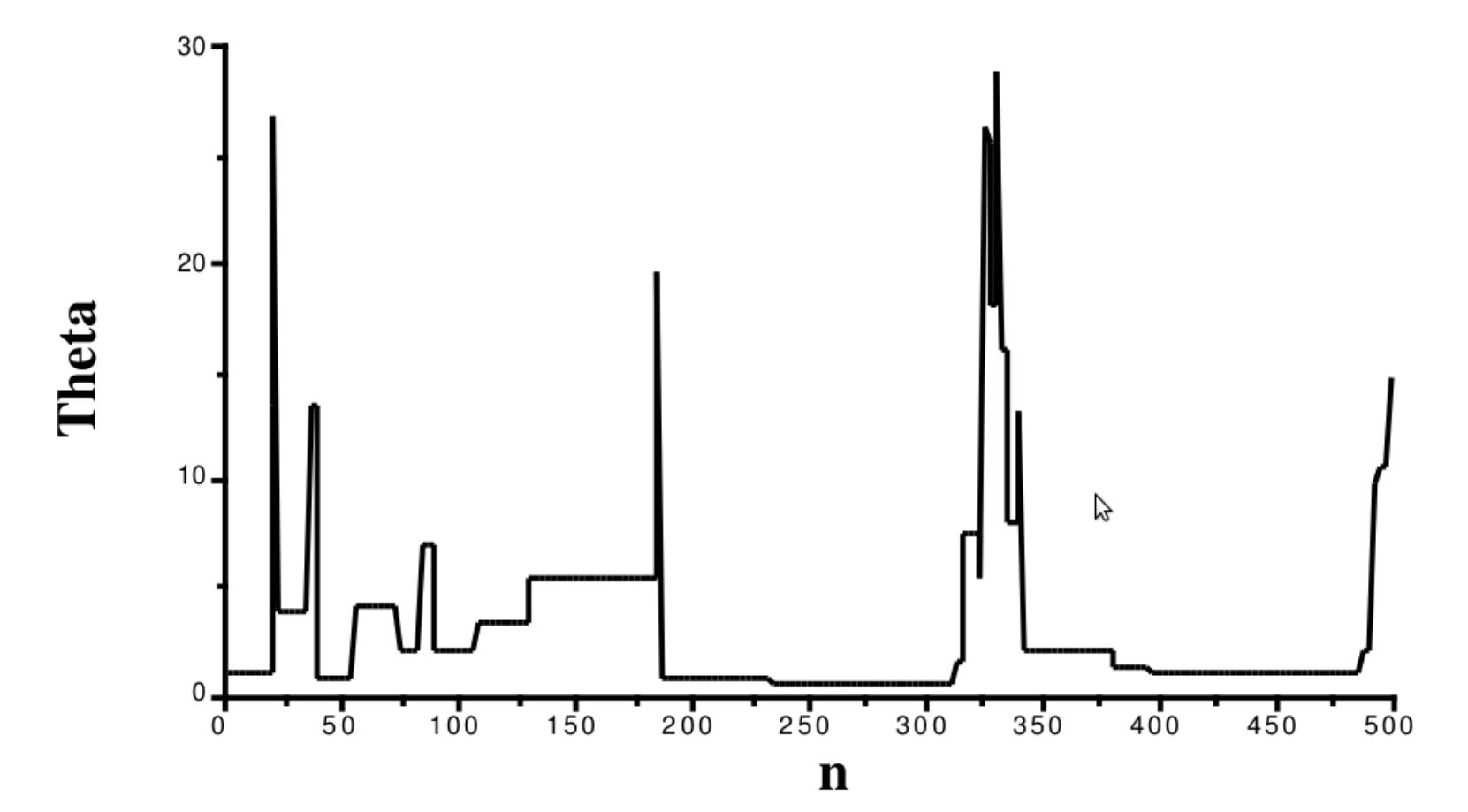
# Problemi aperti

- Quanto è il **burn-in period**? Solitamente, “dopo un po” si comincia a campionare e applicare alcuni test di convergenza.
- Qual’è una scelta ottimale per il punto iniziale $\theta_0$? Solitamente viene usato il centro della distribuzione. Un altra possibilità è iniziare è scegliere a random il punto iniziale ripetendo la simulazione per varie catene.
- Qual’è una scelta ottimale per la distribuzione di salto $q(x, y)$?

Le scelte dei parametri di cui sopra possono dare adito a due possibili risultati...

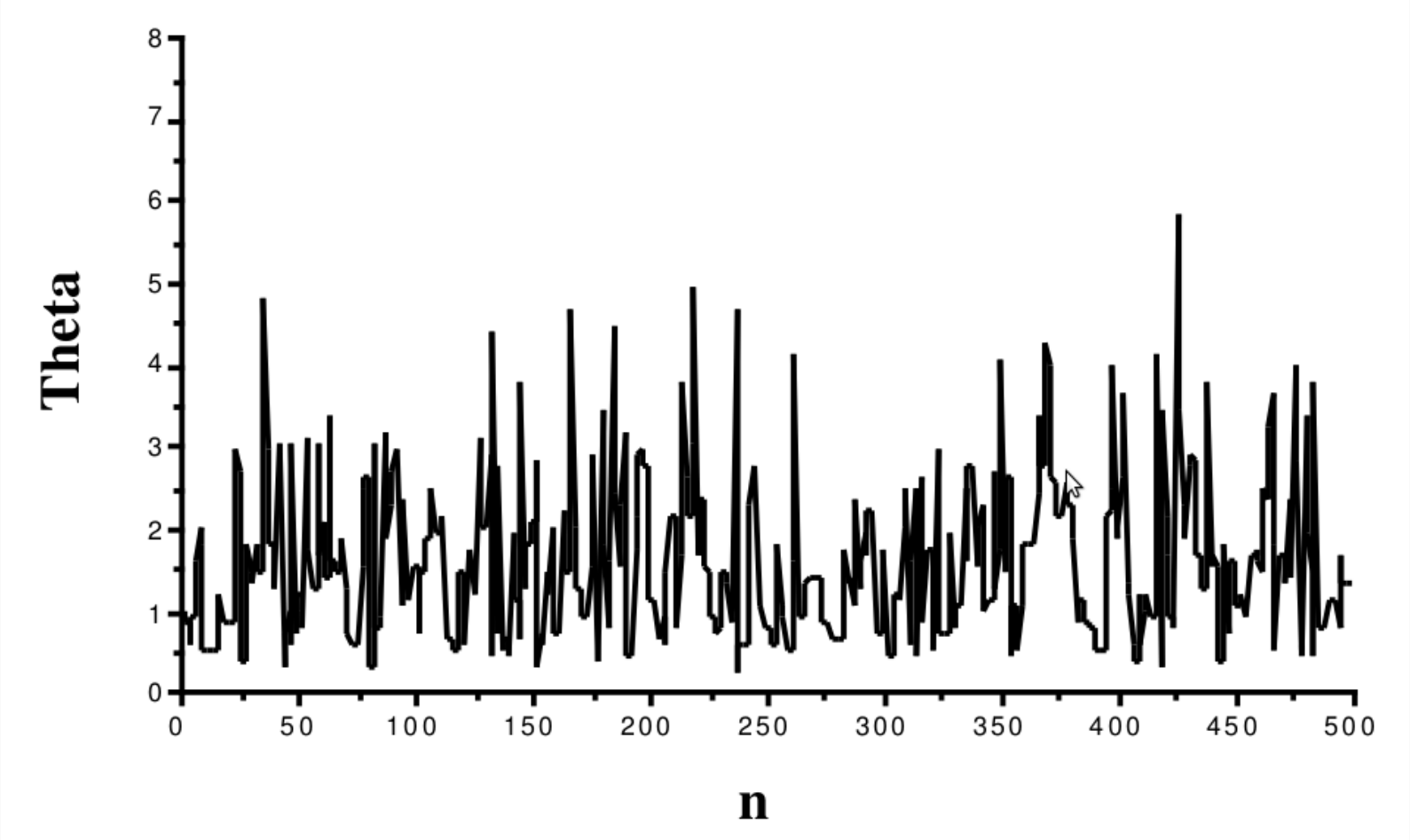
## Poorly mixing chain

Una catena si dice **poorly mixing** se indugia in piccole regioni dello spazio dei parametri per lunghi periodi di tempo.



## Well mixing chain

Una catena si dice **well mixing** se esplora allegramente lo spazio.



Una questione assai rilevante non è ancora stata affrontata...

**Problema**

In statistica Bayesiana se abbiamo una forte distribuzione a priori in conflitto con i dati osservati, può emergere una distribuzione **multimodale**. Questo rende più probabile ottenere una **poorly mixing chain**.

Una possibile soluzione è stata approfonditamente studiata.

# Simulated Annealing

L'esponente$T(t)^{-1}$

# Distribuzione di Salto

Può essere di tipo

- **random walk**: $q(x, y) = g(y - x) = g(z)$ dove $g$ è la densità associata allo spostamento $z$;
- **indipendente**: $q(x, y) = g(y)$ dove $g$ è la densità associata al cadere in $y$.

Nel primo caso la varianza di $g$ gioca il ruolo di *tuning parameter*, e può essere modificata per cercare di correggere la proprietà di *mixing* della catena.

## l'autocorrelazione

**"se il campione è sufficientemente grande"?**

Alcuni indizi sulla dimensione del campione sopra richiesta vengono dalla teoria dei **first-order autoregressive processes** ($AR_1$), dove

$$\theta_t = \mu + \alpha(\theta_{t-1} - \mu) + \epsilon \qquad \text{con } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

# Un *first order autoregressive process* e SSIF

$$\theta_t = \mu + \alpha(\theta_{t-1} - \mu) + \epsilon$$

- $E(\bar{\theta}) = \mu$
- $\rho_1 = \alpha$

- l'errore standard è $SE(\bar{\theta}) =$ $\sigma/\sqrt{n}\sqrt{1+\rho/1-\rho}$ $\sigma/\sqrt{n}$ è l'errore standard dato dal rumore bianco $\sqrt{1+\rho/1-\rho}$ è il **sample size inflaction facto**

# Test di convergenza

Strumenti *(informali)*

- **Time series trace**: grafico delle variabili aleatorie generate vs. il numero di iterazioni. Può evidenziare **poor mixing** e suggerire un periodo minimo di **burn-in**.
- $\alpha_k$ **vs.** $k$ **plot**($k$-esimo ordine di correlazione rispetto il time-lag): mostra un decadimento geometrico nella misura in cui la serie segue un $AR_1$.
- **Partial correlations plot**: la $k$-esima autocorrelazione parziale è l'eccesso di correlazione non presente in un $AR_{k-1}$, quindi se ad esempio la serie segue un $AR_1$ allora l'autocorrelazione parziale di second'ordine è zero.

# Test di convergenza

## Strumenti *(formali)*

- **Geweke test** (1992): spezza il campione in due parti, ad esempio il primo $10\%$ dei valori e l'ultimo $50\%$; se la catena è pressoché stazionaria, la media dei due sottocampioni dev'essere la stessa (per comparare i due sottocampioni possiamo usare uno **z-test modificato** ottenendo il **Geweke z-score**).
- **Raftery-Lewes test**: specificati un quantile $q$, un'accuratezza $\epsilon$ e una potenza $1 - \beta$ di raggiungere accuratezza $\epsilon$ nello specifico quantile, si costruisce la sequenza

$$\tau_t = \begin{cases} 1 & \text{se } \theta_t \leq q \\ 0 & \text{altrimenti} \end{cases}$$

ottenendo una catena di Markov a due stati, della quale si studiano le probabilità di transizione per ricavare eventuali addizionali **burn-in**, **thinning ratio** (i.e. quanti punti scartare ogni punto campionato) e le lunghezza della catena per raggiungere accuratezza $\epsilon$.

### One long or many smaller?

Una questione aperta è se sia più opportuna un'unica lunga catena o più catene meno lunghe.

Naturalmente per macchine parallele sembra preferibile avere più catene.

Usare varie corte catene può risultare inefficiente:

- se lunghi periodi di **burn-in** sono richiesti;
- se la catena ha un'autocorrelazione assai alta.

In generale qualora possibile è bene applicare ambo gli approcci svolgendo i test diagnostici sinora esposti.

# Il Gibbs Sampler

# L'idea

### Un caso speciale

Consideriamo l'algoritmo di Metropolis-Hastings con $\alpha = 1$. Dunque saltiamo sempre!

### Il problema persiste

Come costruire una catena di Markov che converga alla data distribuzione?

### Il campionamento di Gibbs

Considerare solo distribuzioni condizionali **univariate**, fissando tutti i parametri tranne uno. Anziché dover generare un vettore $n$ dimensionale usando l'intera congiunta, abbiamo quindi $n$ variabili aleatorie date da $n$ univariate condizionali.

# L'algoritmo di Gibbs

Sia $p(\Theta)$ una distribuzione multivariata $n$-dimensionale, e si indichi con $\Theta^{-k}$ il vettore contenente tutte le variabili tranne $k$.

1. Siano generati casualmente $\theta_0^{(1)}, \theta_0^{(2)} \ldots, \theta_0^{(n)}$.
2. Il valore di $\theta^{(k)}$ è campionato secondo $p(\theta^{(k)}|\Theta^{-k})$, ovvero

$$\theta_i^{(k)} \sim p(\theta^{(k)}|\theta^{(1)} = \theta_i^{(1)}, \ldots, \theta^{(k-1)} = \theta_{i-1}^{(k-1)}, \ldots, \theta^{(n)} = \theta_{i-1}^{(n)})$$

   Dopo un iniziale **burn-in** per liberarsi dal condizionamento della scelta dei valori iniziali, si campiona ogni $m$ passi ottenendo una sequenza di Gibbs.

Tale sequenza converge a una distribuzione stazionaria indipendente dalla scelta dei valori iniziali, e che per costruzione è la data distribuzione che si vuole simulare (Tierney 1994).

## Esempio

Per $(w, x, y, z)$ all'$i$-esima iterazione si ha...

$$\begin{aligned}
w_i &\sim p(w|x = x_{i-1}, \quad y = y_{i-1}, \quad z = z_{i-1}) \\
x_i &\sim p(w|w = w_i, \quad y = y_{i-1}, \quad z = z_{i-1}) \\
y_i &\sim p(w|w = w_i, \quad x = x_i, \quad z = z_{i-1}) \\
z_i &\sim p(w|w = w_i, \quad x = x_i, \quad y = y_i)
\end{aligned}$$

# La credibilità del Gibbs Sampler

- Gelfand e Smith (1990), hanno illustrato la potenza del Gibbs Sampler;
- Smith e Robers (1993), hanno mostrato la particolare naturalezza dell'uso del Gibbs Sampler in Statistica Bayesiana per ottenere distribuzioni a posteriori;
- una buona introduzione al G.S. si può trovare in Casella e George (1992) e approfondimenti in Tanner (1996), Besag et al. (1995), e Lee (1997).

Inoltre... il Gibbs Sampler può essere pensato come **un analogo stocastico dell'approccio EM** (Expectation-Maximization), dove il campionamento aleatorio rimpiazza i passi di calcolo del valore atteso e di massimizzazione.

# Applicazioni del Gibbs Sampler a distribuzioni marginali

Ogni caratteristica d'interesse delle marginali può essere calcolata dalle $m$ realizzazioni della sequenza di Gibbs.

Ad esempio per il valore atteso di una funzine $f$ della variabile aleatoria $X$ si ha l'approssimazione

$$E[f(X)]_m = {}^1\!/\!_{\mathrm{m}} \sum_{i=1}^{m} f(X_i)$$

detta **Stima Monte Carlo** di $f(X)$, giacché per $m \to \infty$ si ha

$$E[f(X)]_m \to E[f(X)]$$

Similmente, per una funzione $f$ di $n$ variabili aleatorie si ha

$$E[f(\theta^{(1)}, \ldots, \theta^{(n)})]_m = {}^1\!/\!_{\mathrm{m}} \sum_{i=1}^{m} f(\theta_i^{(1)}, \ldots, \theta_i^{(n)})$$

Mentre il calcolo della stima MC di qualsivoglia momento usando il G.S. è ovvia, calcolare la vera e propria *forma* delle densità marginali è assai più complicato.
Gelfand e Smith (1990) e Liu et al. (1991) hanno mostrato che le funzioni delle densità condizionali contengono più informazione sulla forma dell'intera distribuzione di quanta ne contenga la sequenza delle realizzazioni individuali $x_i$ campionate.

Notando che

$$p(x) = \int p(x|y)p(y)dy = E_y[p(x|y)]$$

si può infatti approssimare la densità marginale con

$$\hat{p}_m(x) = {}^{1}\!/\!_{\mathrm{m}} \sum_{i=1}^{m} p(x|y = y_i)$$

# La varianza Monte Carlo di una stima basata sul Gibbs-Sampler

Sia $\theta_1, \ldots, \theta_n$ una sequenza di Gibbs, e sia $h(\theta)$ una funzione della distribuzione che vogliamo stimare. Campionando variabili aleatorie, abbiamo una varianza del campione associata alla stima MC

$$\hat{h} = {}^{1}\!/_{\mathrm{n}} \sum_{i=1}^{n} h(\theta_i)$$

Vediamo alcuni modi di stimarla.

## La varianza campionaria con più catene

Supponiamo di generare diverse catene, allora denotando con $\hat{h_j}$ la varianza campionaria della $j$-esima catena, ponendo $\hat{h}^* = {}^{1}\!/_{\mathrm{m}} \sum_{i=1}^{m} \hat{h}_j$ si ha

$$\mathrm{Var}\left(\hat{h}\right) = {}^{1}\!/_{\mathrm{m\text{-}1}} \sum_{j=1}^{m} \left(\hat{h}_j - \hat{h}^*\right)^2$$

## La varianza campionaria con una sola catena

Usando alcuni risultati dalla teoria delle serie storiche, si ha che, stimando la **lag-$k$ autocovarianza** associata ad $h$ tramite

$$\hat{\gamma}(k) = {}^{1}\!/\!_{\mathrm{n}} \sum_{i=1}^{n-k} \Big( (h(\theta_i) - \hat{h})(h(\theta_{i+k}) - \hat{h}) \Big)$$

che è una generalizzazione dell'autocorrelazione di $k$-esimo ordine applicata alla variabile aleatoria generata da $h(\theta_i)$, la risultante stima della varianza MC è

$$\mathrm{Var}\left(\hat{h}\right) = {}^{1}\!/\!_{\mathrm{n}} \left( \hat{\gamma}(0) + 2 \sum_{i+1}^{2\delta+1} \hat{\gamma}(i) \right)$$

dove $\delta$ è il più piccolo intero positivo soddisfacente $\hat{\gamma}(2\delta) + \hat{\gamma}(2\delta + 1) > 0$.

Una misura degli effetti dell'autocorrelazione è la **lunghezza effettiva della catena**, ovvero $\hat{n} = {}^{(0)}\!/\!_{\mathrm{Var}\left(\hat{h}\right)}$ (in assenza di autocor. si ha $\hat{n} = n$).

# Diagnostica della convergenza: il Gibbs Stopper

Quanto detto per la diagnostica della convergenza riguardo il campionamento di Metropolis-Hasting si applica naturalmente al campionamento di Gibbs, essendo questo un caso speciale del precedente.

Tanner(1996) discute un approccio per monitorare la convergenza basato sul Gibbs Stopper, in cui i pesi basati sul confronto del campionamento di Gibbs e della distribuzione obiettivo sono calcolati e rappresentati graficamente come funzione del numero di iterazioni; approcciando il sampler la stazionarietà, ci si aspetta che la distribuzione dei pesi abbia un picco.