8. Autocorrelazione ed errori nella simulazione dei processi di Markov

Al termine di un esperimento numerico, come nel caso delle simulazioni sulla percolazione o sulle code, è fondamentale fornire sia i risultati che gli errori associati.

Nel caso della simulazione della percolazione, ci troviamo in una situazione semplice. Quando si dispone di un campione di N dati indipendenti, contenuti in un vettore dd con $\mathrm{len}(dd) = N$, il risultato dell'esperimento viene calcolato come $\mathrm{mean}(dd)$, mentre l'errore associato si determina come $\mathrm{std}(dd)/\sqrt{N}$, ovvero:

$$\mathrm{RISULTATO} \leftrightarrow \mathrm{mean}(dd)$$

$$\mathrm{ERRORE} \leftrightarrow \frac{\mathrm{std}(dd)}{\sqrt{N}}$$

Questo approccio è valido nei nostri esperimenti sulla percolazione, poiché i dati generati sono indipendenti, a condizione che il generatore di numeri casuali sia corretto. Si applicano quindi i risultati derivati dalla legge dei grandi numeri: se si osserva un fenomeno descritto da una certa distribuzione, la media aritmetica dei dati è una stima corretta e non distorta del valor medio della distribuzione. Inoltre, la varianza della media è uguale alla varianza del processo divisa per la cardinalità del campione. Di conseguenza, se si assume come stima dell'errore lo scarto (cioè la radice quadrata della varianza), indicato con T, lo scarto della media risulta essere:

$$\frac{T}{\sqrt{N}}$$

dove T è lo scarto del processo, misurato tramite la funzione std di MATLAB.

Simulazioni di processi di Markov

Nel caso delle simulazioni di processi di Markov, come quello della coda, l'applicazione della formula per dati indipendenti $\mathrm{std}(x)/\sqrt{\mathrm{length}(x)}$ porta a una sottostima degli errori, poiché i dati non sono indipendenti.

Notazione

Un processo viene descritto come una successione di configurazioni del sistema $X_t \mid t = 1..T_3$. Se si misura una funzione $f(x)$, si ottiene la successione $f_t = f(X_t)$. Si definisce il valor medio di f_t sulla distribuzione asintotica π come:

$$\mu_f = \langle f \rangle_\pi = \sum_x f(x)\pi_x$$

Si introduce la funzione di autocorrelazione:

$$C_{ff}(t) \equiv \langle f_t f_{t+t'} \rangle - \langle f_t \rangle \langle f_{t'} \rangle = \langle f_t f_{t+t'} \rangle - \mu_f^2$$

Questa funzione rappresenta la correlazione tra due misure distanti t nel tempo della simulazione; è quindi funzione della distanza temporale, e non del tempo assoluto.

Quando $C_{ff}(t) \approx 0$ per una certa distanza temporale t^*, si può ritenere che le misure siano decorrelate, poiché il valor medio del prodotto è approssimativamente uguale al prodotto dei valori medi.

Va sottolineato che la funzione di autocorrelazione $C_{ff}(t)$ dipende dalla funzione f considerata: ogni funzione f ha la sua specifica autocorrelazione.

La funzione può essere espressa anche come:

$$C_{ff}(t) = \langle f_t f_{t+t'} \rangle - \mu_f^2 = \sum_{x,y} f(x) \left(W_{xy}^{t'} \pi_y - \pi_x \pi_y \right) f(y)$$

e tende a zero quando $t' \to \infty$, poiché in tal caso:

$$W_{xy}^{t'} \to \pi_x$$

Decorrelazione delle misure nel tempo

Nel lungo termine, le misure raccolte simulando un processo di Markov tendono a decorrelarsi. Tuttavia, è fondamentale chiedersi: **quanto tempo ci mettono a decorrelarsi?**

Per affrontare questo aspetto, si introduce la funzione di autocorrelazione normalizzata:

$$\rho_{ff}(t) = \frac{C_{ff}(t)}{C_{ff}(0)} \quad (\text{ovvero } \rho_{ff}(0) = 1)$$

Tipicamente, questa funzione decresce esponenzialmente nel tempo secondo la legge:

$$\rho_{ff}(t) \sim e^{-t/t_c}$$

Questo comportamento non sorprende: infatti, si ricorda che la distribuzione di probabilità iniziale $p^{(0)}$ perde memoria nel tempo secondo:

$$p^{(n)} = W^n p^{(0)} \to \pi + \mathcal{O}(\hat{\lambda}^n)$$

dove $\hat{\lambda}$ è l'autovalore in modulo $|\hat{\lambda}|$ più vicino a 1. Considerando $n = t$, si ottiene:

$$\hat{\lambda}^t = e^{\ln \hat{\lambda} \cdot t} = e^{-t/t_c}$$

Poiché $|\hat{\lambda}| < 1$, allora $\ln |\hat{\lambda}| < 0$ e si può scrivere:

$$\ln |\hat{\lambda}| = -\frac{1}{t_c}$$

da cui:

$$|\hat{\lambda}|^t = e^{-t \cdot |\ln \hat{\lambda}|} = e^{-t/t_c}$$

Tempo di autocorrelazione esponenziale

Si definisce quindi il **tempo di autocorrelazione esponenziale** come:

texp=lim supt→∞−tlog|ρf(t)|t{*text{exp}}* = *limsup*{t \to \infty} \frac{-t}{\log |\rho_f(t)|}

e di conseguenza:

texp=supftexp,ft{*text{exp}}* = *sup_f t*{\text{exp}, f}

Il valore t_{exp} rappresenta il **tempo di rilassamento del modo più lento del sistema**.

Questo implica che bisogna attendere un tempo almeno dell'ordine di t_{exp} per considerare il sistema **termalizzato**, ovvero per essere certi che il sistema abbia perso memoria delle condizioni iniziali.

Tuttavia, quanto detto finora non fornisce ancora una risposta su **come valutare correttamente gli errori** nelle misure correlate.

Tempo di autocorrelazione integrato

Si definisce il **tempo di autocorrelazione integrato** come:

$$\tau_{\text{int},f} = \frac{1}{2} \sum_{t=-\infty}^{+\infty} \rho_{ff}(t) = \frac{1}{2} + \sum_{t=1}^{\infty} \rho_{ff}(t)$$

Questa definizione è coerente con la normalizzazione della funzione di autocorrelazione. In particolare, vale l'approssimazione:

$$\frac{1}{2} - \tau_{\text{int},f} \sim -\tau_{\text{exp},f} \cdot \rho_{ff}(t) \cdot e^{-t/t_c}$$

Varianza della media all'equilibrio

Sia:

$$\bar{f} = \frac{1}{N} \sum_{t=1}^{N} f_t$$

la quantità misurata, che tende a μ_f. La media $\bar{f}$ fornisce il **risultato** della simulazione, mentre la **radice quadrata della sua varianza** fornisce la **stima dell'errore**.

La varianza di $\bar{f}$ all'equilibrio è:

$$\text{Var}_\pi(\bar{f}) = \langle \bar{f}^2 \rangle - \mu_f^2 = \frac{1}{N^2} \sum_{t,t'=1}^{N} \langle f_t f_{t'} \rangle - \mu_f^2$$

che si riscrive come:

$$\mathrm{Var}_\pi(\bar{f}) = \frac{1}{N^2} \sum_{t=1}^{N} \sum_{r=-(t-1)}^{N-t} C_{ff}(r)$$

e quindi:

$$\mathrm{Var}_\pi(\bar{f}) = \frac{1}{N^2} \sum_{t=1}^{N-1} (N - t + 1) C_{ff}(t)$$

Infine, si può esprimere come:

$$\mathrm{Var}_\pi(\bar{f}) = \frac{1}{N} \left(2\langle f_t(0) \rangle \cdot \frac{1}{2} \sum_{t=-(N-1)}^{N-1} \left(1 - \frac{t+1}{N}\right) \rho_{ff}(t) \right)$$

Quando $N \gg t, \tau$ (cioè quando si hanno molte misure), si ottiene l'approssimazione:

$$\frac{1}{2} \sum_{t=-(N-1)}^{N-1} \left(1 - \frac{t+1}{N}\right) \rho_{ff}(t) \sim \frac{1}{2} \sum_{t=-\infty}^{\infty} \rho_{ff}(t)$$

Da cui segue che, per $N \gg \tau$:

$$\mathrm{Var}_\pi(\bar{f}) \approx \frac{\sigma^2}{N} (2\tau_{\mathrm{int},f}) = (2\tau_{\mathrm{int},f}) C_{ff}(0)$$

dove $C_{ff}(0)$ è la varianza di f all'equilibrio.

In altri termini:

$$\mathrm{Var}_\pi(\bar{f}) = \frac{1}{N_{\mathrm{eff}}} \mathrm{Var}_\pi(f) \quad \text{dove } N_{\mathrm{eff}} = \frac{N}{2\tau_{\mathrm{int},f}}$$

cioè la varianza della media non si riduce come $1/N$, ma come $1/N_{\mathrm{eff}}$, dove $N_{\mathrm{eff}} < N$ a causa della correlazione temporale rappresentata da $\tau_{\mathrm{int},f}$.

Pertanto, per ottenere misure decorrelate è necessario **spaziare le osservazioni nel tempo**.

Stima dell'errore

L'**errore stimato** (per esempio, per la lunghezza della coda all'equilibrio) è:

$$\mathrm{std}(\mathrm{dd}) \Big/ \sqrt{N/(2\tau_{\mathrm{int},f})}$$

dove `dd` è un campione di N misure di f lungo il processo.

Per stimare correttamente l'errore, è necessario prima calcolare $\tau_{\mathrm{int},f}$ sul campione, tramite la funzione di autocorrelazione di f.

In sintesi

Prima di calcolare medie, è indispensabile **far termalizzare** il sistema, aspettando un tempo dell'ordine di t_{exp}. Una volta raggiunto l'equilibrio, si calcola la **funzione di autocorrelazione**, da cui si ottiene il valore di $\tau_{\mathrm{int},f}$. Solo successivamente è possibile **determinare correttamente l'errore** associato alla misura.