Lezione 8

Quando tutto è lasciato al caso....

Informatica, Corso B - D. Tamascelli

Riassunto puntata precedente

- Definizione di nuovi tipi di dato
- t-uple
- Parametri formali puntatore

(Essenzialmente) completata l'introduzione dei costrutti fondamentali del linguaggio C++ nella sua declinazione procedurale.

Da oggi

- Soluzione di problemi strutturati: integrazione, simulazione esperimenti, valutazione errori.
- Introduzione tecniche algoritmiche e valutazione costo algoritmi

Oggi

- Generazione numeri casuali e tecniche "Monte Carlo"

Misure

- L'investigazione della "Natura" (fenomeni) avviene attraverso l'"esperienza".
- Esperienza: atta a misurare particolari aspetti del fenomeno in oggetto.

La MISURA è quindi lo strumento fondamentale per l'indagine di un fenomeno.

Tuttavia....

Le misure sono INEVITABILMENTE affette da errori, sia sperimentali che intrinseci:

- Precisione strumento di misura
- Intervento umano
- Fluttazioni intrinseche (vibrazioni termiche, fluttuazioni quantistiche)

Misura

UNA MISURA di una grandezza, diciamo v, consiste pertanto nella

RILEVAZIONE RIPETUTA

della grandezza stessa

rilevazioni

$r_1 \quad r_2 \quad \dots \quad r_3 \quad r_N$

stimatore

$f(r_1, r_2, \dots, r_N)$

stima

$\tilde{v}_N$

$$r_i = v + \varepsilon_i$$

Rilevazione

Valore vero

Discrepanza

ATTENZIONE: le rilevazioni r_i sono determinate dalle realizzazioni di un oggetto casuale, l'errore, che nella rilevazione $i-$esima ha assunto il valore ε_i

Misura: modello

$$R_i = v + E_i$$

Rilevazione

Valore vero

Discrepanza

$$\widetilde{V}_N = f(R_1, R_2, \ldots, R_N)$$

E_i : variabile casuale: un oggetto che assume valori in un "certo insieme" con una "certa probabilità"

R_i : variabile casuale:

$$R_i = g(E_i) = v + E_i$$

rilevazioni

R_1 R_2 $\ldots$ R_3 R_N

stimatore

$f(R_1, R_2, \ldots, R_N)$

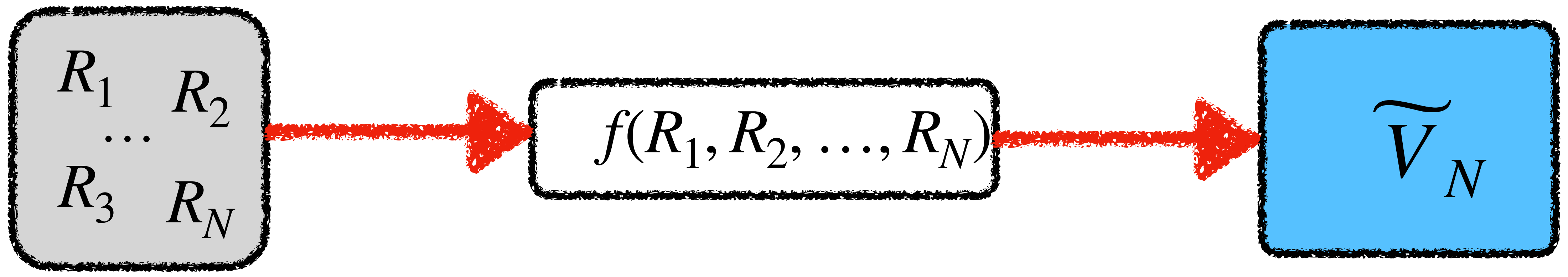
stima

$\widetilde{V}_N$

$$f(R_1, R_2, \ldots, R_N) = \frac{1}{N}\sum_{j=1}^{N} R_i = v + \frac{1}{N}\sum_{j=1}^{N} E_i$$

Media campionaria

Media campionaria



$$f(R_1, R_2, \dots, R_N) = \frac{1}{N}\sum_{j=1}^{N} R_i = v + \frac{1}{N}\sum_{j=1}^{N} E_i$$

Media campionaria

- una variabile casuale (funzione di variabili casuali),
- le cui proprietà dipendono dalle proprietà delle variabili causali che modellano le discrepanze.

Ragionevoli assunzioni sulle discrepanze:

- $E_i \perp E_j,\ i, j = 1,2,\dots,N, i \neq j$ stocasticamente indipendenti (la discrepanza i-esima non influenza la j-esima)
- le $E_i \sim E_j, i, j = 1,2,\dots,N$ identicamente distribuite (hanno le stessa caratteristiche statistiche)

Valore atteso, varianza

- Sia A una variabile casuale che assume valori in $\mathbb{R}$.
- Sia

$$P(a \leq A \leq b) = \int_{-\inf}^{+\inf} I_{[a,b]}(x)\, f_A(x)\ dx = \int_a^b f_A(x)\ dx$$

la probabilità che la variabile casuale A assuma valori nell'intervallo $[a, b]$

Si definiscono allora

- $m(A) = \int_a^b x\, f_A(x)\ dx$ il **valore atteso** di A
- $var(A) = \int_a^b (x - m(A))^2\ f_A(x)\ dx$ la **varianza** di A

- Il valore atteso è una media dei valori assumibili da A pesati per un peso $f_A(x)$.
- La varianza misura le deviazioni dal valore atteso, anch'esse pesate tramite $f_A(x)$.

Teorema Limite Centrale (Revisited)

$$R_i = v + E_i \qquad f(R_1, R_2, \ldots, R_N) = \frac{1}{N}\sum_{j=1}^{N} R_i = v + \frac{1}{N}\sum_{j=1}^{N} E_i = \widetilde{V_N}$$

$$E_i \sim E_j,\ E_i \perp E_j \ \ i,j = 1,2,\ldots,N$$

Siano inoltre

- $m(E_i) = m(E) = \mu$
- $var(E_i) = var(E) = \sigma^2$

Allora:

- $$\frac{\widetilde{V} - \mu}{\sigma/\sqrt{N}} \sim f(N; x) \xrightarrow[N \to \infty]{} f_{N(0,1)}(x)$$

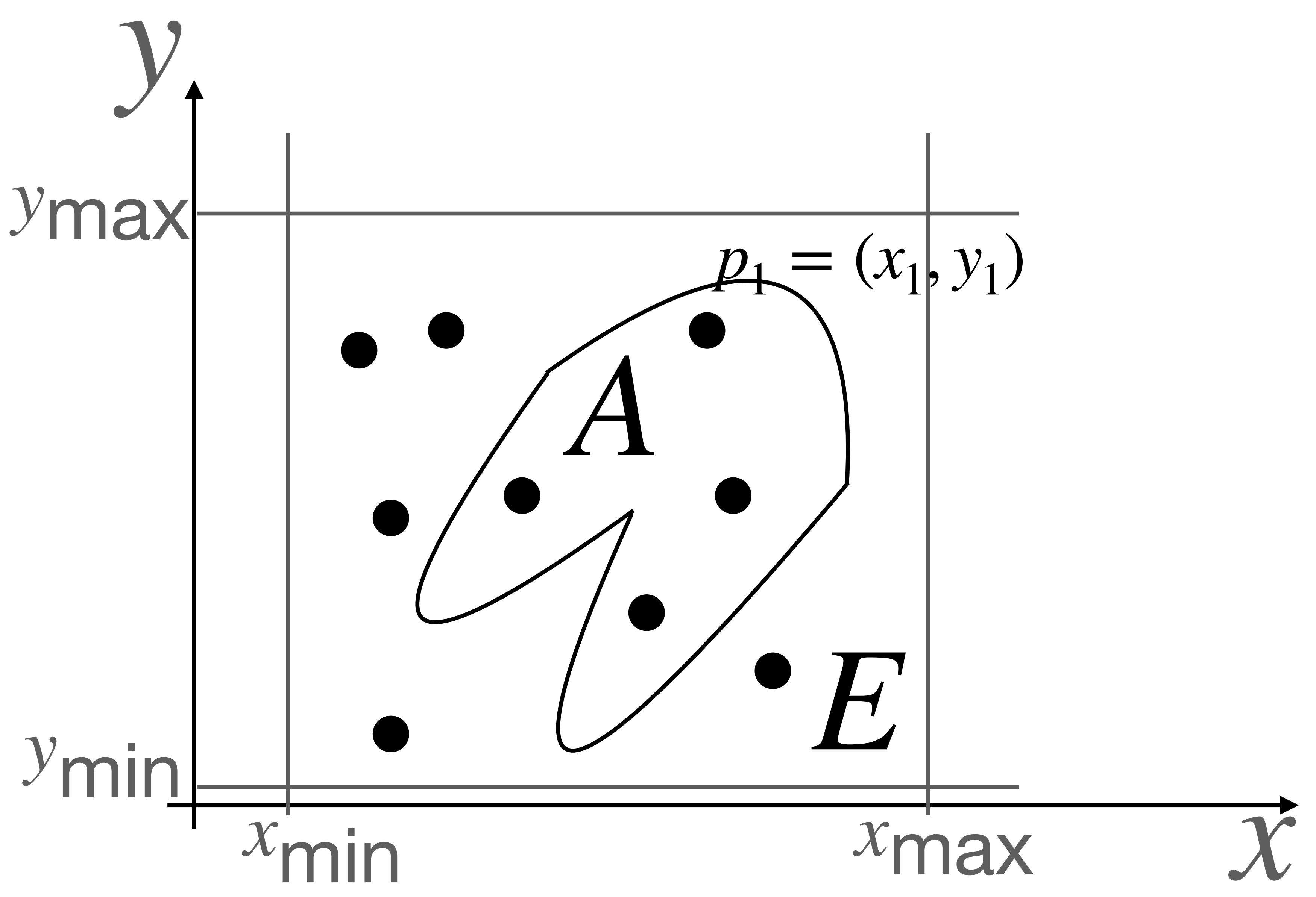
Dove

$$f_{N(v+\mu,\sigma)}(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-(v+\mu))^2}{2\sigma^2}\right]$$

ATTENZIONE: nessuna assunzione fatta sulla distribuzione delle discrepanze se non che abbiano valore atteso e varianza entrambi finiti!

Prima esperienza

- **Obiettivo: misurare area A**
- Idea: individuata un'area $E \supseteq A$...
- ..."lanciati" N punti "a caso" in E



$$\frac{A}{E} \propto \frac{D}{N} \xrightarrow[N \to \infty]{} \frac{D}{N} = \frac{A}{E}$$

D: numero di punti in A

- Determinare se punto è in A: facile.
- Possibilità di estrarre punti a caso in E
- Saper contare....

Prima esperienza: misura di π

Altrimenti detto

- A: quarto di circonferenza di raggio r

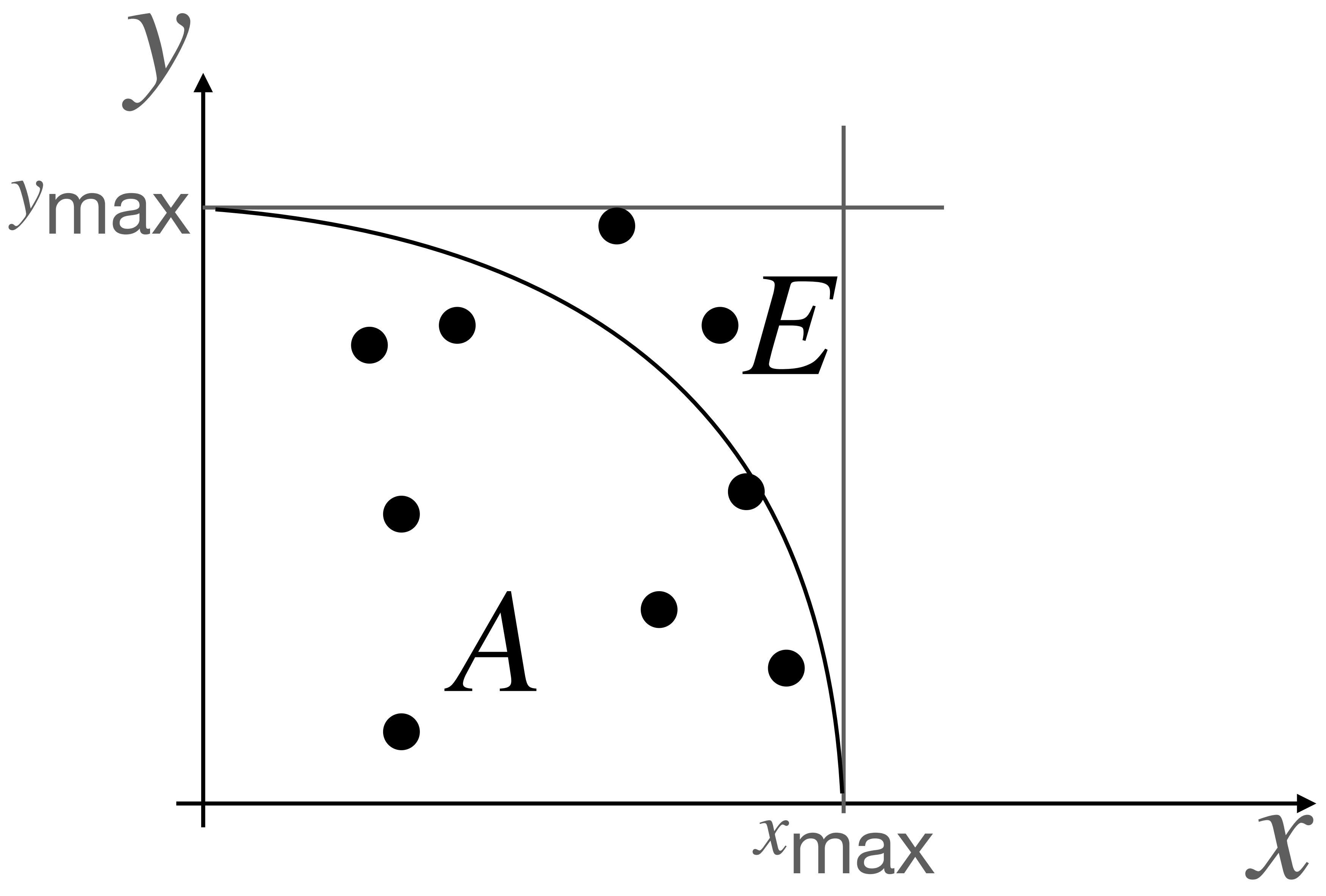
$$E = [0,r] \times [0,r]$$

$$A = \left\{(x, y) : x^2 + y^2 \leq r^2\right\}$$

$$\text{Area } A = \frac{\pi r^2}{4}$$

$$\frac{A}{E} = \frac{\pi}{4} \propto \frac{D}{N} \implies \pi \propto 4\frac{D}{N} \implies \pi = 4\frac{D}{N}$$

$$N \to \infty$$

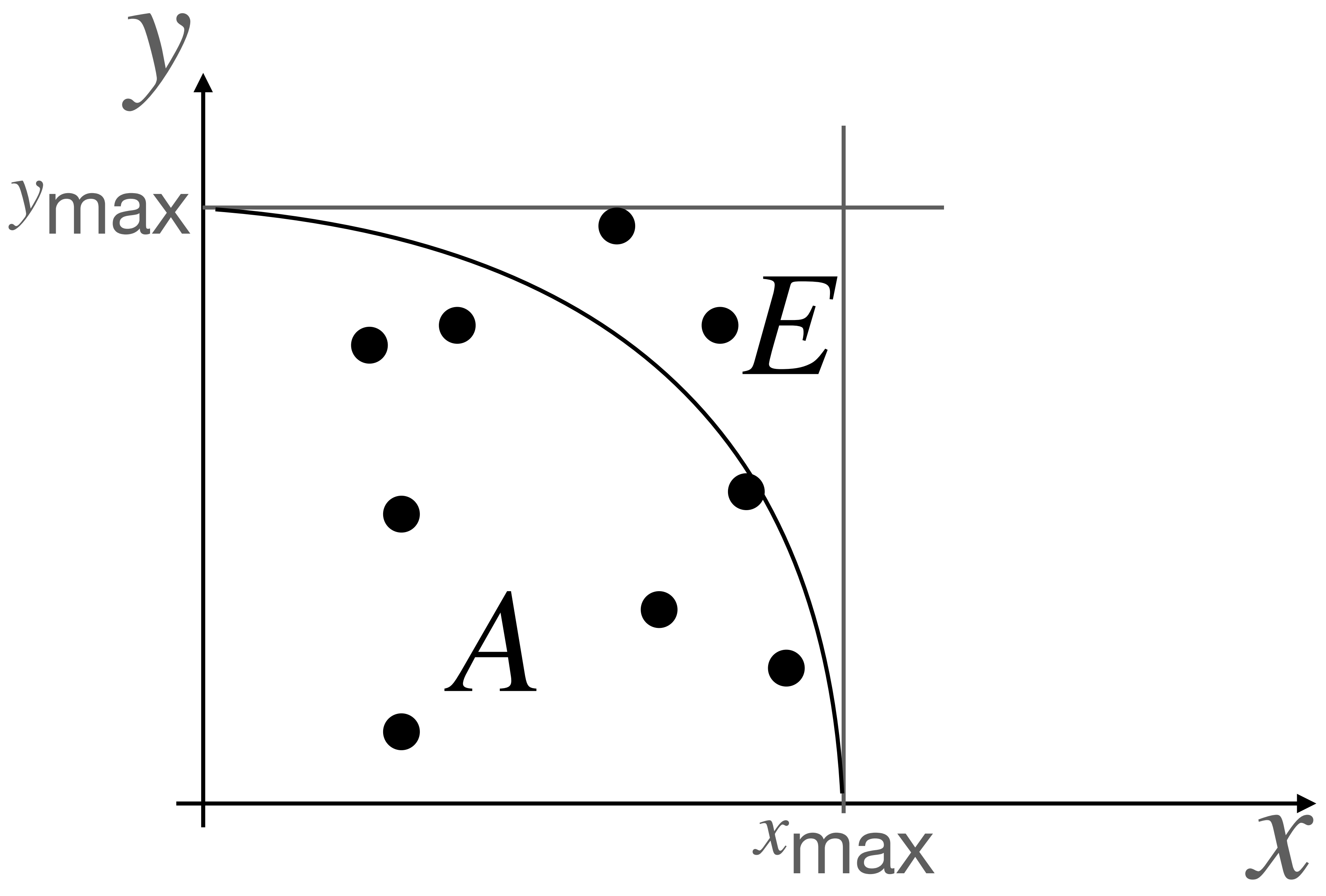


$$D = \sum_{j=1}^{N} X_i$$

$X_i = 1$ se $(x_i, y_i) \in A,\ 0$ altrimenti

$$P(X_i = 1) = \frac{A}{E}$$

Prima esperienza: misura di π



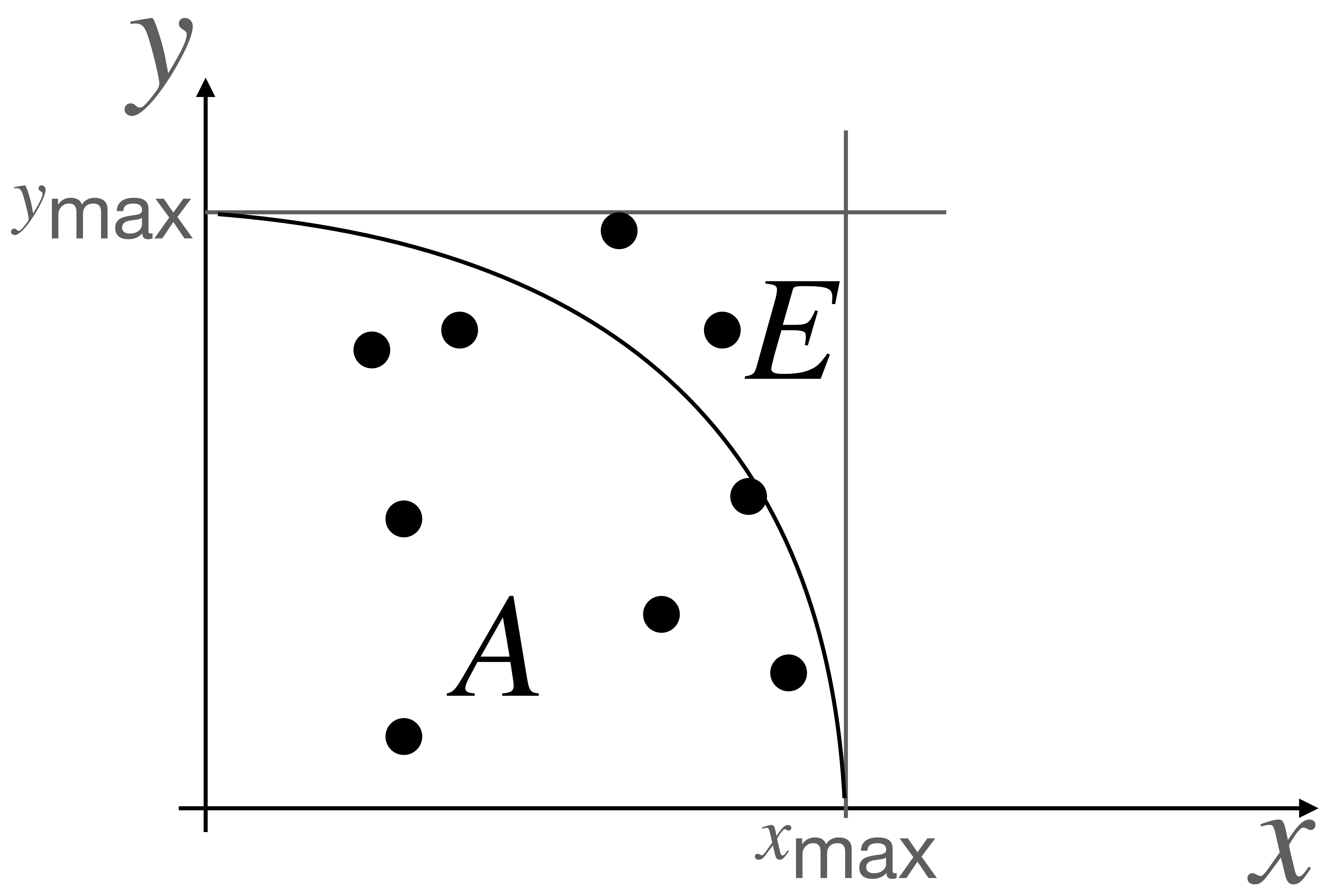
Rilevazione singola:

- Fissiamo N.
- Spariamo N punti in E (vedremo come).
- Ciascun punto determina un valore X_i.
- Determino quindi D.
- Determino quindi una realizzazione di R_i, ovvero

$$r_i = 4\frac{D}{N}$$

$$D = \sum_{j=1}^{N} X_i$$

Prima esperienza: misura di π



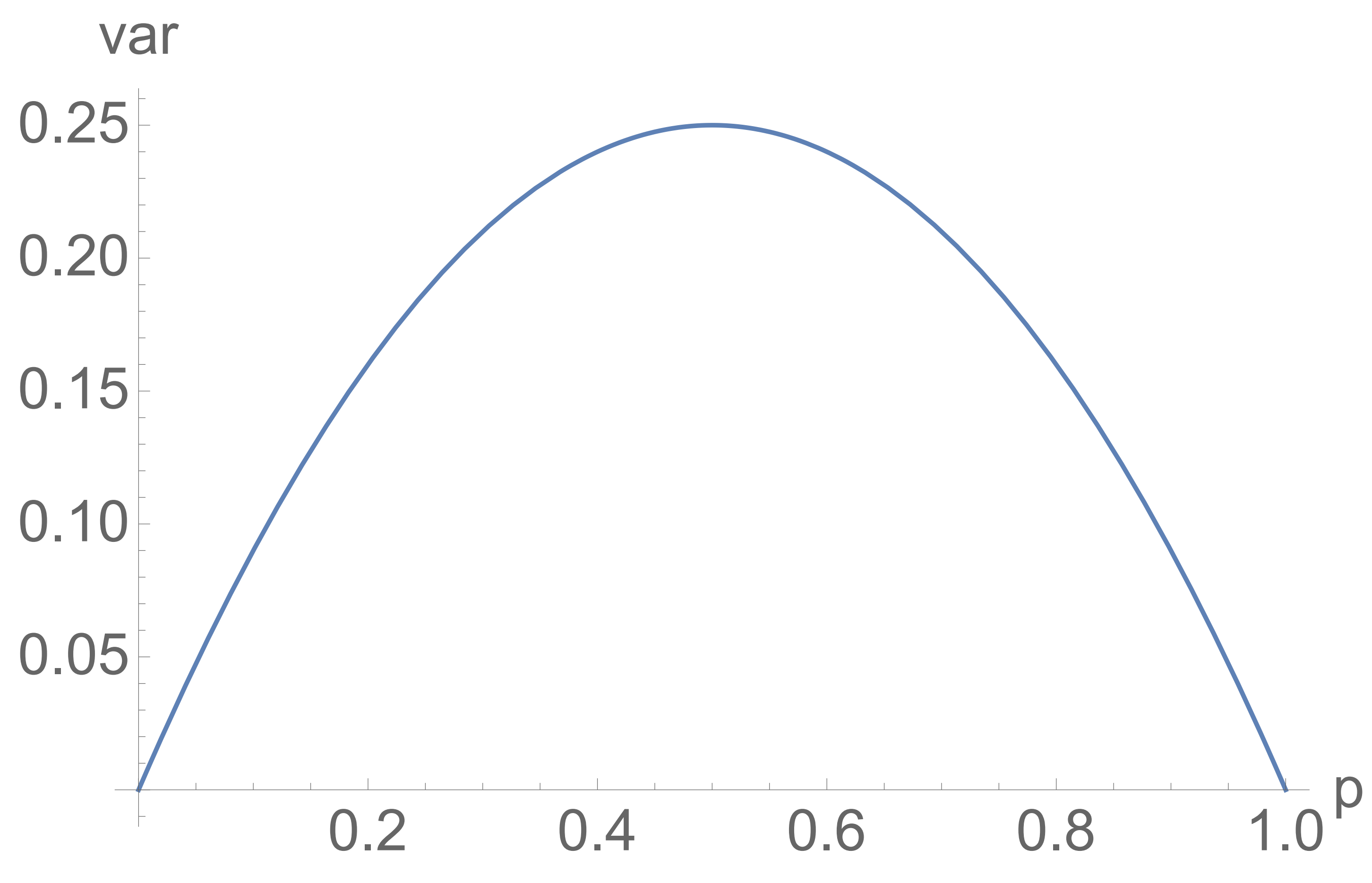
Misura:

Per ottenere una misura di π dobbiamo fare diverse rilevazioni, ovvero.

- Fissiamo il numero di rilevazioni M.
- Determinare realizzazioni r_i delle variabili casuali $(R_1, R_2, \ldots, R_M)$
- Ottenere una realizzazione dello stimatore $\widetilde{\Pi}_M = f(R_1, R_2, \ldots, R_M) = \dfrac{1}{M}\sum_{j=1}^{M} R_j$, ovvero

$$\widetilde{\pi}_M = \frac{1}{M}\sum_{j=1}^{M} r_j$$

Analisi procedura di misura (1)



- X_i: assume valore 1 con $P(X_i = 1) = \frac{A}{E} \equiv p$ e 0 con prob. $1 - p$
- $var(X_i) = p(1-p) = \frac{A}{E} - \left(\frac{A}{E}\right)^2$

$$R_i = \frac{4}{N}\sum_{j=1}^{N} X_j$$

- $E(R_i) = \frac{4}{N}\sum_j E(X_j) = 4p = 4\frac{A}{E} = \pi$
- $var(R_i) = \frac{16}{N^2}\sum_j var(X_j) = \frac{16}{N}p(1-p) = \frac{16}{N}\left[\frac{A}{E} - \left(\frac{A}{E}\right)^2\right]$

indipendenza delle X_j

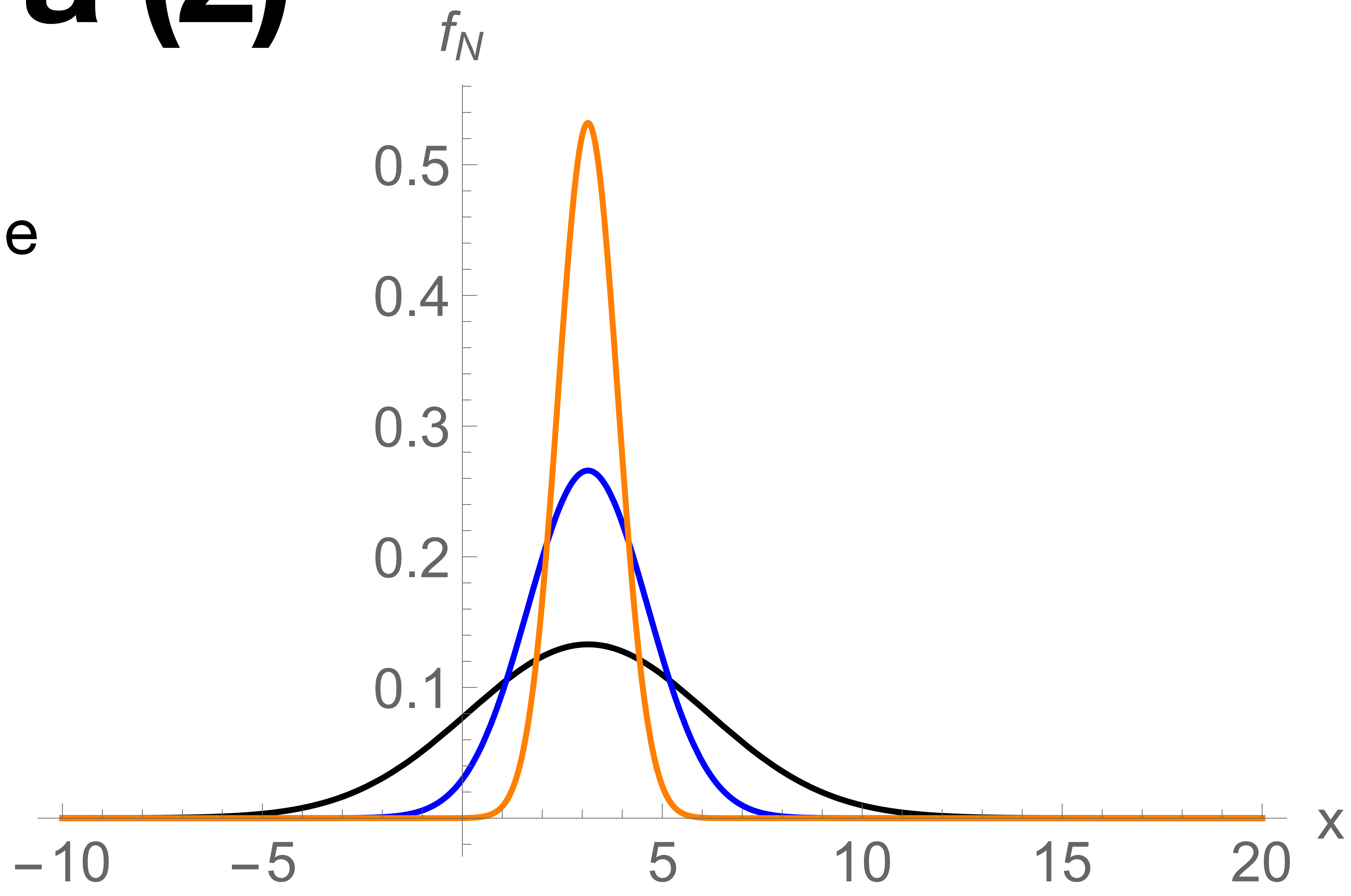
- La varianza della rilevazione diminuisce all'aumentare dei punti usati.
- È "minimizzata" da una scelta di E opportuna, con casi limite dati da $E = A$ o $E \gg A$
- ...che è la stessa scelta che minimizza la varianza di X_j

Analisi procedura di misura (2)

- $R_i = \frac{4}{N}\sum_{j=1}^{N} X_j$, con X_j i.i.d. : per N sufficientemente grande

$$f_R(x) \approx f_{N(\mu,\sigma^2)}(x)$$

- $\mu = 4\frac{A}{E} = \pi$
- $\sigma^2 = \frac{16}{N}\left[\frac{A}{E} - \left(\frac{A}{E}\right)^2\right]$



- Ogni rilevazione, per un numero di punti estratti N sufficientemente grande, può essere modellata come una v.c. Gaussiana $N(\mu, \sigma^2)$
- La varianza delle rilevazioni tende a zero per $N \to \infty$, ma per ogni valore di N finito è finita.

Analisi procedura di misura (3)

$$E(R) = 4\frac{A}{E} = \pi$$

$$var(R) = \frac{16}{N}\left[\frac{A}{E} - \left(\frac{A}{E}\right)^2\right]$$

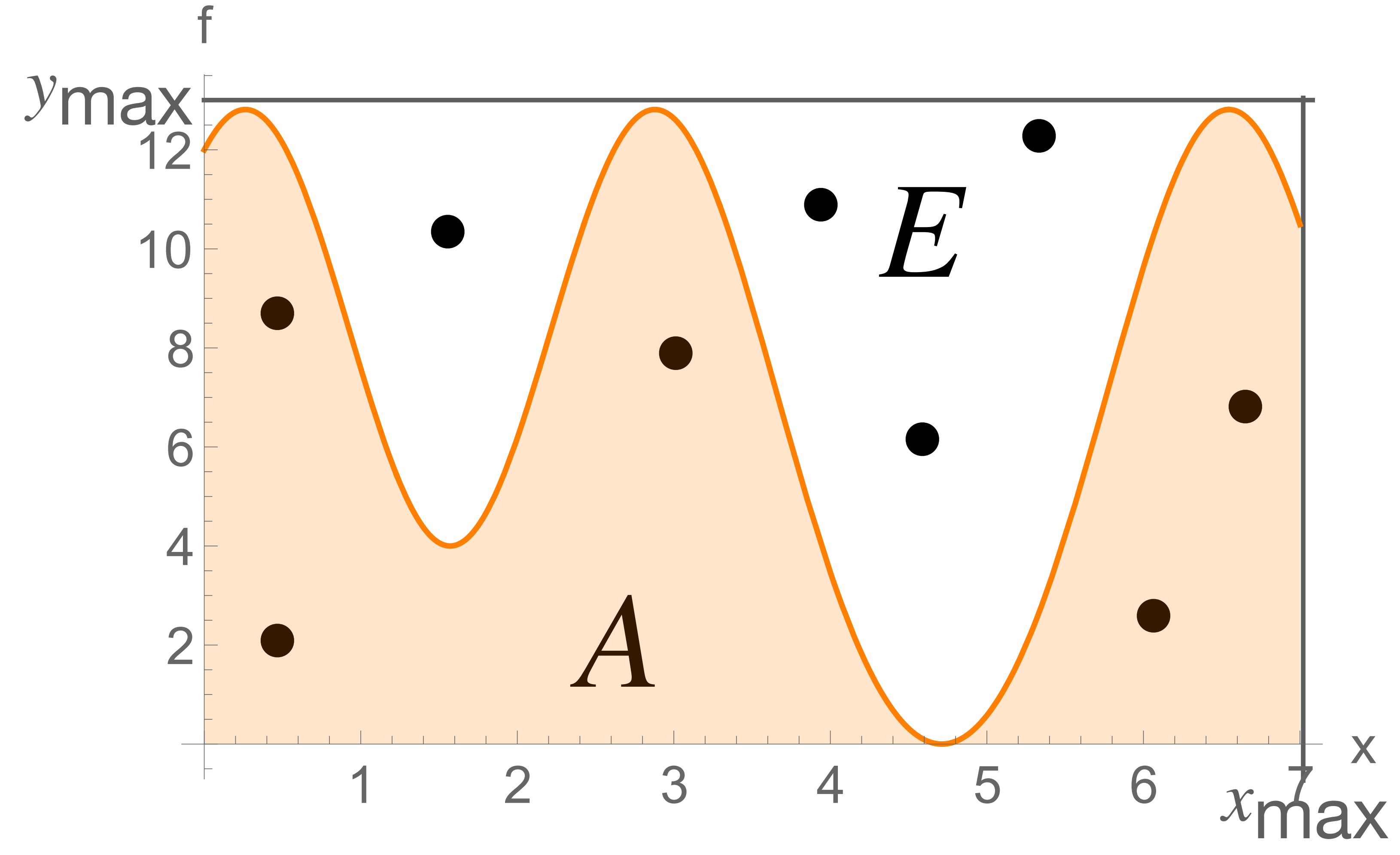
- Lo stimatore di π che proponiamo è $\widetilde{\Pi}_M = \frac{1}{M}\sum_{i=1}^{M} R_i$, con R_i i.i.d.
  - $E(\widetilde{\Pi_M}) = 4\frac{A}{E} = \pi$
  - $var(\widetilde{\Pi_M}) = \frac{1}{M}var(R) = \frac{16}{N \cdot M}\left[\frac{A}{E} - \left(\frac{A}{E}\right)^2\right]$

- Una misura di π è quindi effettuata in termini di molte rilevazioni R_i, ciascuna fatta sotto le stesse condizioni e indipendente dalle altre rilevazioni.
- La precisione di una rilevazione dipende dalla scelta dell'area E e dal numero di punti che vengono estratti.
- Fissati E ed N, la precisione della misura dipende dal numero di rilevazioni M.
- Osserviamo che N, M hanno lo stesso peso nella determinazione dell'accuratezza della misura.

Seconda esperienza: integrale di funzione

- **Obiettivo: misurare integrale**

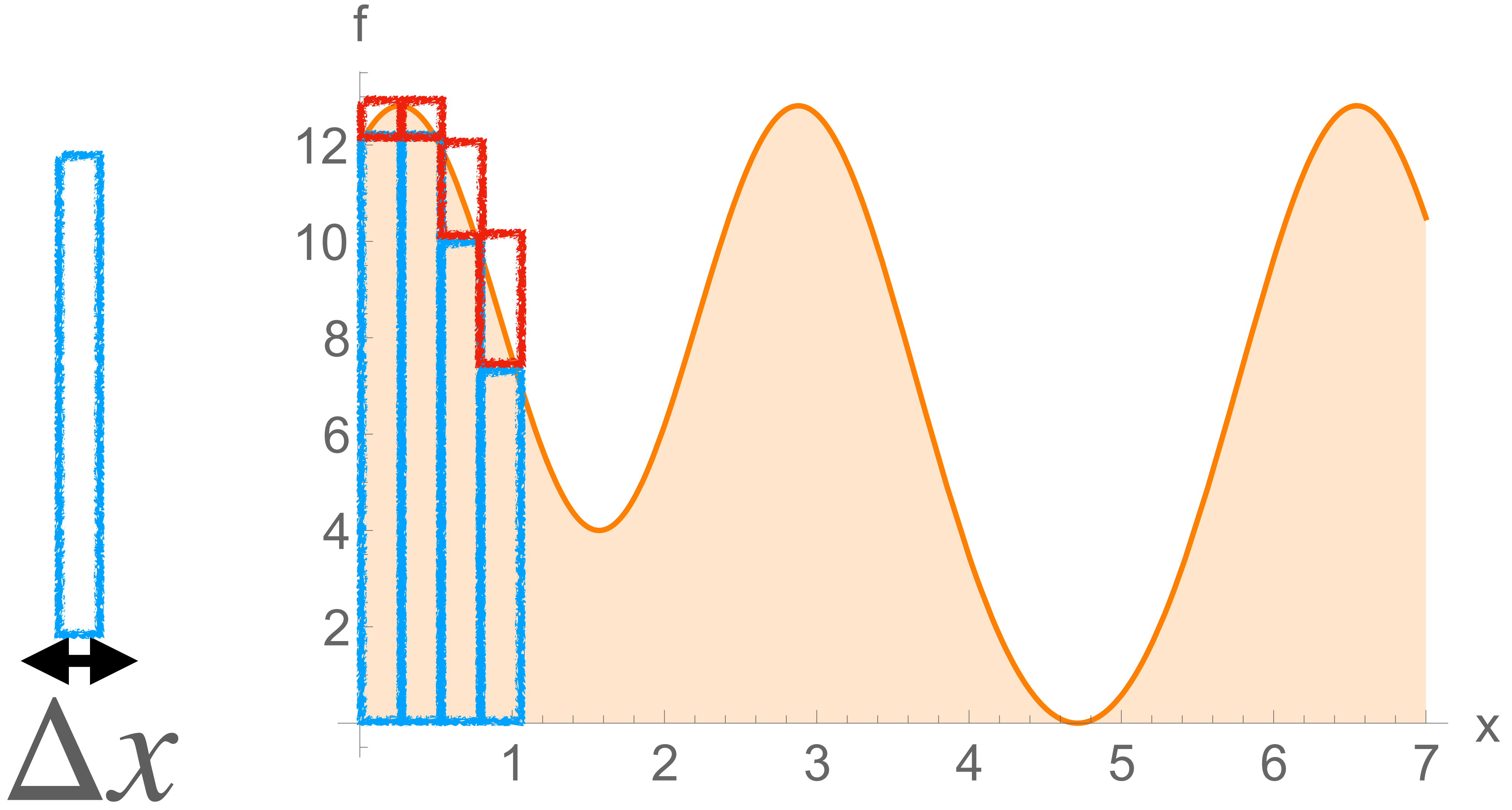
$$\int_a^b f(x)\, dx$$



- Invece di lanciare N punti a caso nel piano...

- ...prendiamo Riemann sul serio

- ... e lanciamo N punti a caso solo sulle ascisse!

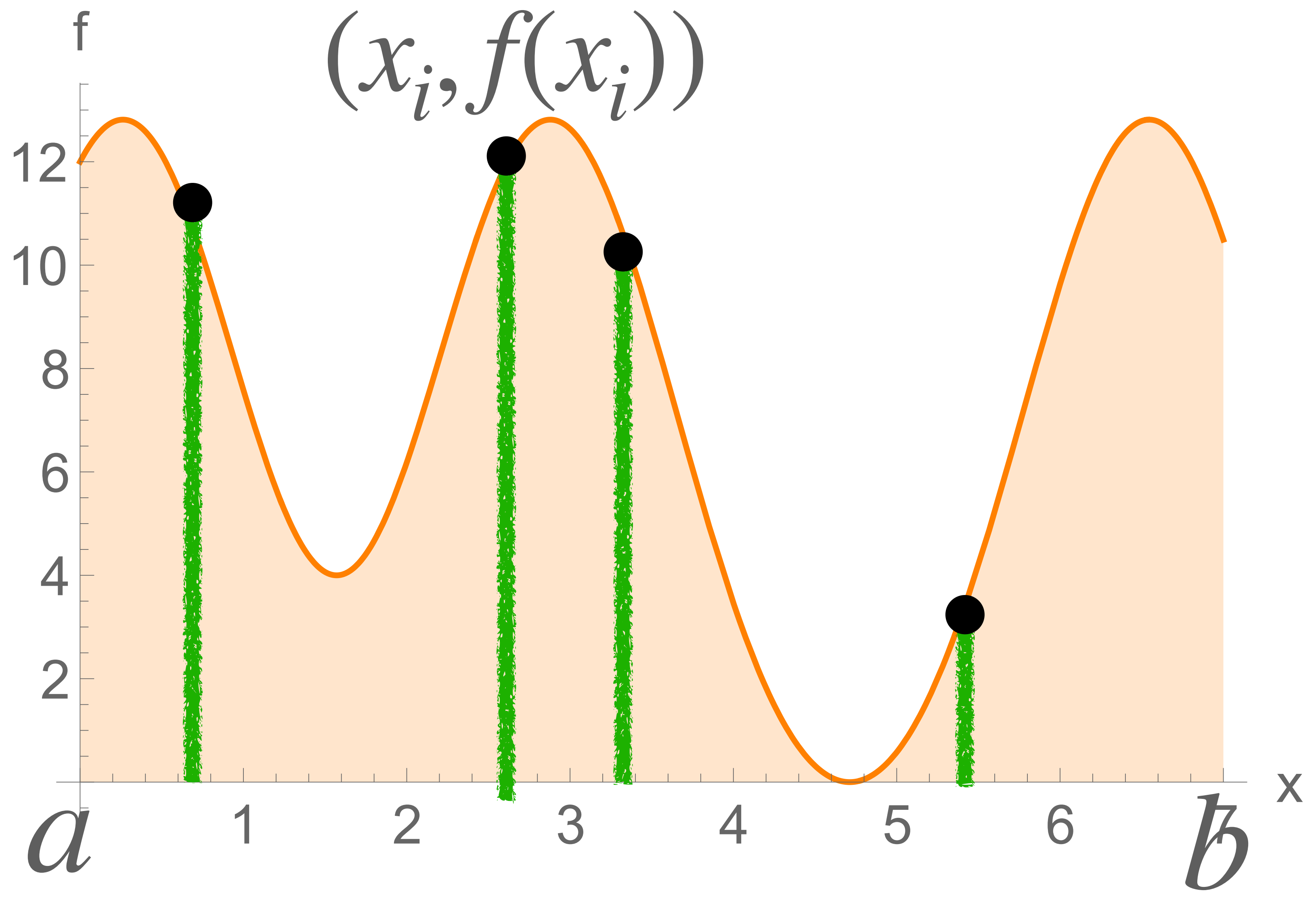


Integrazione Monte Carlo

- Lanciamo N punti a caso solo sulle ascisse.
- A ciascun punto x_i estratto attribuiamo un peso $\dfrac{b-a}{N} f(x_i)$

$$\int_a^b f(x)\, dx \approx \frac{b-a}{N} \sum_{i=1}^{N} f(x_i)$$

$$\frac{(b-a)}{N} \sum_{i=1}^{N} f(x_i) \xrightarrow[N \to \infty]{} \int_a^b f(x)\, dx$$



- Tecniche Monte Carlo introdotte da Ulam (1940) (Los Alamos).
- Svariati usi.
- Integrazione Monte Carlo strumento fondamentale per integrazione funzioni multidimensionali....
- Lo rincontrerete....

Simulazione esperimento

- Sia dato un punto materiale il cui moto è governato dalla legge

$$x(t) = x_0 + vt + at^2$$

di parametri x_0, v e a noti.

- Durante un'esperienza, la posizione del punto materiale viene registrata da un gruppo di studenti ad istanti di tempo $t_0 = 0, t_1, \ldots, t_N,$ con $t_i < t_j$ per $i < j$.

- Le posizioni registrate sono pertanto

$$y_i = x(t_i) + \varepsilon_i$$

- I valori delle posizioni rilevate nell'esperienza da ciascun gruppo vengono registrate su un file "esp_moto_gr_#.dat", dove # indica il numero identificativo del gruppo...

Simulazione fenomeni casuali

- Sia dato un punto materiale il cui moto è governato dalla legge

$$X(t + \delta t) = X(t) + v\delta t + E_i$$

di parametri x_0, v noti.

- È "facile" dimostrare che, sotto opportune ipotesi su E_i (essenzialmente indipendenza e identica distribuzione), la $X(t)$ (attenzione: è una famiglia di variabili casuali ad un parametro, nota anche come **processo stocastico**) descrive un fenomeno di diffusione con drift v
- ....chiaramente lascio un'introduzione decente all'argomento a qualche corso successivo....

Generazione numeri casuali (C++)

- Far generare numeri a caso ad un oggetto deterministico non è un affare semplice. Infatti quello che la macchina può generare sono numeri "pseudo-casuali". (Knuth, The Art of Computer Programming, vol.2 dedica qualche centinaio di pagine all'argomento).
- C++: libreria <random> fornisce molti strumenti per la generazione di numeri estratti da diverse distribuzioni....ma ha una sintassi ad oggetti ed è articolata.
- Noi usiamo uno strumento più basilare, ma che soddisfa le nostre esigenze.

#include <cstdlib>

- `rand()`: genera un numero intero positivo a caso tra o e RAND_MAX
- A caso significa: uniformemente estratto
- `RAND_MAX`: "This value is library-dependent, but is guaranteed to be at least 32767 on any standard library implementation" (documentazione).

Generazione numeri casuali (C++)-2

```
double a;

a = (double) rand()/RAND_MAX;
```

- genera un numero a caso tra 0 e 1

```
double a, b;

a = (double) rand()/RAND_MAX;

b = (max-min)*a + min  //Dilato e traslo...
```

- Genera un numero a caso tra min e max, min<max

Generazione numeri casuali (C++)-3

- rand() usa un metodo di generazione congruenziale.

$u_{n+1} = (u_n k + c) \text{ mod } m$

$u_0 = \text{seed}$

- Questo significa che:
  - una sequenza di chiamate di `rand()` genererà numeri uniformemente distribuiti.
  - Ma diverse istanze della stessa sequenza (e.g. due run dello stesso programma che genera la sequenza), porteranno alla stessa realizzazione (stessa sequenza di numeri).
  - Questo può essere comodo se si vogliono riprodurre gli stessi risultati "casuali" in diverse chiamate del programma, in caso contrario.....

Generazione numeri casuali (C++)

<u>#include <ctime></u>

- `time(NULL)`: restituisce il tempo trascorso dal 1 gennaio 2000, in secondi.
- srand(time(NULL)): imposta il seme, ovvero il primo valore generato, usando il tempo in secondi. EFFETTO: due esecuzioni del programma avverranno a tempi diversi, e quindi le sequenze di valori casuali generate saranno diverse.