# Monte Carlo inference

Alessandro La Farciola

Università di Pisa

28 Gennaio 2021


UNIVERSITÀ DI PISA

# Struttura della presentazione

- Approssimazione *Monte Carlo*
- *Sampling* a partire da distribuzioni standard
- *Rejection Sampling*
- *Importance Sampling*

# Approssimazione Monte Carlo

Sia $X$ una variabile aleatoria. Data una funzione $f$, calcolare la distribuzione di $f(X)$, in generale, può risultare molto complesso. Una valida soluzione è quella della **Approssimazione Monte Carlo**.

Essa avviene in due step:

- Generare $S$ campioni della distribuzione chiamati $x_1, \ldots, x_s$.
- Approssimare la distribuzione di $f(X)$ mediante le distribuzioni empiriche di $\{f(x_s)\}_{s=1}^{S}$.

I metodi Monte Carlo risalgono agli anni '30 e '40 del 900. In particolare, sono stati sviluppati da **Ulam**, **Fermi** e **Von Neumann** durante gli studi per la realizzazione della bomba atomica.

UNIVERSITÀ DI PISA

# Approssimazione Monte Carlo

Sia $X$ una variabile aleatoria. Data una funzione $f$, calcolare la distribuzione di $f(X)$, in generale, può risultare molto complesso. Una valida soluzione è quella della **Approssimazione Monte Carlo**.

Essa avviene in due step:

- Generare $S$ campioni della distribuzione chiamati $x_1, \ldots, x_s$.
- Approssimare la distribuzione di $f(X)$ mediante le distribuzioni empiriche di $\{f(x_s)\}_{s=1}^{S}$.

I metodi Monte Carlo risalgono agli anni '30 e '40 del 900. In particolare, sono stati sviluppati da **Ulam**, **Fermi** e **Von Neumann** durante gli studi per la realizzazione della bomba atomica.

UNIVERSITÀ DI PISA

# Approssimazione Monte Carlo

Sia $X$ una variabile aleatoria. Data una funzione $f$, calcolare la distribuzione di $f(X)$, in generale, può risultare molto complesso. Una valida soluzione è quella della **Approssimazione Monte Carlo**.

Essa avviene in due step:

- Generare $S$ campioni della distribuzione chiamati $x_1, \ldots, x_s$.
- Approssimare la distribuzione di $f(X)$ mediante le distribuzioni empiriche di $\{f(x_s)\}_{s=1}^{S}$.

I metodi Monte Carlo risalgono agli anni '30 e '40 del 900. In particolare, sono stati sviluppati da **Ulam**, **Fermi** e **Von Neumann** durante gli studi per la realizzazione della bomba atomica.


UNIVERSITÀ DI PISA

# Approssimazione Monte Carlo

É possibile utilizzare il metodo *Monte Carlo* per approssimare il valore atteso di una certa v.a. $f(X)$. Se, infatti, campiono $x_s \sim p(x)$, dove $p(x)$ è la densità di probabilità associata a $X$, allora vale:

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s).$$

# Approssimazione Monte Carlo

É possibile utilizzare il metodo *Monte Carlo* per approssimare il valore atteso di una certa v.a. $f(X)$. Se, infatti, campiono $x_s \sim p(x)$, dove $p(x)$ è la densità di probabilità associata a $X$, allora vale:

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s).$$

Tale metodo è detto **Integrazione Monte Carlo** e al variare di $f(\cdot)$ è possibile approssimare molte quantità utili:

- $\bar{x} = \frac{1}{S}\sum_{s=1}^{S} x_s \to \mathbb{E}[X]$
- $\frac{1}{S}\sum_{s=1}^{S}(x_s - \bar{x})^2 \to var(X)$
- $\frac{1}{S}\#\{x_s \leq c\} \to P(X \leq c)$
- $median\{x_1, \ldots, x_S\} \to median(X)$

UNIVERSITÀ DI PISA

# Sampling a partire dalle distribuzioni standard

Il metodo di *sampling* più semplice è quello che sfrutta la funzione di ripartizione (**cdf**) e in particolare la sua **inversa**.

# Sampling a partire dalle distribuzioni standard

Il metodo di *sampling* più semplice è quello che sfrutta la funzione di ripartizione (**cdf**) e in particolare la sua **inversa**.

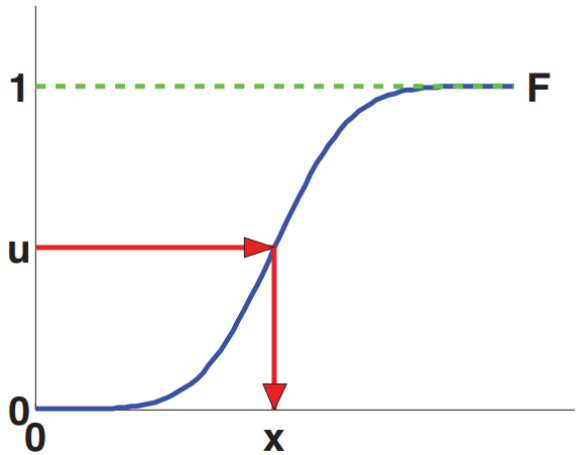
Assumiamo di voler ottenere un campionamento $x_s$ a partire da una certa distribuzione e sia $F(x) : \mathbb{R} \to (0, 1)$ la sua *cdf*. Supponiamo che $F$ sia invertibile e chiamiamo $F^{-1}(x) : (0, 1) \to \mathbb{R}$ l'inversa.

# Sampling a partire dalle distribuzioni standard

Il metodo di *sampling* più semplice è quello che sfrutta la funzione di ripartizione (**cdf**) e in particolare la sua **inversa**.

Assumiamo di voler ottenere un campionamento $x_s$ a partire da una certa distribuzione e sia $F(x) : \mathbb{R} \to (0, 1)$ la sua *cdf*. Supponiamo che $F$ sia invertibile e chiamiamo $F^{-1}(x) : (0, 1) \to \mathbb{R}$ l'inversa.

L'idea è quella di generare secondo la distribuzione uniforme $U(0, 1)$ un campione $u_1, \ldots, u_s$ e poi prendere $F^{-1}(u_1), \ldots, F^{-1}(u_s)$.



# Sampling a partire dalle distribuzioni standard

La correttezza di tale metodo è garantita dal seguente

**Teorema**

*Se $U \sim U(0,1)$ è una v.a. uniforme, allora $F^{-1}(U) \sim F$*

**Proof.**

Poiché $F$ è una cdf, allora è non decrescente. In particolare, essendo invertibile, è strettamente crescente e da questo segue che anche l'inversa è non decrescente.

Inoltre, ricordiamo che la cdf di $U \sim U(0,1)$ è $x\mathbb{1}_{(0,1)}$.

Pertanto,

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

□

# Sampling a partire dalle distribuzioni standard

La correttezza di tale metodo è garantita dal seguente

**Teorema**

*Se $U \sim U(0,1)$ è una v.a. uniforme, allora $F^{-1}(U) \sim F$*

**Proof.**

Poiché $F$ è una cdf, allora è non decrescente. In particolare, essendo invertibile, è strettamente crescente e da questo segue che anche l'inversa è non decrescente.

Inoltre, ricordiamo che la cdf di $U \sim U(0,1)$ è $x\mathbb{1}_{(0,1)}$.

Pertanto,

$$P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$$

□

# Sampling a partire dalle distribuzioni standard

Se $F$ non è invertibile possiamo definire una sorta di inversa nel seguente modo.

## Definizione

Sia $F : \mathbb{R} \to [0, 1]$. Si definisce **Smirnov transform** la funzione $G : (0, 1) \to \mathbb{R}$ tale che

$$G(a) := \inf\{x \in \mathbb{R} \ : \ a \leq F(x)\}$$

# Sampling a partire dalle distribuzioni standard

Se $F$ non è invertibile possiamo definire una sorta di inversa nel seguente modo.

## Definizione

Sia $F : \mathbb{R} \to [0, 1]$. Si definisce **Smirnov transform** la funzione $G : (0, 1) \to \mathbb{R}$ tale che

$$G(a) := \inf\{x \in \mathbb{R} \ : \ a \leq F(x)\}$$

1. Se $F$ è invertibile allora $G = F^{-1}$.
2. Affinché $G$ sia ben definita, bisogna che l'insieme sia $\neq \emptyset$ e che l'inf sia $\neq \pm\infty$; ma se $F$ è una cdf queste sono verificate.


UNIVERSITÀ DI PISA

# Sampling a partire dalle distribuzioni standard

## Teorema

*Siano $F$ una cdf e $G$ la sua Smirnov transform. Se $U \sim U(0,1)$, allora $G(U) \sim F$.*

## Proof.

Claim: $\forall\ a \in (0,1)\ \forall\ y \in \mathbb{R}$ si ha che $a \leq F(y)$ *sse* $G(a) \leq y$.

$\Rightarrow$) $a \leq F(y) \Rightarrow G(a) \leq y$ in quanto $y$ appartiene all'insieme di cui faccio l'inf.

$\Leftarrow$) Esiste $x_n \geq G(a)$ tale che $x_n \to G(a)$. Ma allora $a \leq F(x_n)$ per ogni $n$. Quindi per la continuità a dx di $F$

$$a \leq \lim_n F(x_n) = F(G(a)) \leq F(y).$$

Ora,

$$P(G(U) \leq y) = P(U \leq F(y)) = F(y).$$

□

# Sampling a partire dalle distribuzioni standard

## Teorema

*Siano $F$ una cdf e $G$ la sua Smirnov transform. Se $U \sim U(0,1)$, allora $G(U) \sim F$.*

## Proof.

Claim: $\forall\ a \in (0,1)\ \forall\ y \in \mathbb{R}$ si ha che $a \leq F(y)$ *sse* $G(a) \leq y$.

$\Rightarrow$) $a \leq F(y) \Rightarrow G(a) \leq y$ in quanto $y$ appartiene all'insieme di cui faccio l'inf.

$\Leftarrow$) Esiste $x_n \geq G(a)$ tale che $x_n \to G(a)$. Ma allora $a \leq F(x_n)$ per ogni $n$. Quindi per la continuità a dx di $F$

$$a \leq \lim_n F(x_n) = F(G(a)) \leq F(y).$$

Ora,

$$P(G(U) \leq y) = P(U \leq F(y)) = F(y).$$

# Sampling a partire dalle distribuzioni standard

Ad esempio, consideriamo la distribuzione esponenziale $Expon_\lambda$. La sua cdf è pari a $F(x) = 1 - e^{-\lambda x}\mathbb{1}_{\{x \geq 0\}}$, la cui inversa (più precisamente *Smirnov transform*) è $G(x) = -ln(1-x)/\lambda$.

Dal precedente teorema se $U \sim Unif(0,1)$, allora $G(U) \sim Expon_\lambda$.

Inoltre, se $U \sim Unif(0,1)$, allora anche $1 - U$ lo è e quindi posso trasformare il campione uniforme nel campione esponenziale applicando $-ln(x)/\lambda$.

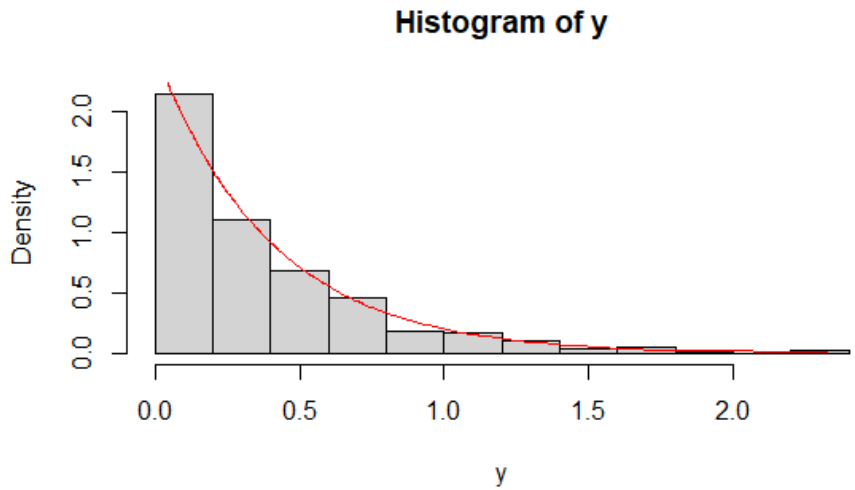
# Sampling a partire dalle distribuzioni standard

Ad esempio, consideriamo la distribuzione esponenziale $Expon_\lambda$. La sua cdf è pari a $F(x) = 1 - e^{-\lambda x}\mathbb{1}_{\{x \geq 0\}}$, la cui inversa (più precisamente *Smirnov transform*) è $G(x) = -ln(1-x)/\lambda$.

Dal precedente teorema se $U \sim Unif(0,1)$, allora $G(U) \sim Expon_\lambda$.

Inoltre, se $U \sim Unif(0,1)$, allora anche $1 - U$ lo è e quindi posso trasformare il campione uniforme nel campione esponenziale applicando $-ln(x)/\lambda$.

```
x=runif(500,0,1)
y=-log(x)/2.5
hist(y,10,freq=F)
lines(sort(y),dexp(sort(y),2.5), col="red")
```



# Metodo di Box-Muller

Il **Metodo di Box-Muller** è un modo di campionare a partire da una distribuzione gaussiana. L'idea è la seguente:

1. Partire un campionamento uniforme su un cerchio di raggio 1
2. Applicare un cambio di variabili per ottenere un campionamento su una gaussiana 2-dimensionale
3. Pensare alla gaussiana 2d come prodotto di due gaussiane 1d e restringerci ad una dimensione

Nei dettagli, si campionano $z_1$ e $z_2 \in (-1, 1)$ in maniera uniforme e, scartando quelle coppie che non soddisfano la condizione $z_1^2 + z_2^2 \leq 1$, si ottiene un campionamento uniforme sul cerchio unitario con distribuzione $p(\mathbf{z}) = \frac{1}{\pi} \mathbb{1}_{B(0,1)}$.

UNIVERSITÀ DI PISA

# Metodo di Box-Muller

Il **Metodo di Box-Muller** è un modo di campionare a partire da una distribuzione gaussiana. L'idea è la seguente:

1. Partire un campionamento uniforme su un cerchio di raggio 1
2. Applicare un cambio di variabili per ottenere un campionamento su una gaussiana 2-dimensionale
3. Pensare alla gaussiana 2d come prodotto di due gaussiane 1d e restringerci ad una dimensione

Nei dettagli, si campionano $z_1$ e $z_2 \in (-1, 1)$ in maniera uniforme e, scartando quelle coppie che non soddisfano la condizione $z_1^2 + z_2^2 \leq 1$, si ottiene un campionamento uniforme sul cerchio unitario con distribuzione $p(\mathbf{z}) = \frac{1}{\pi} \mathbb{1}_{B(0,1)}$.


UNIVERSITÀ DI PISA

# Metodo di Box-Muller

Il **Metodo di Box-Muller** è un modo di campionare a partire da una distribuzione gaussiana. L'idea è la seguente:

1. Partire un campionamento uniforme su un cerchio di raggio 1
2. Applicare un cambio di variabili per ottenere un campionamento su una gaussiana 2-dimensionale
3. Pensare alla gaussiana 2d come prodotto di due gaussiane 1d e restringerci ad una dimensione

Nei dettagli, si campionano $z_1$ e $z_2 \in (-1, 1)$ in maniera uniforme e, scartando quelle coppie che non soddisfano la condizione $z_1^2 + z_2^2 \leq 1$, si ottiene un campionamento uniforme sul cerchio unitario con distribuzione $p(\mathbf{z}) = \frac{1}{\pi} \mathbb{1}_{B(0,1)}$.

UNIVERSITÀ DI PISA

# Metodo di Box-Muller

Il **Metodo di Box-Muller** è un modo di campionare a partire da una distribuzione gaussiana. L'idea è la seguente:

1. Partire un campionamento uniforme su un cerchio di raggio 1
2. Applicare un cambio di variabili per ottenere un campionamento su una gaussiana 2-dimensionale
3. Pensare alla gaussiana 2d come prodotto di due gaussiane 1d e restringerci ad una dimensione

Nei dettagli, si campionano $z_1$ e $z_2 \in (-1, 1)$ in maniera uniforme e, scartando quelle coppie che non soddisfano la condizione $z_1^2 + z_2^2 \leq 1$, si ottiene un campionamento uniforme sul cerchio unitario con distribuzione $p(\mathbf{z}) = \frac{1}{\pi} \mathbb{1}_{B(0,1)}$.


UNIVERSITÀ DI PISA

# Metodo di Box-Muller

Ora, si definisce

$$x_i = z_i \left( \frac{-2lnr^2}{r^2} \right)^{\frac{1}{2}}$$

per $i = 1, 2$, dove $r^2 = z_1^2 + z_2^2$.

Grazie alla formula del cambio di variabili, si ottiene

$$p(x_1, x_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(x_1, x_2)} \right| = \left[ \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x_1^2 \right) \right] \left[ \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x_2^2 \right) \right]$$

Quindi $x_1$ e $x_2$ sono due campioni indipendenti di una gaussiana standard uno-dimensionale.


UNIVERSITÀ DI PISA

# Metodo di Box-Muller

Ora, si definisce

$$x_i = z_i \left( \frac{-2lnr^2}{r^2} \right)^{\frac{1}{2}}$$

per $i = 1, 2$, dove $r^2 = z_1^2 + z_2^2$.

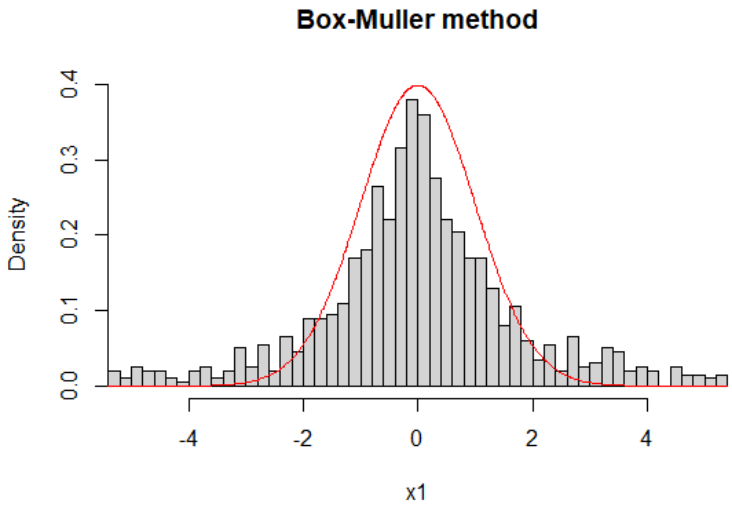
Grazie alla formula del cambio di variabili, si ottiene

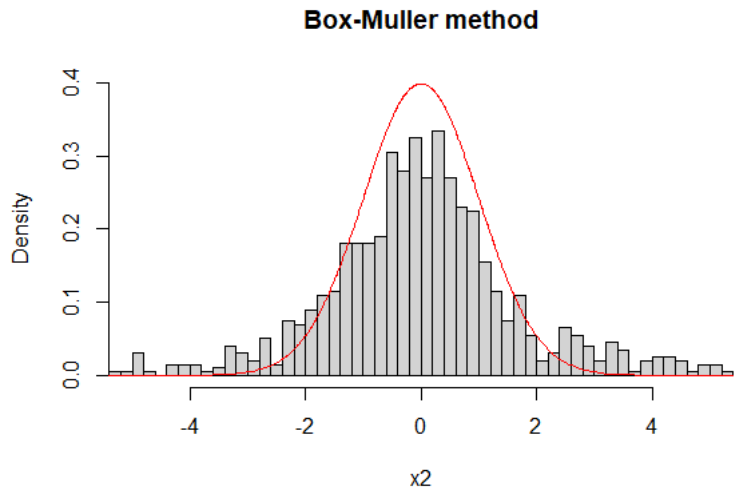
$$p(x_1, x_2) = p(z_1, z_2) \left| \frac{\partial(z_1, z_2)}{\partial(x_1, x_2)} \right| = \left[ \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x_1^2 \right) \right] \left[ \frac{1}{\sqrt{2\pi}} \exp\left( -\frac{1}{2} x_2^2 \right) \right]$$

Quindi $x_1$ e $x_2$ sono due campioni indipendenti di una gaussiana standard uno-dimensionale.


UNIVERSITÀ DI PISA

```r
x1=rep(0,1000)
x2=rep(0,1000)
i=1
while(length(x1[which(x1==0)]>0)){
  z1=runif(1,-1,1)
  z2=runif(1,-1,1)
  if(z1^2+z2^2 <= 1){
    x1[i]=z1*(-2*log(z1^2+z2^2)/(z1^2+z2^2)^2)^(1/2)
    x2[i]=z2*(-2*log(z1^2+z2^2)/(z1^2+z2^2)^2)^(1/2)
    i=i+1
  }
}
hist(x1, 700, xlim=c(-5,5), ylim=c(0,0.4), freq=F, main="Box-Muller method"
lines(sort(x1),dnorm(sort(x1),0,1),type = "l", col="red")
hist(x2, 700, xlim=c(-5,5), ylim=c(0,0.4), freq=F, main="Box-Muller method"
lines(sort(x2),dnorm(sort(x2),0,1),type = "l", col="red")
```





UNIVERSITÀ DI PISA

# Rejection Sampling

Per i casi in cui la cdf è molto complicata da trattare, una valida alternativa è il **Rejection Sampling**.

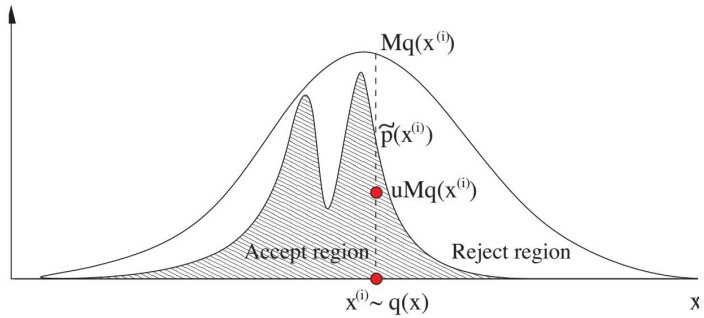
Supponiamo di voler ottenere un campionamento a partire da una distribuzione di densità $p(x)$. Sia $\tilde{p}(x)$ la sua versione non normalizzata, cioè $p(x) = \tilde{p}(x)/Z_p$. Prendiamo un'altra densità $q(x)$ che soddisfa la condizione $Mq(x) \geq \tilde{p}(x)$ per qualche costante $M > 0$. Tale $q(x)$ è detta **proposal distribution**.

# Rejection Sampling

Per i casi in cui la cdf è molto complicata da trattare, una valida alternativa è il **Rejection Sampling**.

Supponiamo di voler ottenere un campionamento a partire da una distribuzione di densità $p(x)$. Sia $\tilde{p}(x)$ la sua versione non normalizzata, cioè $p(x) = \tilde{p}(x)/Z_p$. Prendiamo un'altra densità $q(x)$ che soddisfa la condizione $Mq(x) \geq \tilde{p}(x)$ per qualche costante $M > 0$. Tale $q(x)$ è detta **proposal distribution**.

- Campioniamo $x \sim q(x)$ (random $x$) e $u \sim U(0, 1)$ (random $y$)
- Se $uMq(x) > \tilde{p}(x)$ rifiutiamo $x$
- Altrimenti accettiamo $x$



# Rejection Sampling

La procedura appena presentata è corretta.

## Proof.

Siano

$$S = \{(x,u) \ : \ uMq(x) \leq \tilde{p}(x)\}, \quad S_0 = \{(x,u) \ : \ x \leq x_0, \ uMq(x) \leq \tilde{p}(x)\}$$

Allora, la cdf dei punti accettati è

$$P_q(x \leq x_0 \mid x \text{ accettato}) = \frac{P_q(x \leq x_0, \ x \text{ accettato})}{P_q(x \text{ accettato})}$$
$$= \frac{\int\int \mathbb{1}_{\{(x,u)\in S_0\}} q(x) du dx}{\int\int \mathbb{1}_{\{(x,u)\in S\}} q(x) du dx} = \frac{\int_{-\infty}^{x_0} \tilde{p}(x) dx}{\int_{-\infty}^{\infty} \tilde{p}(x) dx} = P_p(x \leq x_0)$$

dove abbiamo usato che

$$\int\int \mathbb{1}_{\{(x,u)\in S_0\}} q(x) du dx = \int_{-\infty}^{x_0} q(x) \left( \int_0^{\frac{\tilde{p}(x)}{Mq(x)}} du \right) dx = \frac{1}{M} \int_{-\infty}^{x_0} \tilde{p}(x) dx.$$

# Rejection Sampling

Quanto è efficiente il metodo? Dipende da quanto sia alta la probabilità di accettare il campione $x$, così da effettuare meno rifiuti e ottenere il campionamento in meno passaggi. Infatti,

$$P(x \text{ accettato}) = \int \int \mathbb{1}_{\{(x,u)\in S\}} q(x) du dx = \int \frac{\tilde{p}(x)}{Mq(x)} q(x) dx = \frac{1}{M} \int \tilde{p}(x) dx.$$

# Rejection Sampling

Quanto è efficiente il metodo? Dipende da quanto sia alta la probabilità di accettare il campione $x$, così da effettuare meno rifiuti e ottenere il campionamento in meno passaggi. Infatti,

$$P(x \text{ accettato}) = \int \int \mathbb{1}_{\{(x,u)\in S\}} q(x) du dx = \int \frac{\tilde{p}(x)}{Mq(x)} q(x) dx = \frac{1}{M} \int \tilde{p}(x) dx.$$

Quindi, vogliamo scegliere $M$ più piccolo possibile, ma che continui a verificare la condizione $Mq(x) \geq \tilde{p}(x)$.


UNIVERSITÀ DI PISA

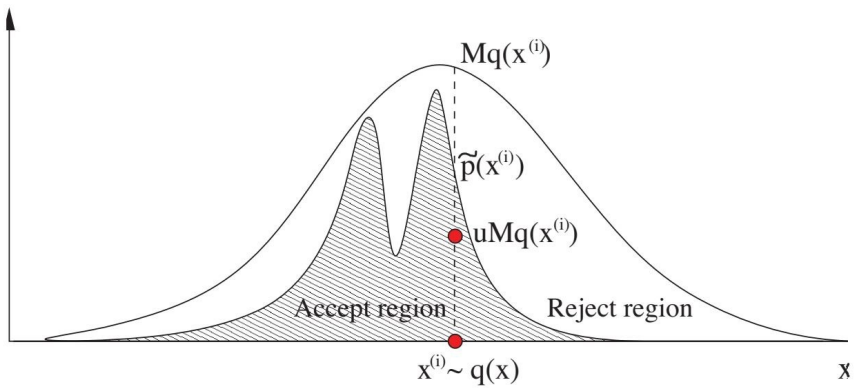
# Rejection Sampling

Quanto è efficiente il metodo? Dipende da quanto sia alta la probabilità di accettare il campione $x$, così da effettuare meno rifiuti e ottenere il campionamento in meno passaggi. Infatti,

$$P(x \text{ accettato}) = \int \int \mathbb{1}_{\{(x,u)\in S\}} q(x) du dx = \int \frac{\tilde{p}(x)}{Mq(x)} q(x) dx = \frac{1}{M} \int \tilde{p}(x) dx.$$

Quindi, vogliamo scegliere $M$ più piccolo possibile, ma che continui a verificare la condizione $Mq(x) \geq \tilde{p}(x)$.

Ciò è coerente anche con l'interpretazione geometrica della regione rifiutata.



# Rejection Sampling

Vediamo un esempio di *Rejection Sampling* a partire da una distribuzione **Gamma** di paramteri $\alpha$ e $\lambda$, la cui densità è pari a

$$Ga_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x).$$

Per $\alpha = k$ intero si può usare il fatto noto che se $X_i \overset{iid}{\sim} Expon_\lambda$ e $Y = X_1 + \cdots + X_k$, allora $Y \sim Ga_{k,\lambda}$.

Per $\alpha$ non intero si può usare il il metodo del *rejection sampling*. Ad esempio, possiamo usare come *proposal distribution* $q(x) = Ga_{k,\lambda-1}(x)$, dove $k = \lfloor \alpha \rfloor$. Cerchiamo, ora, l'$M$ ottimale, ovvero il più piccolo $M$ tale che $Mq(x) \geq \tilde{p}(x)$. Ciò coincide con trovare il massimo del rapporto $\frac{p(x)}{q(x)}$.

UNIVERSITÀ DI PISA

# Rejection Sampling

Vediamo un esempio di *Rejection Sampling* a partire da una distribuzione **Gamma** di paramteri $\alpha$ e $\lambda$, la cui densità è pari a

$$Ga_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x).$$

Per $\alpha = k$ intero si può usare il fatto noto che se $X_i \overset{iid}{\sim} Expon_\lambda$ e $Y = X_1 + \cdots + X_k$, allora $Y \sim Ga_{k,\lambda}$.

Per $\alpha$ non intero si può usare il il metodo del *rejection sampling*. Ad esempio, possiamo usare come *proposal distribution* $q(x) = Ga_{k,\lambda-1}(x)$, dove $k = \lfloor \alpha \rfloor$. Cerchiamo, ora, l'$M$ ottimale, ovvero il più piccolo $M$ tale che $Mq(x) \geq \tilde{p}(x)$. Ciò coincide con trovare il massimo del rapporto $\frac{p(x)}{q(x)}$.

UNIVERSITÀ DI PISA

# Rejection Sampling

Vediamo un esempio di *Rejection Sampling* a partire da una distribuzione **Gamma** di paramteri $\alpha$ e $\lambda$, la cui densità è pari a

$$Ga_{\alpha,\lambda}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x).$$

Per $\alpha = k$ intero si può usare il fatto noto che se $X_i \overset{iid}{\sim} Expon_\lambda$ e $Y = X_1 + \cdots + X_k$, allora $Y \sim Ga_{k,\lambda}$.

Per $\alpha$ non intero si può usare il il metodo del *rejection sampling*. Ad esempio, possiamo usare come *proposal distribution* $q(x) = Ga_{k,\lambda-1}(x)$, dove $k = \lfloor \alpha \rfloor$. Cerchiamo, ora, l'$M$ ottimale, ovvero il più piccolo $M$ tale che $Mq(x) \geq \tilde{p}(x)$. Ciò coincide con trovare il massimo del rapporto $\frac{p(x)}{q(x)}$.


UNIVERSITÀ DI PISA

# Rejection Sampling

$$\frac{p(x)}{q(x)} = \frac{Ga_{\alpha,\lambda}(x)}{Ga_{k,\lambda-1}(x)} = \frac{x^{\alpha-1}\lambda^{\alpha}\exp(-\lambda x)/\Gamma(\alpha)}{x^{k-1}(\lambda-1)^{k}\exp(-(\lambda-1)x)/\Gamma(k)}$$

$$= \frac{\Gamma(k)\lambda^{\alpha}}{\Gamma(\alpha)(\lambda-1)^{k}} x^{\alpha-k}\exp(-x)$$

e assume il massimo quando $x = \alpha - k$.

# Rejection Sampling

$$\frac{p(x)}{q(x)} = \frac{Ga_{\alpha,\lambda}(x)}{Ga_{k,\lambda-1}(x)} = \frac{x^{\alpha-1}\lambda^{\alpha}\exp(-\lambda x)/\Gamma(\alpha)}{x^{k-1}(\lambda-1)^{k}\exp(-(\lambda-1)x)/\Gamma(k)}$$

$$= \frac{\Gamma(k)\lambda^{\alpha}}{\Gamma(\alpha)(\lambda-1)^{k}} x^{\alpha-k}\exp(-x)$$

e assume il massimo quando $x = \alpha - k$.

Si può fare di meglio, invece, se come *proposal distriution* si prende la distribuzione di Cauchy di parametri $x_0, y_0$, ovvero $q(x) = \frac{1}{\pi}\frac{y_0}{(x-x_0)^2+y_0^2}$.


UNIVERSITÀ DI PISA

# Rejection Sampling

$$\frac{p(x)}{q(x)} = \frac{Ga_{\alpha,\lambda}(x)}{Ga_{k,\lambda-1}(x)} = \frac{x^{\alpha-1}\lambda^{\alpha}\exp(-\lambda x)/\Gamma(\alpha)}{x^{k-1}(\lambda-1)^{k}\exp(-(\lambda-1)x)/\Gamma(k)}$$

$$= \frac{\Gamma(k)\lambda^{\alpha}}{\Gamma(\alpha)(\lambda-1)^{k}} x^{\alpha-k}\exp(-x)$$

e assume il massimo quando $x = \alpha - k$.

Si può fare di meglio, invece, se come *proposal distriution* si prende la distribuzione di Cauchy di parametri $x_0, y_0$, ovvero $q(x) = \frac{1}{\pi}\frac{y_0}{(x-x_0)^2+y_0^2}$.

Allo stesso modo, fissati i parametri $\alpha, \lambda, x_0, y_0$, vale che

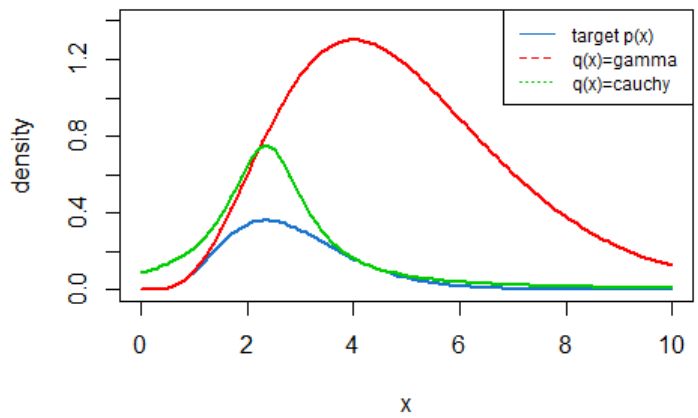
$$\frac{p(x)}{q(x)} = \frac{x^{\alpha-1}\lambda^{\alpha}\exp(-\lambda x)/\Gamma(\alpha)}{\frac{1}{\pi}\frac{y_0}{(x-x_0)^2+y_0^2}} = \frac{\pi\lambda^{\alpha}}{\Gamma(\alpha)y_0} x^{\alpha-1}\exp(-\lambda x)\left[(x-x_0)^2+y_0^2\right] \leq N$$

UNIVERSITÀ DI PISA

# Rejection Sampling

Fissiamo alcuni parametri per visualizzare le differenze:

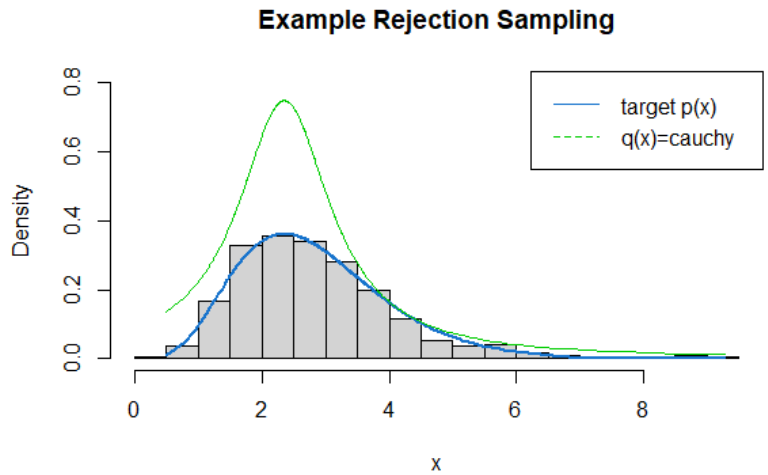
```
#parametri
alpha=5.7
lambda=2
k=5

#plot target p(x)
x=seq(0,10,0.1)
plot(x,dgamma(x,alpha,lambda),type = "l", ylab="density", ylim=c(0,1.4),
     col="dodgerblue3", lwd=2)

#proposal distribution q(x)=gamma
M=dgamma(alpha-k,alpha,lambda)/dgamma(alpha-k,k,lambda-1)
lines(x,M*dgamma(x,k,lambda-1),type = "l", col="red", lwd=2 )

#proposal distribution q(x)=cauchy
x0=4.7/2
y0=1/(pi*(dgamma(x0,alpha,lambda)))
N=2.07
lines(x,N*dcauchy(x,x0,y0),type = "l", col="green3", lwd=2)

legend("topright", legend=c("target p(x)", "q(x)=gamma", "q(x)=cauchy"),
       col=c("dodgerblue3","red","green3"),lty=1:3,  cex = 0.75)
```



# Rejection Sampling

Alla luce di quanto ottenuto, proviamo a calcolare un campione con tale metodo.

```
x=rep(0,500)
i=1
while(length(x[which(x==0)]>0)){
y=rcauchy(1,x0,y0)
u=runif(1,0,1)
if(u*N*dcauchy(y,x0,y0) <= dgamma(y,alpha,lambda)){
  x[i]=y
  i=i+1
}
}
hist(x,15,freq=F,ylim=c(0,0.8), main="Example Rejection Sampling")
lines(sort(x),dgamma(sort(x),alpha,lambda),type = "l", col="dodgerblue3",lwd=2)
lines(sort(x),N*dcauchy(sort(x),x0,y0),type = "l", col="green3")
legend("topright", legend=c("target p(x)", "q(x)=cauchy"),
       col=c("dodgerblue3","green3"),lty=1:2)
```



# Importance Sampling

**Importance Sampling** è un metodo Monte Carlo utilizzato per approssimare integrali della forma

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx.$$

Avevamo visto il metodo dell'approssimazione Monte Carlo in cui

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

e $x_s \sim p(x)$, la stessa distribuzione di $X$.

Nell'*Importance Sampling*, invece, l'idea è quella di campionare $x$ rispetto ad un'altra distribuzione $q(x)$, detta **proposal distribution**, e poi procedere con l'approssimazione.

UNIVERSITÀ DI PISA

# Importance Sampling

**Importance Sampling** è un metodo Monte Carlo utilizzato per approssimare integrali della forma

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx.$$

Avevamo visto il metodo dell'approssimazione Monte Carlo in cui

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

e $x_s \sim p(x)$, la stessa distribuzione di $X$.

Nell'*Importance Sampling*, invece, l'idea è quella di campionare $x$ rispetto ad un'altra distribuzione $q(x)$, detta **proposal distribution**, e poi procedere con l'approssimazione.

UNIVERSITÀ DI PISA

# Importance Sampling

**Importance Sampling** è un metodo Monte Carlo utilizzato per approssimare integrali della forma

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx.$$

Avevamo visto il metodo dell'approssimazione Monte Carlo in cui

$$\mathbb{E}[f(X)] = \int f(x)p(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} f(x_s)$$

e $x_s \sim p(x)$, la stessa distribuzione di $X$.

Nell'*Importance Sampling*, invece, l'idea è quella di campionare $x$ rispetto ad un'altra distribuzione $q(x)$, detta **proposal distribution**, e poi procedere con l'approssimazione.

UNIVERSITÀ DI PISA

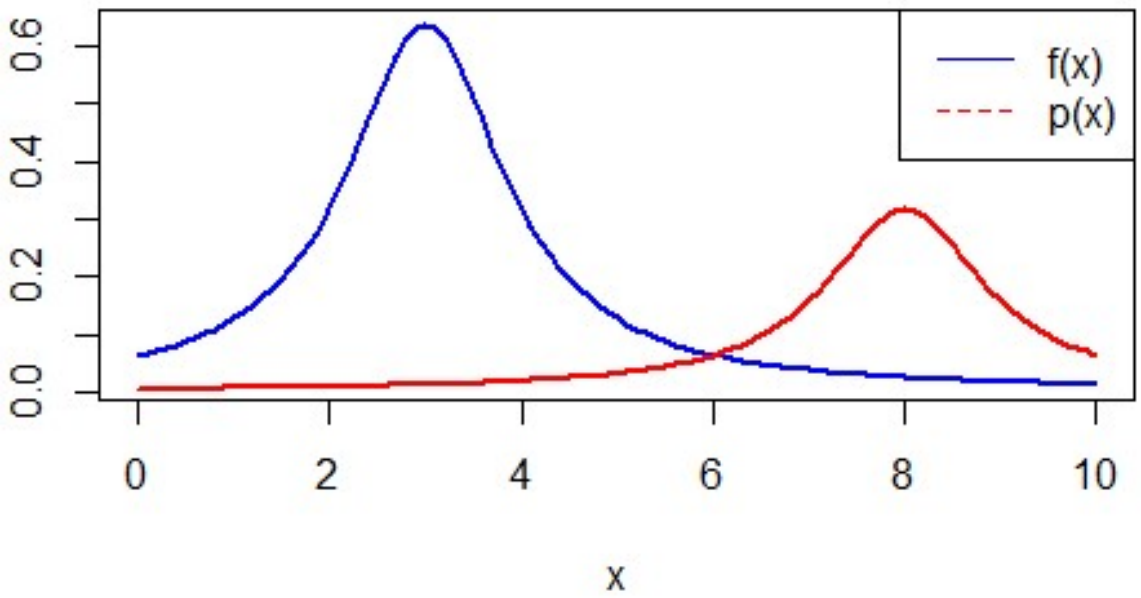
# Importance Sampling

Come scegliere $q(x)$? Poiché il nostro obiettivo è approssimare l'integrale di $f(x)p(x)$, l'idea è quella di scegliere una densità $q(x) \propto f(x)p(x)$ Ciò significa avere una densità alta non solo quando è alta $p(x)$, ma anche quando è grande $|f(x)|$, in modo da ottenere maggiore efficienza.

UNIVERSITÀ DI PISA

# Importance Sampling

Come scegliere $q(x)$? Poiché il nostro obiettivo è approssimare l'integrale di $f(x)p(x)$, l'idea è quella di scegliere una densità $q(x) \propto f(x)p(x)$ Ciò significa avere una densità alta non solo quando è alta $p(x)$, ma anche quando è grande $|f(x)|$, in modo da ottenere maggiore efficienza.



# Importance Sampling

A questo punto, otteniamo che

$$I = \mathbb{E}[f(X)] = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} w_s f(x_s) = \hat{I}$$

dove $w_s = \frac{p(x_s)}{q(x_s)}$ sono detti **importance weights** e $x_s \sim q(x)$.

1. Possiamo fare questi passaggi se la distribuzione $q(x)$ soddisfa la proprietà che $q(x) = 0 \Rightarrow p(x) = 0$, cioè $p$ assolutamente continua rispetto a $q$.
2. Come scegliere $q(x)$? Se non ho in mente nessuna $f(x)$ in particolare, conviene scegliere $q(x)$ più simile possibile a $p(x)$. Negli altri casi, un criterio naturale è quello di minimizzare la varianza dell'approssimazione $\hat{I}$, così da rendere il metodo più efficiente rispetto al Monte Carlo standard.

UNIVERSITÀ DI PISA

# Importance Sampling

A questo punto, otteniamo che

$$I = \mathbb{E}[f(X)] = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} w_s f(x_s) = \hat{I}$$

dove $w_s = \frac{p(x_s)}{q(x_s)}$ sono detti **importance weights** e $x_s \sim q(x)$.

1. Possiamo fare questi passaggi se la distribuzione $q(x)$ soddisfa la proprietà che $q(x) = 0 \Rightarrow p(x) = 0$, cioè $p$ assolutamente continua rispetto a $q$.
2. Come scegliere $q(x)$? Se non ho in mente nessuna $f(x)$ in particolare, conviene scegliere $q(x)$ più simile possibile a $p(x)$. Negli altri casi, un criterio naturale è quello di minimizzare la varianza dell'approssimazione $\hat{I}$, così da rendere il metodo più efficiente rispetto al Monte Carlo standard.


UNIVERSITÀ DI PISA

# Importance Sampling

A questo punto, otteniamo che

$$I = \mathbb{E}[f(X)] = \int f(x)\frac{p(x)}{q(x)}q(x)dx \approx \frac{1}{S}\sum_{s=1}^{S} w_s f(x_s) = \hat{I}$$

dove $w_s = \frac{p(x_s)}{q(x_s)}$ sono detti **importance weights** e $x_s \sim q(x)$.

1. Possiamo fare questi passaggi se la distribuzione $q(x)$ soddisfa la proprietà che $q(x) = 0 \Rightarrow p(x) = 0$, cioè $p$ assolutamente continua rispetto a $q$.
2. Come scegliere $q(x)$? Se non ho in mente nessuna $f(x)$ in particolare, conviene scegliere $q(x)$ più simile possibile a $p(x)$. Negli altri casi, un criterio naturale è quello di minimizzare la varianza dell'approssimazione $\hat{I}$, così da rendere il metodo più efficiente rispetto al Monte Carlo standard.


UNIVERSITÀ DI PISA

# Importance Sampling

Ora,

$$var_{q(x)}[f(x)w(x)] = \mathbb{E}_{q(x)}[f^2(x)w^2(x)] - I^2.$$

Ma poiché $I$ è indipendente da $q$ si può ignorare. In più, per la disuguaglianza di Jensen, abbiamo la stima

$$\mathbb{E}_{q(x)}[f^2(x)w^2(x)] \geq \left(\mathbb{E}_{q(x)}[|f(x)w(x)|]\right)^2 = \left(\int |f(x)|p(x)dx\right)^2,$$

e l'ultimo termine non dipende da $q$. In particolare è possibile raggiungere tale lower bound se si prende come *proposal distribution*

$$q^*(x) = \frac{|f(x)|p(x)}{\int |f(x')|p(x')dx'}.$$

UNIVERSITÀ DI PISA

# Importance Sampling

Spesso, la distribuzione target è facile da valutare ma non da integrare e pertanto si considera la relativa distribuzione non normalizzata $p(x) = \frac{1}{Z_p}\tilde{p}(x)$. É ragionevole considerare anche la *proposal distribution* non normalizzata $\tilde{q}(x)$ con costante di normalizzazione $Z_q$.
Allora,

$$\mathbb{E}[f(X)] = \frac{Z_q}{Z_p}\int f(x)\frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \approx \frac{Z_q}{Z_p}\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s) = \hat{I}$$

dove $\tilde{w}_s = \frac{\tilde{p}(x_s)}{\tilde{q}(x_s)}$.

# Importance Sampling

Spesso, la distribuzione target è facile da valutare ma non da integrare e pertanto si considera la relativa distribuzione non normalizzata $p(x) = \frac{1}{Z_p}\tilde{p}(x)$. É ragionevole considerare anche la *proposal distribution* non normalizzata $\tilde{q}(x)$ con costante di normalizzazione $Z_q$.
Allora,

$$\mathbb{E}[f(X)] = \frac{Z_q}{Z_p}\int f(x)\frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \approx \frac{Z_q}{Z_p}\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s) = \hat{I}$$

dove $\tilde{w}_s = \frac{\tilde{p}(x_s)}{\tilde{q}(x_s)}$.

Ora, poiché non conosciamo $Z_p, Z_q$ possiamo approssimare anche loro:

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(x)dx = \int \frac{\tilde{p}(x)}{\tilde{q}(x)}q(x)dx \approx \frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s$$

UNIVERSITÀ DI PISA

# Importance Sampling

Infine, abbiamo ottenuto che

$$\hat{I} = \frac{Z_q}{Z_p}\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s) = \frac{\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s)}{\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s} = \sum_{s=1}^{S} w_s f(x_s)$$

dove $w_s = \frac{\tilde{w}_s}{\sum_{s'}\tilde{w}_{s'}}$, sono detti **normalized importance weights**.

É chiaro come nel caso non normalizzato il metodo consiste nel compiere due approssimazioni e pertanto risulterà meno efficace rispetto al primo caso.

# Importance Sampling

Infine, abbiamo ottenuto che

$$\hat{I} = \frac{Z_q}{Z_p}\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s) = \frac{\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s f(x_s)}{\frac{1}{S}\sum_{s=1}^{S}\tilde{w}_s} = \sum_{s=1}^{S} w_s f(x_s)$$

dove $w_s = \frac{\tilde{w}_s}{\sum_{s'}\tilde{w}_{s'}}$, sono detti **normalized importance weights**.

É chiaro come nel caso non normalizzato il metodo consiste nel compiere due approssimazioni e pertanto risulterà meno efficace rispetto al primo caso.

*GRAZIE PER L'ATTENZIONE*


UNIVERSITÀ DI PISA