PSC: Progettazione di sistemi di controllo	a.a. 2010-2011
Lezione 26 — 11 Gennaio	
Docente: Luca Schenato	*Stesori: Dario Volpato, Giorgio Pattarello*

26.1 Importance Sampling e Sequential Sampling

26.1.1 Importance Sampling

L'importance Sampling e una tecnica molto utile nell'ambito dell'applicabilità dei filtri particellari; molte volte non è possibile implementare degli algoritmi che possono generare campioni da una densità di probabilità generica e la tecnica dell'importance sampling risolve questo problema.
Sia $p(\cdot)$ una densità di probabilità che si vuole approssimare attraverso una serie di funzioni δ:

$$p(x) \cong \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i) = \hat{p}(x)$$

dove $x \in \mathbb{R}^N$ ed $x_i \sim p(x)$ sono i.i.d.

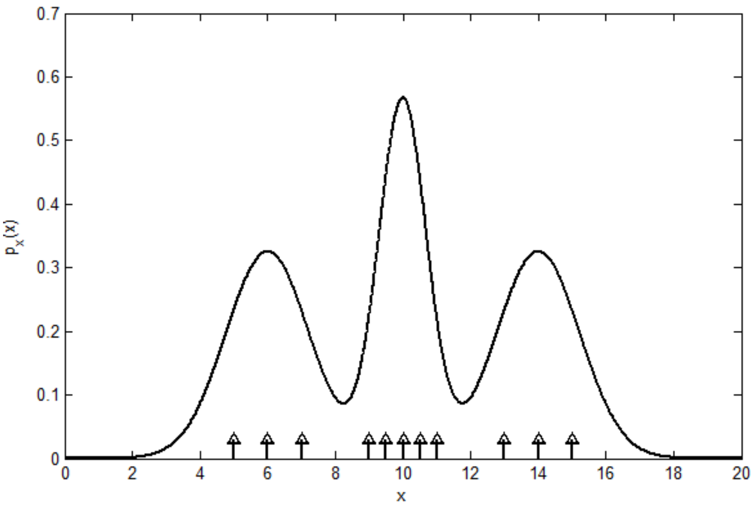

Figura 26.1. Approssimazione di una densità tramite funzioni δ

L'obiettivo quello di calcolare:

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

che può essere approssimata dalla media campionaria:

$$\begin{aligned}\mathbb{E}_{\hat{p}}[f(x)] &= \frac{1}{N}\int\sum_{i=1}^{N} f(x)\delta(x-x_i)dx \\ &= \frac{1}{N}\sum_{i=1}^{N} f(x_i) = \hat{\mu}_p(f)\end{aligned}$$

Si faccia presente che, per la legge dei grandi numeri, vale:

$$\lim_{N\to+\infty}\ \hat{\mu}_p = \mu_p \qquad q.c.$$

Dalla diseguaglianza di Chebychev si evince, dal calcolo della varianza di $f(\cdot)$, che l'incertezza dell'approssimazione della media di $f(\cdot)$ diminuisce al crescere di N e si ricava inoltre la stima di quanti campioni N utilizzare, una volta fissato il valore di tolleranza ϵ.

$$\mathbb{P}[|\hat{\mu}-\mu| > \epsilon] \leq \ \frac{1}{N}\frac{Var_p[f(x)]}{\epsilon^2}$$

A causa delle difficoltà di calcolo di $Var_p[f(x)]$ si può calcolare il valore approssimato da inserire nella disuguaglianza di Chebychev:

$$\widehat{Var}_p[f(x)] = \frac{1}{N}\sum_{i=1}^{N}|f(x_i)-\hat{\mu}_p(f)|^2$$

ESEMPIO 1:
Sia $p(x)$:

$$p(x) = c\,\mathrm{sinc}^2(x) = c\frac{\sin^2(x)}{x^2}$$

dove c è la costante di normalizzazione, cioè $1/c = \int \mathrm{sinc}(x)\,dx$.

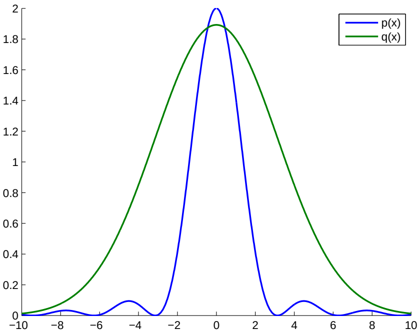

Figura 26.2. Importance sampling con $p(x) = c\,\mathrm{sinc}^2$ e $q(x)$ gaussiana

questa densità di probabilità è impossibile da campionare in Matlab quindi l'operazione di approssimazione tramite il campionamento si effettua attraverso l'uso di un'altra densità di probabilità, per esempio una densità $q(x)$ gaussiana.

Se la nuova densità di probabilità scelta $q(x)$ soddisfa la proprietà:

$$p(x) > 0 \qquad \Rightarrow \qquad q(x) > 0$$

si ha che:

$$\begin{aligned}\mu_p(f) &= \int f(x)p(x)dx \\ &= \int f(x)p(x)\frac{q(x)}{q(x)}dx \\ &= \int f(x)\pi(x)q(x)dx = \mathbb{E}_q[f(x)\pi(x)]\end{aligned}$$

dove $\pi(x) = \frac{p(x)}{q(x)}$. Questa quantità puo' essere approssimata tramite la sua media campionaria $\hat{\mu}_q(f\pi)$ secondo la densità $q(x)$:

$$\mathbb{E}_q[f(x)\pi(x)] \cong \frac{1}{N}\sum_{i=1}^{N} f(x_i)\pi(x_i) := \hat{\mu}_q(f\pi)$$

dove i $\pi(x_i)$ sono detti importance weight e $x_i \sim q(x)$ i.i.d.
Per la legge dei grandi numeri, come precedentemente visto, vale

$$\lim_{N\to+\infty} \hat{\mu}_p = \mu_p \qquad q.c.$$

ma anche

$$\lim_{N\to+\infty} \hat{\mu}_q(f\pi) = \mu_p \qquad q.c.$$

Con questa implementazione, le prestazioni peggiorano a parità di numero di campioni, quindi bisogna valutare quanti campioni utilizzare.

$$\mathbb{P}[\ |\hat{\mu}_q(f\pi) - \mu_p(f)| > \epsilon\ ]\ \leq\ \frac{1}{N}\frac{c(f,p,q)}{\epsilon^2}\ \leq\ \frac{1}{N}\frac{c}{\epsilon^2}$$

Nella disuguaglianza precedente $c(f,p,q)$ è una costante che dipende da $f(x), p(x), q(x)$.
Questo tipo di approcio nella realtà viene modificato: qui si suppone di conoscere sia le densità di probabilità che i loro coefficienti di normalizzazione, informazioni che, in molti casi, risultano difficili da ottenere.
Si suppone quindi $p(x) \propto \gamma(x)$ e $q(x) \propto \beta(x)$, si sa inoltre campionare $q(x)$. Da $\pi(x) = p(x)/q(x)$ si ha che

$$\mathbb{E}_q[\pi] = \int \frac{p(x)}{q(x)}q(x)dx = 1$$

che campionata diventa

$$\mathbb{E}_q[\pi] \cong \mu_q(\pi) := \frac{1}{N}\sum_{i=1}^{N} \pi(x_i)$$

con $x_i \sim q(x)$ i.i.d.. La varianza risulta:

$$\begin{aligned} Var_q(\pi) &= \int (\pi(x) - 1)^2 q(x)dx \\ &= \int \pi(x)^2 q(x) + q(x) - 2\pi(x)q(x)dx \\ &= \int \frac{p^2(x)}{q(x)}dx - 1 \geq 0 \end{aligned}$$

Si definisce quindi

$$\hat{\hat{\mu}}_q(f) = \frac{\frac{1}{N}\sum_{i=1}^{N} \pi(x_i)f(x_i)}{\frac{1}{N}\sum_{i=1}^{N} \pi(x_i)}$$

che a livello intuitivo per $N \to +\infty$ converge ad $\mu_p(f)$ in quanto il denominatore converge a $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \pi(x_i) = 1$ ed il numeratore a $\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \pi(x_i)f(x_i) = \mu_p(f)$.
Si può riscrivere $\hat{\hat{\mu}}_q$ anche nel seguente modo:

$$\hat{\hat{\mu}}_q(f) = \sum_{i=1}^{N} w_i f(x_i)$$

dove

$$w_i = \frac{\pi(x_i)}{\sum_{i=1}^{N} \pi(x_i)} \geq 0 \qquad \text{con} \qquad \sum_i w_i = 1$$

Riscritta in questo modo si capisce meglio il vantaggio che si ha nell'usare questa implementazione, infatti si può notare che definendo

$$\tilde{\pi}(x) = \frac{\gamma(x)}{\beta(x)}$$

si vede facilmente che $\tilde{\pi}(x) \propto \pi(x)$ e quindi w_i risulta:

$$w_i = \frac{\pi(x_i)}{\sum_{i=1}^{N} \pi(x_i)} = \frac{\tilde{\pi}(x_i)}{\sum_{i=1}^{N} \tilde{\pi}(x_i)}$$

Questa ultima uguaglianza è importante perchè mostra che come di calcolare $\hat{\hat{\mu}}_q(f)$ senza conoscere le costante di normalizzazione e questo risulta essere un enorme vantaggio.
Le prestazioni di questo stimatore sono peggiori rispetto al caso generale ed inoltre non si ottiene nemmeno uno stimatore che è corretto solo asintoticamente:

$$\lim_{N\to+\infty} N\big(\mathbb{E}[\hat{\hat{\mu}}_q(f)] - \mu_p(f)\big) = c$$

dove BIAS$(\hat{\hat{\mu}}_q(f)) = \mathbb{E}[\hat{\hat{\mu}}_q(f)] - \mu_p(f)$ è il bias dello stimatore.
Per risolvere ora la disuguaglianza di Chebychev occorre calcolare il membro di destra della disuguaglianza e per ricavarlo si calcola:

$$\mathbb{E}[|\hat{\hat{\mu}}_q(f) - \mu_p(f)|^2] = \text{BIAS}^2(\hat{\hat{\mu}}_q(f)) + Var(\hat{\hat{\mu}}_q(f))$$

con $BIAS^2$ che diventa trascurabile rispetto alla $Var(\hat{\hat{\mu}}_q(f))$ perchè con $N \to +\infty$

$$BIAS \cong \frac{c_1^2}{N^2} \text{ e } Var(\hat{\hat{\mu}}_q(f)) \cong \frac{c_2^2}{N}$$

La $Var(\hat{\hat{\mu}}_q(f))$ è difficile da calcolare, ma si può dimostrare almeno nel caso $p(x) \approx q(x)$, che la varianza è approssimabile a:

$$Var(\hat{\hat{\mu}}_q(f)) \cong \frac{1}{N} Var_q(f\pi) \cong \frac{1}{N} Var_p(f)(1 + Var_q(\pi))$$

Si ottiene quindi la seguente disuguglianza di Chebychev

$$\mathbb{P}[|\hat{\hat{\mu}}_q(f) - \mu_p(f)|^2] \lessapprox \frac{1}{N}\frac{Var_p(f)}{\epsilon^2}(1 + Var_q(\pi))$$

ESEMPIO 2:
Sia $x = (x_1, \cdots, x_n)$ un vettore con $x_i \in \mathbb{R}$ $\forall i$ e sia $\bar{p}(x)$ una densità di probabilità gaussiana di media nulla e varianza unitaria. definiamo la densità $p(x)$ come segue:

$$p(x) = p(x_1, \cdots, x_n) = \prod_{i=1}^{N} \bar{p}(x_i) = \prod_{i=1}^{N} \mathcal{N}_{x_i}(0, 1) \tag{26.1}$$

sia inoltre $\bar{q}(x)$ simile ad $\bar{p}(x)$ ma con varianza leggermente differente dall'unità:

$$q(x) = \prod_{i=1}^{N} \bar{q}(x_i) = \prod_{i=1}^{N} \mathcal{N}_{x_i}(0, \sigma^2) \qquad \text{con} \qquad \sigma^2 \simeq 1 \tag{26.2}$$

Si vuole calcolare $Var(\hat{\hat{\mu}}_q(f))$, sapendo che:

$$\mathcal{N}_{x_i}(0, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2}\frac{x^2}{\sigma^2}) \tag{26.3}$$

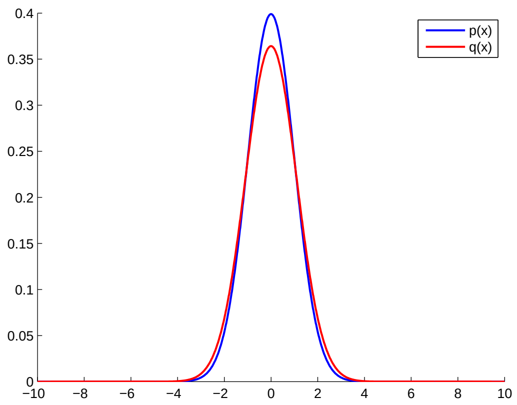

Figura 26.3. Importance sampling con $p(x)$ e $q(x)$ gaussiane

Si può ricavare che

$$
\begin{aligned}
Var(\hat{\mu}_q(f)) &= \int \frac{p^2(x)}{q(x)}dx - 1 \\
&= \int \frac{\prod_i \mathcal{N}_{x_i}(0,1)}{\prod_i \mathcal{N}_{x_i}(0,\sigma^2)}dx - 1 \\
&= \int \frac{[\prod_i \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}x_i^2)]^2}{\prod_i \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\frac{x_i^2}{\sigma^2})}dx - 1 \\
&= \int \frac{[\frac{1}{(2\pi)^n} \exp(-\sum_i x_i^2)]^2}{\frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} \exp(-\frac{1}{2\sigma^2}\sum_i x_i^2)}dx - 1 \\
&= \int \frac{1}{(2\pi)^{\frac{n}{2}}}\sigma^n \exp(-\frac{1}{2}\sum_i (2-\frac{1}{\sigma^2})x_i dx - 1 \\
&= \left[\int \frac{\sigma}{\sqrt{2\pi}} \exp(-\frac{1}{2\sigma^2}(2\sigma^2-1)x_i^2 dx\right]^n - 1 \\
&= \left(\frac{\sigma}{\sqrt{2\pi}}\sqrt{2\pi}\sqrt{\frac{\sigma^2}{2\sigma^2-1}}\right)^n - 1 \\
&= \left(\frac{\sigma^4}{2\sigma^2-1}\right)^{\frac{n}{2}} - 1
\end{aligned}
$$

Come si può vedere dal risultato finale per $\sigma = 1$ si ottiene $Var(\hat{\mu}_q(f)) = 0$ come è ovvio risulti in quanto p e q sarebbero uguali, mentre se $p(x) \simeq q(x)$ allora $Var(\hat{\mu}_q(f))$ cresce esponenzialmente al cresce di n, quindi con un numero di campioni elevato sarebbe meglio avere una densità di probabilità $q(x)$ il più possibile simile a $p(x)$, per mantenere la $Var(\hat{\mu}_q(f))$ bassa e quindi ottenere delle buone prestazioni.

26.1.2 Sequential Sampling

Nel caso di variabili aleatorie con $p(x) = p(x_1, \cdots, x_n)$ e cioè con $x \in \mathbb{R}^n$, $p(x)$ può essere scritta come:

$$\begin{aligned} p(x) &= p(x_1, \cdots, x_n) \\ &= p(x_2, \cdots, x_n | x_1) p(x_1) \\ &= p(x_3, \cdots, x_n | x_n) p(x_2 | x_1) p(x_1) \\ &= \prod_{k=2}^{n} p(x_k | x_{k-1}, \cdots, x_1) p(x_1) \end{aligned}$$

allo stesso modo, procedendo in modo analogo all'importance sampling, si può scrivere $q(x)$ come

$$q(x) = \prod_{k=2}^{n} q(x_k | x_{k-1}, \cdots, x_1) q(x_1)$$

avendo

$$\pi(x) = \frac{p(x)}{q(x)} = \prod_{k=2}^{n} \frac{p(x_k | x_{k-1}, \cdots, x_1)}{q(x_k | x_{k-1}, \cdots, x_1)} \quad \frac{p(x_1)}{q(x_1)}$$

dove

$$\pi_1(x) = \frac{p(x_1)}{q(x_1)} \qquad \text{e} \qquad \pi_k(x) = \frac{p(x_k | x_{k-1}, \cdots, x_1)}{q(x_k | x_{k-1}, \cdots, x_1)}$$

si ha che

$$\pi(x) = \prod_{k=1}^{n} \pi_k(x)$$

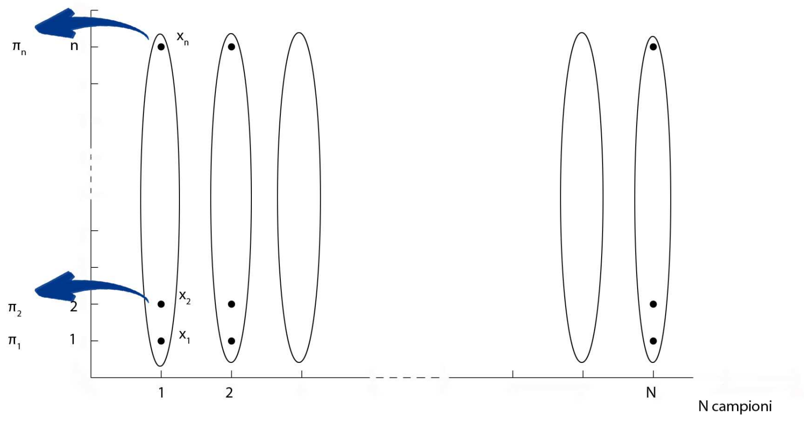

Figura 26.4. Sequential sampling

Come si può notare in Fig. 26.4 vengono generati in modo sequenziale N campioni di dimensione n: genero N campioni $i = 1, \ldots, N$ $x_1^i \sim q(x_1)$ calcolando π_1^i, poi genero $x_2^i \sim q(x_2|x_1^i)$ calcolando π_2^i, poi $x_3^i \sim q(x_3|x_1^i, x_2^i)$ calcolando π_3^i e così via.