Sampling methods

Intelligent Systems for Pattern Recognition (ISPR)

Daniele Castellana

Department of Computer Science
Università di Pisa

daniele.castellana@di.unipi.it

Outline

- **Recap probabilistic concepts**
- **Sampling**
  - What is it?
  - Why do we need it?
- **Sampling from univariate distribution**
- **Sampling from multivariate distribution**
  - Ancestor sampling
  - Gibbs Sampling
  - Monte Carlo Markov Chain (MCMC)
  - Other methods

Probability recap

- **Discrete Random Variable**
  - x a **discrete random variable** with C state;
  - $p(x = i),\ i \in [1, C]$ is its **probability distribution**;
  - $p(x_1, \ldots, x_n)$ **joint distribution** of n discrete random variable;
- **Expectation**
  - let $f(\cdot)$ a **function** over a random variable x;
  - $E_{p(x)}\left[f(x)\right] = \sum_{i=1}^{C} f(i) \times p(x = i)$ is its **expected value**;
- **Unbiased Estimator**
  Let $\mathcal{X} = \{x_1, \ldots, x_L\}$ i.i.d. samples from $p(x|\theta)$, $\hat{\theta}(\mathcal{X})$ is an **unbiased estimator** of θ if

$$E_{p(\mathcal{X}|\theta)}\left[\hat{\theta}(\mathcal{X})\right] = \theta.$$

In this lesson we will focus only on **discrete variable**, but the same holds in the continue case.

What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

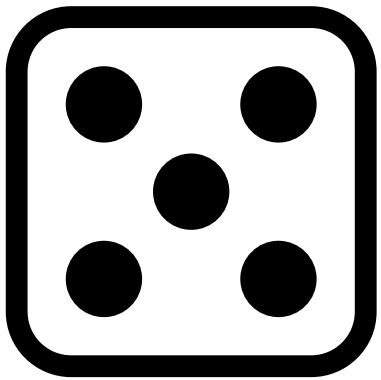
Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	$\mathbf{x_l}$
1	5

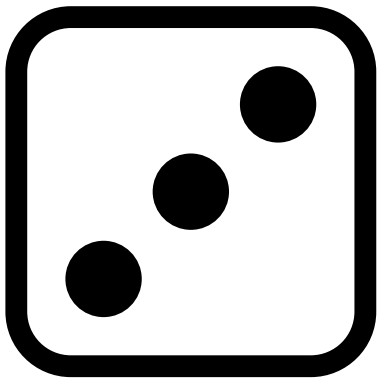


What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	x_l
1	5
2	3

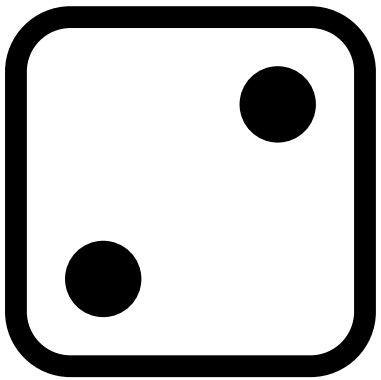


What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	$\mathbf{x_l}$
1	5
2	3
3	2

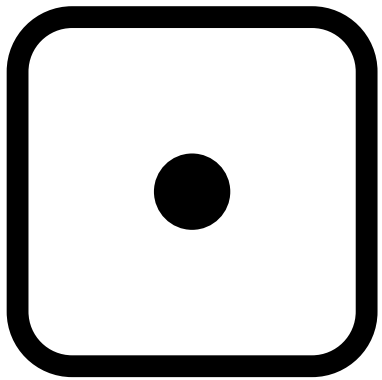


What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	$\mathbf{x_l}$
1	5
2	3
3	2
4	1

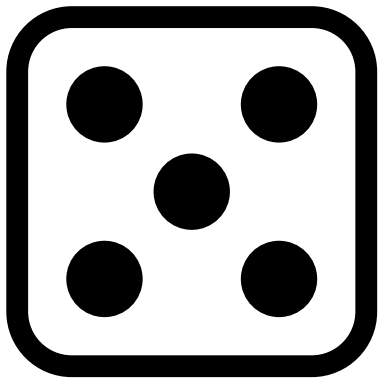


What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	x_l
1	5
2	3
3	2
4	1
5	5

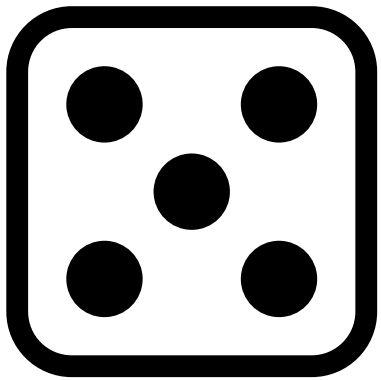


What is sampling?

Sampling consists drawing a set of **realisation** $\mathcal{X} = \{x_1, \ldots, x_L\}$ of a random variable x with distribution $p(x)$.

Example.
We would like to sample a dice: $p(x = i) = 1/6, \quad i \in [1, 6]$.

l	x_l
1	5
2	3
3	2
4	1
5	5



The **set** $\mathcal{X} = \{5, 3, 2, 1, 5\}$ contains $L = 5$ **samples**.

Why do we need sampling?

Sampling schema are useful to approximate **expectations** and **integrals**.

- if the distribution $p(x)$ is **intractable**.

In General Boltzmann Machine, the computation of Z requires **exponential time**!

- if the distribution $p(x)$ has no **closed form**.

In Bayesian statistics, the **posterior** has no closed form if a **non-conjugate priori** is used!

But why sampling?

The use of sampling is justified by two main reasons:

- the empirical distribution **converges** almost surely to the true distribution, i.e

$$\lim_{L\to\infty} \frac{1}{L} \sum_{l=1}^{L} \mathbb{I}[x^l = i] = p(x = i),$$

  where $\mathbb{I}[c] = 1$ if and only if c is true;
- the sampling approximation

$$E_{p(x)}[f(x)] \approx \frac{1}{L} \sum_{l=1}^{L} f(x_l) \equiv \hat{f}_{\mathcal{X}} \tag{1}$$

  can be an **unbiased estimator**.

Sampling Approximation

Let $\tilde{p}(\mathcal{X})$ the distribution over all possible realisations of the sampling set $\mathcal{X}$, then $\hat{f}_{\mathcal{X}}$ is an **unbiased estimator** if

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

Sampling Approximation

Let $\tilde{p}(\mathcal{X})$ the distribution over all possible realisations of the sampling set $\mathcal{X}$, then $\hat{f}_{\mathcal{X}}$ is an **unbiased estimator** if

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

This is **true** provided that $\tilde{p}(x_l) = p(x_l)$!

The proof is given in the Appendix.

Variance of Sampling Approximation
Definition

The variance of $\hat{f}(\mathcal{X})$ tell us **how much we can rely on the approximation** computed using the sampling set $\mathcal{X}$.

Let

$$\Delta\hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right], \tag{3}$$

the variance of $\hat{f}(\mathcal{X})$ is given by:

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right].$$

Variance of Sampling Approximation
Definition

The variance of $\hat{f}(\mathcal{X})$ tell us **how much we can rely on the approximation** computed using the sampling set $\mathcal{X}$.

Let

$$\Delta\hat{f}_{\mathcal{X}} = \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right], \tag{3}$$

the variance of $\hat{f}(\mathcal{X})$ is given by:

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right].$$

If the variance is **low**, $\hat{f}(\mathcal{X})$ is (quite) always **close** to $E_{p(x)}[f(x)]$!

Variance of Sampling Approximation

If we assume:

- $\tilde{p}(x_l) = p(x_l)$ (**same marginals**);
- $\tilde{p}(x_l, x_{l'}) = \tilde{p}(x_l)\tilde{p}(x_{l'})$ (**samples independence**);

we obtain

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \tag{4}$$

The proof is given in the Appendix.

Variance of Sampling Approximation

If we assume:

- $\tilde{p}(x_l) = p(x_l)$ (**same marginals**);
- $\tilde{p}(x_l, x_{l'}) = \tilde{p}(x_l)\tilde{p}(x_{l'})$ (**samples independence**);

we obtain

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \tag{4}$$

The proof is given in the Appendix.

We can reduce the variance using a **small** number of samples!

Provided that $\mathrm{Var}_{p(x)}[f(x)]$ is finite.

Small recap

So far, we have shown that:

- **we need sampling!**
- $\tilde{p}(x_l) = p(x_l)$ must holds to have a **valid sampler**.
- if $\tilde{p}(x_l, x_{l'}) = \tilde{p}(x_l)\tilde{p}(x_{l'})$ holds, **we need less samples**.

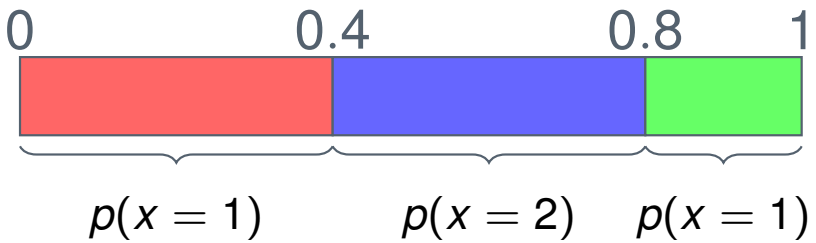
In the next slides we will introduce examples of sampling schema.

Univariate Sampling

Draw samples of an univariate variable is **easy**!

We only need a random number generator R which produces a value uniformly at random in $[0, 1]$.

$$p(x) = \begin{cases} 0.4 & x = 1 \\ 0.4 & x = 2 \\ 0.2 & x = 3 \end{cases}$$



R	x
0.19	1
0.24	1
0.47	2
0.88	3
0.73	2
0.63	2
0.52	2
0.96	3

Multivariate Sampling

In the multivariate case, $p(x)$ represents the **joint distribution** of a set of variables $\{s_1, \ldots, s_n\}$, where each s_i is a **discrete variable**.

Hence, each sample x_l contains n values.

$\mathcal{X}$	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	4	3	2	1	2
x_3	5	2	5	3	4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_L	3	5	6	6	1

Multivariate Sampling
Naive Approach 1

We build an **univariate distribution** $p(S)$, where S is a discrete variable with C^n states (i.e. **all possible combination** of s_i variable states).

S	s_1	s_2	s_3	s_4	s_5	$p(S)$
1	1	1	1	1	1	$p(1,1,1,1,1)$
2	1	1	1	1	2	$p(2,2,2,2,2)$
3	1	1	1	3	3	$p(3,3,3,3,3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C^n	C	C	C	C	C	$p(C,C,C,C,C)$

We can sample from $p(S)$ using the **univariate schema**!

Multivariate Sampling
Naive Approach 1

We build an **univariate distribution** $p(S)$, where S is a discrete variable with C^n states (i.e. **all possible combination** of s_i variable states).

S	s_1	s_2	s_3	s_4	s_5	$p(S)$
1	1	1	1	1	1	$p(1,1,1,1,1)$
2	1	1	1	1	2	$p(2,2,2,2,2)$
3	1	1	1	3	3	$p(3,3,3,3,3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C^n	C	C	C	C	C	$p(C,C,C,C,C)$

We can sample from $p(S)$ using the **univariate schema**!

This approach works only for **small values** of n!

Multivariate Sampling
Naive Approach 2

Using the **chain rule**, we can rewrite the joint distribution as:

$$p(s_1, \ldots, s_n) = p(s_1)p(s_2 \mid s_1)p(s_3 \mid s_1, s_2) \ldots p(s_n \mid s_1, \ldots, s_{n-1})$$

Them, we sample variables in the following order:

1. sample $\tilde{s}_1 \sim p(s_1)$;
2. sample $\tilde{s}_2 \sim p(s_2 \mid \tilde{s}_1)$;
3. sample $\tilde{s}_3 \sim p(s_3 \mid \tilde{s}_1, \tilde{s}_2)$;

$\vdots$

n. sample $\tilde{s}_n \sim p(s_n \mid \tilde{s}_1, \ldots, \tilde{s_{n-1}})$.

Multivariate Sampling
Naive Approach 2

Using the **chain rule**, we can rewrite the joint distribution as:

$$p(s_1, \ldots, s_n) = p(s_1)p(s_2 \mid s_1)p(s_3 \mid s_1, s_2) \ldots p(s_n \mid s_1, \ldots, s_{n-1})$$

Them, we sample variables in the following order:

1. sample $\tilde{s}_1 \sim p(s_1)$;
2. sample $\tilde{s}_2 \sim p(s_2 \mid \tilde{s}_1)$;
3. sample $\tilde{s}_3 \sim p(s_3 \mid \tilde{s}_1, \tilde{s}_2)$;

$\vdots$

n. sample $\tilde{s}_n \sim p(s_n \mid \tilde{s}_1, \ldots, \tilde{s}_{n-1})$.

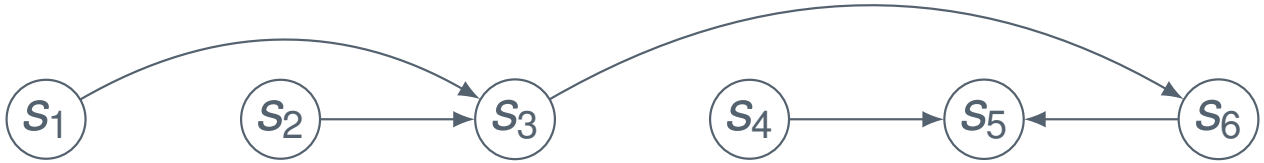
Easy because **univariate**!

Unfortunately, computing the distribution $p(s_i \mid s_{j<i})$ can require summation over an **exponential number of states**!

Ancestral Sampling

The approach used in the previous slide is called **Ancestral Sampling** (AS).

If the distribution $p(s_1, \ldots, s_n)$ is represented by a Belief Network (BN), we can apply it directly!

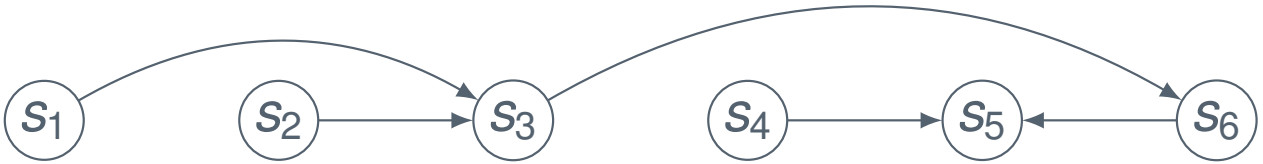


The **BN ancestral order** tell us the sampling order.

$$\{s_1, s_2, s_4\} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

Ancestral Sampling
Example

$$\{s_1, s_2, s_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$



Ancestral Sampling
Example

$$\{s_1, s_2, s_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

Sample $\tilde{s_1} \sim p(s_1)$

Ancestral Sampling
Example

$$\{\not{s}_1, s_2, \not{s}_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

Sample $\tilde{s}_4 \sim p(s_4)$

Ancestral Sampling
Example

$$\{\not{s}_1, \not{s}_2, \not{s}_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

Sample $\tilde{s}_2 \sim p(s_2)$

Ancestral Sampling
Example

$$\{\not{s}_1, \not{s}_2, \not{s}_4, \} \prec \{\not{s}_3\} \prec \{s_6\} \prec \{s_5\}$$

Sample $\tilde{s}_3 \sim p(s_3 \mid \tilde{s}_1, \tilde{s}_2)$

$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_3$ $\tilde{s}_4$

Ancestral Sampling
Example

$$\{\not{s}_1, \not{s}_2, \not{s}_4, \} \prec \{\not{s}_3\} \prec \{\not{s}_6\} \prec \{s_5\}$$

Sample $\tilde{s}_6 \sim p(s_6 \mid \tilde{s}_3)$

$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_3$ $\tilde{s}_4$ $\textcircled{s_5}$ $\tilde{s}_6$

Ancestral Sampling
Example

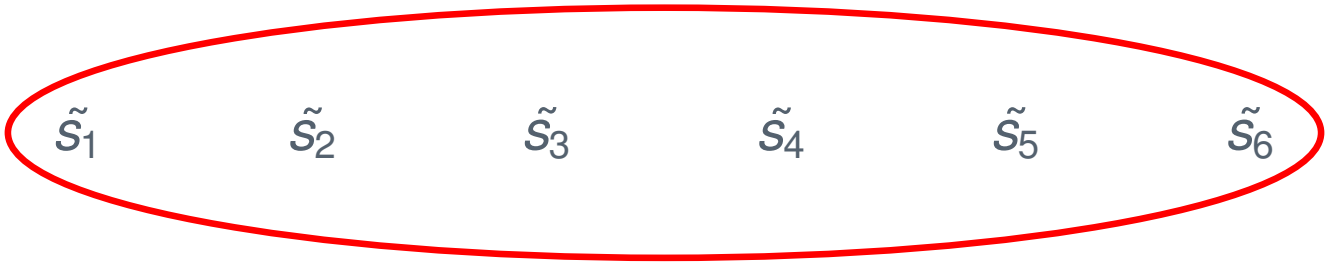
$$\{s_1, s_2, s_4, \} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

Sample $\tilde{s}_5 \sim p(s_5 \mid \tilde{s}_4, \tilde{s}_6)$

$\tilde{s}_1$ $\tilde{s}_2$ $\tilde{s}_3$ $\tilde{s}_4$ $\tilde{s}_5$ $\tilde{s}_6$

Ancestral Sampling
Example

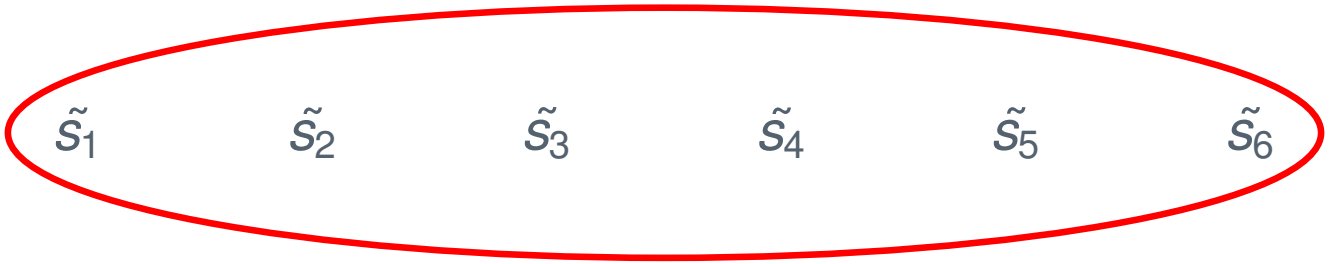
$$\{\not{s}_1, \not{s}_2, \not{s}_4, \} \prec \{\not{s}_3\} \prec \{\not{s}_6\} \prec \{\not{s}_5\}$$



This is a single sample x_l!

Ancestral Sampling
Example

$$\{s\!\!\not{1}, s\!\!\not{2}, s\!\!\not{4}, \} \prec \{s\!\!\not{3}\} \prec \{s\!\!\not{6}\} \prec \{s\!\!\not{5}\}$$



This is a single sample x_l!

AS is an **exact sampling** procedure since each sample x_l is indeed **independently** drawn from the **required distribution**.

Sampling with evidence

Suppose we that a subset of variables s_ϵ are clamped to **evidential states**; writing $s = s_\epsilon \cup s_{\setminus\epsilon}$, we would like to sample from:

$$p(s_{\setminus\epsilon} \mid s_\epsilon) = \frac{p(s_{\setminus\epsilon}, s_\epsilon)}{p(s_\epsilon)}$$

Sampling with evidence

Suppose we that a subset of variables s_ϵ are clamped to **evidential states**; writing $s = s_\epsilon \cup s_{\setminus\epsilon}$, we would like to sample from:

$$p(s_{\setminus\epsilon} \mid s_\epsilon) = \frac{p(s_{\setminus\epsilon}, s_\epsilon)}{p(s_\epsilon)}$$

Clamping variables **changes the structure** of the distribution (in previous example $s_1 \perp\!\!\!\perp s_2$, but $s_1 \not\perp\!\!\!\perp s_2 \mid s_3$).

Computing the new structure is complex as running exact inference!

Sampling with evidence

Suppose we that a subset of variables s_ϵ are clamped to **evidential states**; writing $s = s_\epsilon \cup s_{\setminus\epsilon}$, we would like to sample from:

$$p(s_{\setminus\epsilon} \mid s_\epsilon) = \frac{p(s_{\setminus\epsilon}, s_\epsilon)}{p(s_\epsilon)}$$

Clamping variables **changes the structure** of the distribution (in previous example $s_1 \perp\!\!\!\perp s_2$, but $s_1 \not\perp\!\!\!\perp s_2 \mid s_3$).

Computing the new structure is complex as running exact inference!

An alternative is to proceed with AS and then **discard** any samples which do not match the evidential states.

We discard **a lot** of samples!

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	3	1	2	4	5

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	3	1	2	4	5
x_3	3	4	2	4	5

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	3	1	2	4	5
x_3	3	4	2	4	5
x_4	3	4	2	1	5

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	3	1	2	4	5
x_3	3	4	2	4	5
x_4	3	4	2	1	5
x_5	3	4	6	1	5

Gibbs Sampling

The idea is to start from a sample $x_1 = \{s_1^1, \ldots, s_n^1\}$ and to **update only one variable** at a time.

Sample	s_1	s_2	s_3	s_4	s_5
x_1	1	1	2	4	5
x_2	3	1	2	4	5
x_3	3	4	2	4	5
x_4	3	4	2	1	5
x_5	3	4	6	1	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Gibbs Sampling
Pros

During the $(i+1)$-th iteration,

- we **select a variable** s_j;
- we **update its value** according to

$$p(s_j \mid s_{\setminus j}) = \frac{1}{Z} p(s_j \mid pa(s_j)) \prod_{k \in ch(j)} p(s_j \mid pa(s_j)),$$

where variables in $s_{\setminus j}$ are clamped to $\{s_1^i, \ldots, s_{j-1}^i, s_{j+1}^i, \ldots, s_n^i\}$

Gibbs Sampling
Pros

During the $(i+1)$-th iteration,

- we **select a variable** s_j;
- we **update its value** according to

$$p(s_j \mid s_{\setminus j}) = \frac{1}{Z} p(s_j \mid pa(s_j)) \prod_{k \in ch(j)} p(s_j \mid pa(s_j)),$$

where variables in $s_{\setminus j}$ are clamped to $\{s_1^i, \ldots, s_{j-1}^i, s_{j+1}^i, \ldots, s_n^i\}$

It depends only on the Markov blanket of s_j! **Easy to sample!**

Gibbs Sampling
Pros

During the $(i+1)$-th iteration,

- we **select a variable** s_j;
- we **update its value** according to

$$p(s_j \mid s_{\setminus j}) = \frac{1}{Z} p(s_j \mid pa(s_j)) \prod_{k \in ch(j)} p(s_j \mid pa(s_j)),$$

where variables in $s_{\setminus j}$ are clamped to $\{s_1^i, \ldots, s_{j-1}^i, s_{j+1}^i, \ldots, s_n^i\}$

It depends only on the Markov blanket of s_j! **Easy to sample!**

Dealing with evidence is easy! We just do not select a variable!

Gibbs Sampling
Cons

The Gibbs sampling draws a new sample x_l from $q(x_l \mid x_{l-1})$.

Samples are highly dependent! This lead to high variance!

Gibbs Sampling
Cons

The Gibbs sampling draws a new sample x_l from $q(x_l \mid x_{l-1})$.

Samples are highly dependent! This lead to high variance!

We are **not sampling** from $p(x)$ directly, so we cannot ensure that the sampling distribution has the same marginals of $p(x)$.

However, if we compute the limit to $l \to \infty$, the series $\{x_1, x_2, \dots\}$ **converges** to samples taken from $p(x)$!

In the limit of a large number of samples, **the Gibbs sampler is valid!**

MCMC Sampling

The idea in Markov Chain Monte Carlo (MCMC) sampling is to **build a Markov Chain** whose stationary distribution is $p(x)$.

Let $q(x^{l+1} \mid x^l)$ the transition distribution, we **must ensure** that the Markov Chain is:

- **irreducible** $\rightarrow$ it is possible to get to any state from any state;
- **aperiodic** $\rightarrow$ at each time-step we can be anywhere.

Hence, the Markov Chain has a **unique stationary distribution**.

However, there are different $q(\cdot)$ which converge to $p(\cdot)$.

MCMC Sampling

The idea in Markov Chain Monte Carlo (MCMC) sampling is to **build a Markov Chain** whose stationary distribution is $p(x)$.

Let $q(x^{l+1} \mid x^l)$ the transition distribution, we **must ensure** that the Markov Chain is:

- **irreducible** $\rightarrow$ it is possible to get to any state from any state;
- **aperiodic** $\rightarrow$ at each time-step we can be anywhere.

Hence, the Markov Chain has a **unique stationary distribution**.

However, there are different $q(\cdot)$ which converge to $p(\cdot)$.

We obtain different MCMC sampling procedure!

MCMC Sampling
Examples

There are many sampling procedure in the MCMC framework:

- Gibbs Sampling
- Metropolis-Hastings Sampling
- Hybrid Monte Carlo
- Swendson-Wang
- Slice Sampling

⋮

Each of them has different characteristics!
We should choose the most suitable for our purpose!

MCMC Sampling
Examples

There are many sampling procedure in the MCMC framework:

- Gibbs Sampling
  - Relies on **marginals** $p(s_j \mid s_{\setminus j})$!
  - Works well when variables are **not strogly related**!
- Metropolis-Hastings Sampling
- Hybrid Monte Carlo
- Swendson-Wang
- Slice Sampling

$\vdots$

Each of them has different characteristics!
We should choose the most suitable for our purpose!

MCMC Sampling
Examples

There are many sampling procedure in the MCMC framework:

- Gibbs Sampling
  - Relies on **marginals** $p(s_j \mid s_{\setminus j})$!
  - Works well when variables are **not strogly related**!
- Metropolis-Hastings Sampling
  - Relies on **unnormalised** $p^*(x) \sim p(x)$!
  - The choice of $\tilde{q}(\cdot)$ **is crucial**!
- Hybrid Monte Carlo
- Swendson-Wang
- Slice Sampling

⋮

Each of them has different characteristics!
We should choose the most suitable for our purpose!

Importance Sampling

Important Sampling (IS) approximates expectations w.r.t. $p(x)$ using **importance distribution** $q(x)$:

- we draw samples from $q(x)$;
- we assign weights s.t. $E_{p(x)}[f(x)] = \frac{1}{L}\sum_{l=1}^{L} f(x_l)w_l$.

Importance Sampling

Important Sampling (IS) approximates expectations w.r.t. $p(x)$ using **importance distribution** $q(x)$:

- we draw samples from $q(x)$;
- we assign weights s.t. $E_{p(x)}[f(x)] = \frac{1}{L}\sum_{l=1}^{L} f(x_l)w_l$.

No samples are drawn from $p(x)$!

This allow to define recursive sampling procedure, known as **Sequential Importance Filtering** (or Sequential Monte Carlo or Particle Filtering).

We can draw samples from recursive models (e.g. HMC)!
MCMC sampling assumes the number of variables n is known.

Take home messages

- **Sampling is needed** to work with intractable distribution
- Sampling methods can be **unbiased estimators with low variance**
- **In the multivariate case**, it is not easy to derive sampling procedure with these characteristics
  - **Ancestral Sampling** for Bayesian Network
- We can **approximate** the sampling procedure using the **MCMC**
  - **Gibbs Sampling**
  - **Metropolis-Hastings Sampling**
- We can use an **IS** approach to deal with **recursive distribution**
  - **Particle Filtering**

See Chapter 27 of BRML book!

Appendix

In the following slides, we provide:

- some properties of the **expectation** that are useful in the proofs;
- the proof of the **average approximation**;
- the proof of the **variance approximation**;

Expectation properties

The following property are used during the proofs:

- **Linearity**

$$\begin{aligned} E_{p(x)}\left[f(x)+g(x)\right] &= E_{p(x)}\left[f(x)\right]+E_{p(x)}\left[g(x)\right] \\ E_{p(x)}\left[c\,f(x)\right] &= c\,E_{p(x)}\left[f(x)\right]. \end{aligned} \tag{5}$$

- **Expected value of a constant**

$$E_{p(x)}\left[c\right] = c. \tag{6}$$

Also, we use the symbol $\overset{(n)}{=}$ to indicate statement in equation n is used to make a step in the proof.

Proof Average Approximation

We want to prove that

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] = E_{p(x)}\left[f(x)\right]. \tag{2}$$

assuming

$$\tilde{p}(x) = p(x) \tag{7}$$

Proof.

$$E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] \overset{(1)}{=} E_{\tilde{p}(\mathcal{X})}\left[\frac{1}{L}\sum_{l=1}^{L} f(x_l)\right] \overset{(5)}{=} \frac{1}{L}\sum_{l=1}^{L} E_{\tilde{p}(x_l)}\left[f(x_l)\right] =$$

$$\overset{(7)}{=} \frac{1}{L}\sum_{l=1}^{L} E_{p(x)}\left[f(x)\right] = \frac{1}{L} \times L \times E_{p(x)}\left[f(x)\right] = E_{p(x)}\left[f(x)\right].$$

Proof Variance Approximation I

We want to prove that

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)]. \tag{4}$$

assuming

$$\tilde{p}(x) = p(x) \tag{7}$$
$$\tilde{p}(x_i, x_j) = \tilde{p}(x_i)\tilde{p}(x_j) \tag{8}$$

Proof.

This holds due to assumption (7)!

$$\Delta\hat{f}_{\mathcal{X}} \overset{(3)}{=} \hat{f}_{\mathcal{X}} - E_{\tilde{p}(\mathcal{X})}\left[\hat{f}_{\mathcal{X}}\right] \overset{(1)\pm(2)}{=} \frac{1}{L}\sum_{l=1}^{L} f(x_l) - E_{p(x)}\left[f(x)\right] = \tag{9}$$
$$= \frac{1}{L}\sum_{l=1}^{L} f(x_l) - \frac{1}{L}\sum_{l=1}^{L} E_{p(x)}\left[f(x)\right] = \frac{1}{L}\sum_{l=1}^{L}\left(f(x_l) - E_{p(x)}\left[f(x)\right]\right)$$

Proof Variance Approximation II

Then, naming

$$\Delta f(x_l) = f(x_l) - E_{p(x)}\left[f(x)\right], \tag{10}$$

we obtain

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] \overset{(9)}{=} E_{\tilde{p}_{\mathcal{X}}}\left[\left(\frac{1}{L}\sum_{l=1}^{L}\left(f(x_l) - E_{p(x)}\left[f(x)\right]\right)\right)^2\right] \overset{(10)}{=}$$

$$= E_{\tilde{p}_{\mathcal{X}}}\left[\left(\frac{1}{L}\sum_{l=1}^{L}\Delta f(x_l)\right)^2\right] = E_{\tilde{p}_{\mathcal{X}}}\left[\frac{1}{L^2}\sum_{l,l'}^{L}\Delta f(x_l)\Delta f(x_{l'})\right] \overset{(5)}{=} \tag{11}$$

$$= \frac{1}{L}E_{\tilde{p}(x)}\left[(\Delta f(x))^2\right] + \frac{1}{L^2}\sum_{l\neq l'}E_{\tilde{p}(x_l,x_{l'})}\left[\Delta f(x_l)\Delta f(x_{l'})\right].$$

Proof Variance Approximation III

The term

$$\frac{1}{L^2}\sum_{l\neq l'} E_{\tilde{p}(x_l,x_{l'})}\left[\Delta f(x_l)\Delta f(x_{l'})\right] \overset{(8)}{=}$$
$$= \frac{1}{L^2}\sum_{l\neq l'} E_{p(x_l)}\left[\Delta f(x_l)\right] E_{p(x_{l'})}\left[\Delta f(x_{l'})\right] = 0, \tag{12}$$

since

$$E_{p(x_l)}\left[\Delta f(x_l)\right] \overset{(10)}{=} E_{p(x_l)}\left[f(x_l) - E_{p(x)}\left[f(x)\right]\right] \overset{(5)+(6)}{=}$$
$$= E_{p(x)}\left[f(x)\right] - E_{p(x)}\left[f(x)\right] = 0.$$

Finally, combining (11) and (12), we obtain

$$E_{\tilde{p}_{\mathcal{X}}}\left[\left(\Delta\hat{f}_{\mathcal{X}}\right)^2\right] = \frac{1}{L}E_{\tilde{p}(x)}\left[(\Delta f(x))^2\right] = \frac{1}{L}\mathrm{Var}_{p(x)}[f(x)].$$