Sampling methods

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA' DI PISA

DAVIDE.BACCIU@UNIPI.IT

Outline

- Sampling
  - What is it?
  - Why do we need it?
  - Properties of samplers
- Sampling from univariate distributions
- Sampling from multivariate distributions
  - Ancestor sampling
  - Gibbs Sampling
  - (Elements of) Markov Chain Monte Carlo (MCMC)

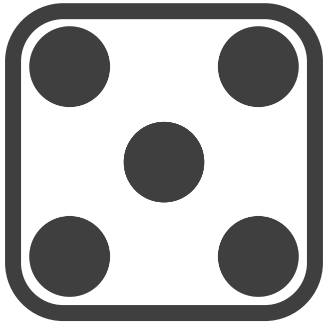
What is sampling?

Sampling consists in drawing a set of realizations $\mathrm{X} = \{x_1, \ldots, x_L\}$ 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



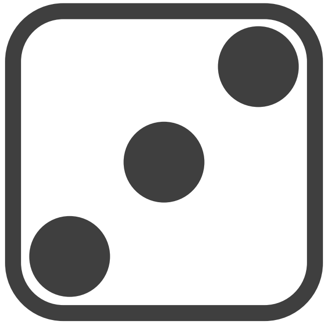
What is sampling?

Sampling consists in drawing a set of realizations $\mathrm{X} = \{x_1, \ldots, x_L\}$ 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

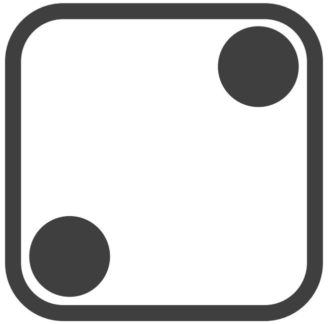


What is sampling?

Sampling consists in drawing a set of realizations $\mathrm{X} = \{x_1, \ldots, x_L\}$ 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



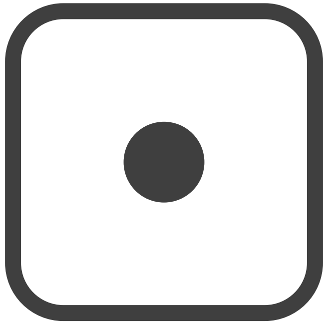
What is sampling?

Sampling consists in drawing a set of realizations $\mathrm{X} = \{x_1, \ldots, x_L\}$ 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



UNIVERSITÀ DI PISA

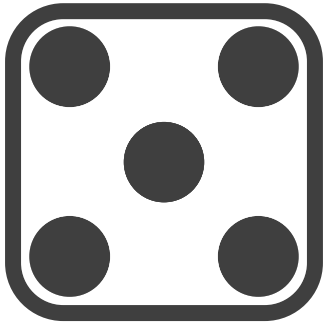
What is sampling?

Sampling consists in drawing a set of realizations $\mathrm{X} = \{x_1, \ldots, x_L\}$ 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** $X = \{5, 3, 2, 1, 5\}$ contains $L = 5$ **samples**.

Why do we need sampling?
Approximating expectations

Suppose that we want to compute the expectation $E_{p(x)}[f(x)]$.

If $p(x)$ is intractable, we cannot compute it enumerating all the states of x.

If we have a sample set $X = \{x_1, \ldots, x_L\}$, then we can approximate the expectation as:

$$E_{p(x)}[f(x)] \approx \frac{1}{L}\sum_{l=1}^{L} f(x^l) \equiv \hat{f}_X \quad (1)$$

There are many cases where $p(x)$ is intractable:

- the distribution of a Boltzmann Machine;
- the posterior $P(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$ in LDA;
- posteriors when non-conjugate priors are used.

Why do we need sampling?
Learning parameters

In Bayesian models, parameters are random variables.

We can **learn** the model parameters by **sampling** their **posteriors**!

In LDA, we can learn the model parameters by sampling:

$$\theta, \boldsymbol{z}, \beta \sim P(\theta, \boldsymbol{z}, \beta | \boldsymbol{w}, \alpha)$$

Sampling from the posterior is also useful to **classify** new instances!

In this case, we sample:

$$\theta^*, \boldsymbol{z}^* \sim P(\theta, \boldsymbol{z} | \boldsymbol{w}^*, \alpha, \beta),$$

where $\boldsymbol{w}^*$ are the words in the new documents.

Properties of sampling

The most important properties of sampling are:

- the empirical distribution converges almost surely to the true distribution:

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

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

- the sampling approximation $\hat{f}_X$ of the expectation can be an unbiased estimator;
- the sampling approximation $\hat{f}_X$ of the expectation can have low variance;

The last two properties are **desirable but difficult** to ensure!

Unbiased Sampling Approximation

Quick refresher:
Unbiased estimator $\hat{\theta}$ of the unknown θ. $\Longrightarrow$ the approximation is exact on average.

Let $\tilde{p}(X)$ the distribution over all possible realizations of the sampling set X, then $\hat{f}_X$ is an unbiased estimator if:

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

This is **true** provided that $\tilde{p}\left(x^l\right) = p(x)$!

The equality ensure us that we are **sampling** the **desired distribution**, i.e. we are using a **valid sampler**!

Variance of Sampling Approximation
Definition

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

Let

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

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

$$E_{\tilde{p}_X}\left[\left(\Delta \hat{f}_X\right)^2\right]$$

If the variance is **low**, $\hat{f}(X)$ is (quite) always **close** to its expected value, i.e $E_{p(x)}[f(x)]$!

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

$$E_{\tilde{p}_X}\left[\left(\Delta \hat{f}_X\right)^2\right] = \frac{1}{L} Var_{p(x)}[f(x)]. \tag{4}$$

We can use a **small** number of samples to accurately estimate the expectation!

Provided that $Var_{p(x)}[f(x)]$ is finite.

Sampling procedures as distributions

The quality of the sampling approximation depends on the properties of $\tilde{p}(X)$, i.e. the probability to obtain a sample set X.

$$\boldsymbol{p}(X) \neq \boldsymbol{p}(\boldsymbol{x})$$

$p(x) \longrightarrow$ The distribution to sample $\longrightarrow$ Does not depend on the sampling procedure.

$\tilde{p}(X) \longrightarrow$ The distribution of the samples $\longrightarrow$ Depends on the sampling procedure.

Small recap

So far, we have shown that:

- we need sampling:
  - to approximate expectations;
  - to do inference in Bayesian models.
- properties of the sampling procedure depends on $\tilde{p}(X)$:
  - $\tilde{p}(x^l) = p(x^l) \Longrightarrow$ valid sampler;
  - $\tilde{p}(x^l, x^{l'}) = \tilde{p}(x^l)\tilde{p}(x^{l'}) \Longrightarrow$ low approximation variance.

In the next slides, we introduce examples of sampling procedures:

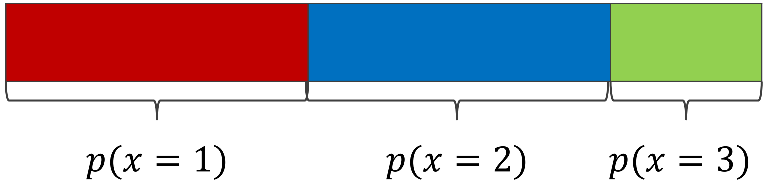
- sampling from univariate distributions;
- sampling from multivariate distributions:
  - naive approaches;
  - exact procedures;
  - approximated procedures.

Univariate Sampling

Drawing samples from a univariate distribution 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 with C states.

Hence, each sample x^l contains n values.

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

How can we sample from $p(x)$?

Naive Multivariate Sampling - 1

We build a 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	$p(1,1,1,1,2)$
3	1	1	1	1	3	$p(1,1,1,1,3)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
C^n	C	C	C	C	C	$p(C,C,C,C,C)$

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

S has $O(C^n)$ states! **Computationally infeasible**!

Naive Multivariate Sampling - 2

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

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

Them, we sample the variables in the following order:

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

⋮

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

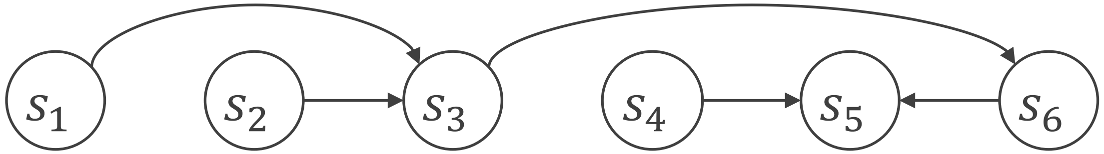
Easy because univariate!

Unfortunately, computing the distribution $p(s_i|s_{j<i})$ easily becomes **exponential w.r.t. 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 already represented as a Belief Network (BN), we can apply it directly!



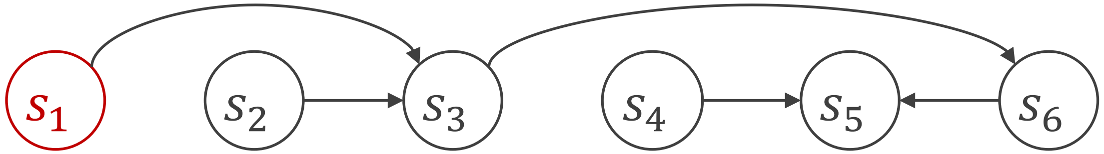
The BN ancestral order tell us the sampling order.

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

Ancestral Sampling
Example

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

Sample $\tilde{s}_1 \sim p(s_1)$



Ancestral Sampling
Example

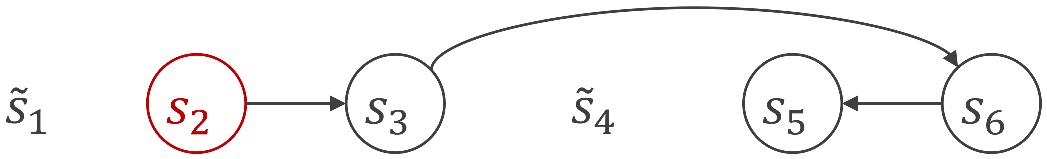
$$\{\cancel{s_1}, s_2, \color{red}{s_4}\} \prec \{s_3\} \prec \{s_6\} \prec \{s_5\}$$

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

Ancestral Sampling
Example

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

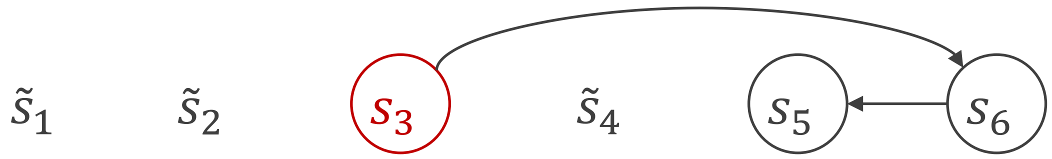
Sample $\tilde{s}_2 \sim p(s_2)$



Ancestral Sampling
Example

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

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



Ancestral Sampling

Example

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

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

Ancestral Sampling
Example

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

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

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

Ancestral Sampling

Example

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

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

This is a single sample x^l!

AS performs **exact sampling** since each sample x^l is drawn from $p(x)$.

The samples are also **independent**!
Thus, AS has **low variance**!

Sampling with evidence

Suppose that a subset of variables $s_{\in}$ are visible; writing $s = s_{\in} \cup s_{\setminus\in}$, we would like to sample from:

$$p(s_{\setminus\in}|s_{\in}) = \frac{p(s_{\setminus\in}, s_{\in})}{p(s_{\in})}$$

Can we still use AS?

- Clamping variables changes the structure of the BN.
  In previous example $s_1 \perp\!\!\!\perp s_2$, but $s_1 \not\perp\!\!\!\perp s_2 | s_3$.

Computing the new structure is complex as running exact inference!

- We can run AS on the old structure and then discard any samples which do not match the evidence.

We discard **a lot** of samples!

The needing of new sampling procedures

Sampling under evidence is important!

In probabilistic models, the inference is based on the posterior:

$$p(h|v) = \frac{p(h,v)}{p(v)},$$

where:

- where h is set of hidden variables
- where v is set of visible variables (i.e. the data)

We need an **efficient** method to sample under evidence!

In the next slide, we introduce the Gibbs sampling procedure.

Gibbs Sampling

Example

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
Example

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

UNIVERSITÀ DI PISA

Gibbs Sampling
Example

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
Example

The idea is to start from a sample $x_1 = \{s_1^1, \dots, s_n^1\}$ 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
Example

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
Example

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.

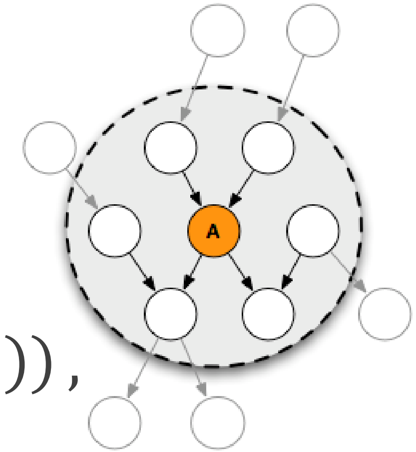
Sample	s_1	s_2	s_3	s_4	s_5
x^1	1	1	2	4	5
x^2	3	1	2	4	5
x^3	3	4	2	4	5
x^4	3	4	2	1	5
x^5	3	4	6	1	5
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Gibbs Sampling
Definition

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

- we select a variable s_j;
- we sample its value according to

$$s_j^{l+1} \sim p(s_j | s_{\setminus j}) = \frac{1}{Z} p(s_j | pa(s_j)) \prod_{k \in ch(j)} p(s_k | pa(s_k)),$$

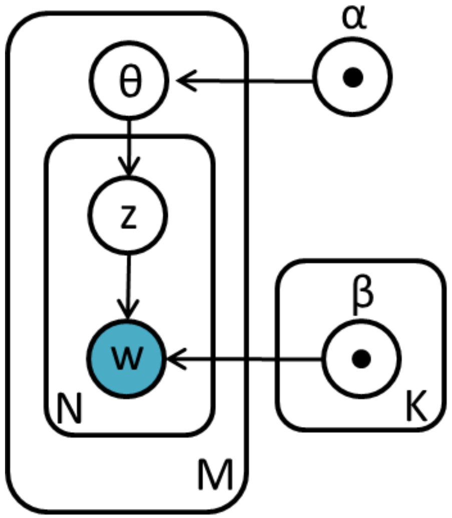


where variables $s_{\setminus j}$ are clamped to $\{s_1^l, \ldots, s_{j-1}^l, s_{j+1}^l, \ldots, s_n^l\}$.

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!

LDA Gibbs Sampling



Start from an initial guess $\{z_{ij}^0, \theta_i^0, \beta^0\}$.

Do:

1. $z_{ij}^{l+1} \sim P(z_{ij}|\mathbf{w}, \mathbf{z_{ij}^{l-1}}, \theta^l, \beta^l, \alpha)$
2. $\theta_i^{l+1} \sim P(\theta_i|\mathbf{w}, \mathbf{z}^{l+1}, \beta^l, \alpha)$
3. $\beta^{l+1} \sim P(\beta|\mathbf{w}, \mathbf{z}^{l+1}, \theta^{l+1}, \alpha)$

Repeat until convergence.

The **derivation** of the sampling formulas can be quite **mathematically involved**

The convergence criteria is based on $P(\mathbf{z}, \mathbf{w}, \theta|\beta, \alpha)$.
The procedure terminates when the likelihood stops increasing.

Gibbs Sampling
Properties

The Gibbs sampling draws a new sample x^l from $q(x^l | x^{l-1})$.

- **Is the Gibbs sampling a valid sampling procedure?**

We are **not sampling** from p(x)!

We cannot ensure that the sampling distribution has the same marginals of $p(x)$.
However, if we compute the limit to $L \rightarrow \infty$, the series $\{x_1, \dots, x_L\}$ converges to samples taken from $p(x)$!

In the limit of **infinite samples, the Gibbs sampler is valid!**

- **Has the Gibbs sampler low variance?**

No, samples are **highly dependent**!

MCMC Sampling Framework

Gibbs sampling is a specialization of the Markov Chain Monte Carlo (MCMC) sampling framework.

The idea is to build a **Markov Chain** whose stationary distribution is $p(x)$.

Let $q(x^{l+1}|x^{l})$ the MC state-transition distribution, we must ensure that the Markov Chain is:

- irreducible → it is possible to reach any state from anywhere;
- aperiodic → at each time-step, we can be anywhere.

Hence, the Markov Chain has a unique stationary distribution.

There are different $q(\cdot)$ which converge to $p(\cdot)$.

UNIVERSITÀ DI PISA

MCMC Sampling Procedures

There are many sampling procedures in the MCMC framework, defining different state-transition distribution $q(\cdot)$:

- Gibbs Sampling
  - $q(\cdot)$ relies on marginals $p(s_j|s_{\setminus j})$!
  - it works well when variables are not strongly related!
- Metropolis-Hastings Sampling
  - based on a proposal distribution $\tilde{q}(x^{l+1}|x^l)$!
  - the choice of $\tilde{q}(\cdot)$ is crucial!
- Particle Filtering
  - it is used in recursive models such as HMM!
- Hybrid Monte Carlo
- Swendson-Wang

  ⋮

Each of them has different characteristics!

We should choose the most suitable for our purpose!

Take home messages

- Sampling is useful to deal with intractable $p(x)$:
  - we can approximate expectations;
  - we can perform inference in Bayesian models;
- $p(x)$ univariate $\rightarrow$ sampling is easy!
- $p(x)$ multivariate $\rightarrow$ sampling is difficult!
  - naive approaches are not feasible;
  - if $p(x)$ is a BN, we can use AS (valid and with low variance);
  - AS does not work with evidence (we always have it!);
- MCMC framework approximates the sampling procedure:
  - we can easily deal with evidence;
  - the sampler defines a MC whose stationary distribution is $p(x)$;
    - the sampler is valid in the limit $l \rightarrow \infty$;
  - different state-transition q leads to different procedures:
    - Gibbs Sampling, Metropolis-Hastings Sampling, . . .

Take home messages

- Sampling is useful to deal with intractable $p(x)$:
  - we can approximate expectations;
  - we can perform inference in Bayesian models;
- $p(x)$ univariate $\rightarrow$ sampling is easy!
- $p(x)$ multivariate $\rightarrow$ sampling is difficult!
  - naive approaches are not feasible;
  - if $p(x)$ is a BN, we can use AS (valid and with low variance);
  - AS does not work with evidence (we always have it!);
- Gibbs sampling approximates the sampling procedure:
  - we can easily deal with evidence;
  - the sampler defines a Markov Chain whose stationary distribution is $p(x)$;
    - the sampler is valid in the limit $l \rightarrow \infty$;

Next Lecture (April 02)

There is no lecture on April 01!
(Will be recovered on April 11 - h. 14.00)

Boltzmann machines

- The missing link between probabilistic (MRF) and neural models (RNNs)
- The originator of the whole deep learning fuzz
- A good way to see Gibbs sampling at work