Probabilistic Graphical Models and Their Applications

Bjoern Andres and Bernt Schiele

Max Planck Institute for Informatics

slides adapted from Peter Gehler

January 4, 2o17

Today's topics

- Sampling
  - Barber Sections 27.1, 27.2, 27.3, 27.4

What to infer?

- Mean

$$\mathbb{E}_{p(x)}[x] = \sum_{x \in \mathcal{X}} x p(x)$$

- Mode (most likely state)

$$x^* = \operatorname*{argmax}_{x \in \mathcal{X}} p(x)$$

- Conditional Distributions

$$p(x_i, x_j \mid x_k, x_l) \quad \text{or} \quad p(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$$

- Max-Marginals

$$x_i^* = \operatorname*{argmax}_{x_i \in \mathcal{X}_i} p(x_i) = \operatorname*{argmax}_{x_i \in \mathcal{X}_i} \sum_{j \neq i} p(x_1, \ldots, x_n)$$

Inference in General Graphs – Approximate Inference

Approximate Inference?

- Approximate Inference comes into play whenever exact inference is not tractable.
  - E.g. the model is not tree structured
- What would we like to approximate?
  - E.g. posterior distribution $p(z \mid x)$
  - Expectations
    - continuous: integrals may be intractable
    - discrete: sum over exponentially many states $\Rightarrow$ infeasible
- Conceptually there are two approaches
  - Deterministic Approximation
  - Numerical Sampling (e.g. Markov Chain Monte Carlo)

Two approaches

1. Deterministic Approximation
   - Approximate the quantity of interest
   - Solve the approximation analytically
   - Results depends on the quality of the approximation
2. Numerical Sampling
   - Take the quantity of interest
   - Use random samples to approximate it
   - Results depends on the quality and amount of random samples

- The correct answer to the wrong question, or the wrong answer to the correct question?
- Only sampling allows to get the *golden standard*

Different methods

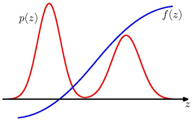
- For trees: one algorithm only (efficient)
- In general graphs: difficult, therefore many algorithms have been proposed
- Sampling
  - Markov Chain Monte Carlo
  - Gibbs Sampling
  - ...
- Deterministic Approximate Inference
  - Variational Bounds
  - Loopy Belief Propagation
  - Mean field
  - Junction Tree
  - Expectation Propagation

Approximate Inference: Sampling

Motivation: Sampling

- Draw random samples from some distribution $p(x)$
  - discrete or continuous
  - univariate or multi-variate
- For example Gaussian, Poisson, Uniform, Dirichlet, ...
  - All of the above already available in Matlab
- More general: what about sampling from some joint distribution $p(x)$ e.g. defined by a graphical model?
  - e.g. a distribution over body parts, we want to find likely body poses
  - e.g. a distribution over images, we want to look at likely images.

Example: Expectation



- We want to evaluate
$$\mathbb{E}[f] = \int f(x)p(x)dx \quad \text{or } \mathbb{E}[f] = \sum_{x \in \mathcal{X}} f(x)p(x)$$
- Sampling idea:
  - draw L independent samples $x^1, x^2, \ldots, x^L$ from $p(\cdot)$: $x^l \sim p(\cdot)$
  - replace the integral/sum with the finite set of samples
$$\hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(x^l)$$
  - as long as $x^l \sim p(\cdot)$ then
$$\mathbb{E}[\hat{f}] = \mathbb{E}[f]$$

So how to sample? A Simple case

Just to get an idea of what's going on

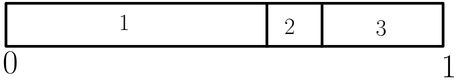
Pre-Requiste

- Assume we can draw a value uniformly at random from the unit interval $[0, 1]$
- How? Pseudo-Random number generators

Univariate Sampling – discrete example

- Target distribution with $K = 3$ states

$$p(x) = \begin{cases} 0.6 & x = 1 \\ 0.1 & x = 2 \\ 0.3 & x = 3 \end{cases} \tag{1}$$



Univariate Sampling – discrete

Slightly more formal:

- Consider we want to sample from a univariate discrete distribution p
  - one-dimensional
  - K states
- So we have $p(x = k) = p_k$
- Calculate the cumulant

$$c_i = \sum_{j \leq i} p_j \tag{2}$$

- Draw $u \sim [0, 1]$
- Find that i for which $c_{i-1} < u \leq c_i$
- Return state i as sample from p

Univariate Sampling – continuous

Extension to continuous variable is clear

- Compute the cumulant

$$C(y) = \int_{-\infty}^{y} p(x)dx \tag{3}$$

- Then sample $u \sim [0, 1]$
- Compute $x = C^{-1}(u)$
- So sampling is possible if we can compute the integral
  - e.g. Gaussian distribution

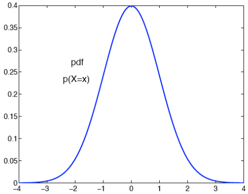
Univariate Sampling Example: Gaussian

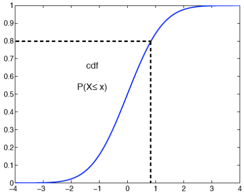
- 1-dimensional Gaussian pdf (probability density function) $p(x|\mu, \sigma^2)$ and the corresponding cumulative distribution:

$$F_{\mu,\sigma^2}(x) = \int_{-\infty}^{x} p(z|\mu, \sigma^2)dz$$

- to draw a sample from a Gaussian, we invert the cumulative distribution function

$$u \sim uniform(0,1) \Rightarrow x = F_{\mu,\sigma^2}^{-1}(u) \sim p(x|\mu, \sigma^2)$$





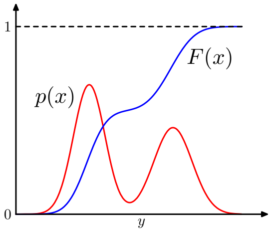
Univariate Sampling

- assume pdf (probability density function) $p(x)$ and the corresponding cumulative distribution:

$$F(x) = \int_{-\infty}^{x} p(z)dz$$

- to draw a sample from this pdf, we invert the cumulative distribution function

$$u \sim uniform(0,1) \Rightarrow x = F^{-1}(u) \sim p(x)$$



Overview: Sampling Methods

- Rejection Sampling
- Ancestral Sampling
- Importance Sampling
- Gibbs Sampling
- Markov Chain Monte Carlo methods
- Metropolis-Hastings
- Hybrid Monte Carlo

- Do I need to know them all?
- Yes! Sampling is an "art", most efficient technique depends on model structure

Rejection Sampling

Rejection Sampling

- Suppose we want to sample from $p(x)$ (but that is difficult)
- Furthermore assume we can evaluate $p(x)$ up to a constant (think of Markov Networks)

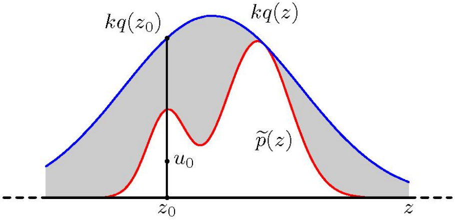
$$p(x) = \frac{1}{Z}\tilde{p}(x) = \frac{1}{Z}\prod_c \phi_c(\mathcal{X}_c) \tag{4}$$

- Instead sample from a proposal distribution $q(x)$
- Choose q such that we can easily sample and a k exists such that

$$kq(x) \geq \tilde{p}(x)\ \forall x \tag{5}$$

Rejection Sampling

- Sample two random variables:
  1. $z_0 \sim q(x)$
  2. $u_0 \sim [0, kq(z_0)]$ uniform
- reject sample z_0 if $u_0 > \tilde{p}(z_0)$

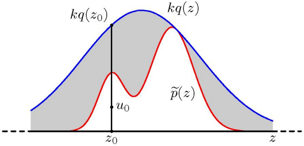


Probability of acceptance

- Sample z drawn from q and accepted with probability $\tilde{p}(z)/kq(z)$
- So (overall) acceptance probability

$$p(accept) = \int \frac{\tilde{p}(z)}{kq(z)} q(z) dz = \frac{1}{k} \int \tilde{p}(z) dz \tag{6}$$

- So the lower k the better (more acceptance)
  - subject to constraint $kq(z) \geq \tilde{p}(z)$



Efficiency of Rejection Sampling: Example

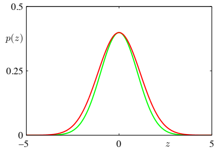
- Depends on k
- If $q(x) = p(x)$ and $k = 1$ then $p(accept) = 1$
- But $k > 1$ is typical
- For the easiest case of factorizing distribution $p(x) = \prod_{i=1}^{D} p(x_i)$ we have

$$p(accept \mid x) = \prod_{i=1}^{D} p(accept \mid x_i) = \mathcal{O}(\gamma^D) \tag{7}$$

  where $0 \leq \gamma \leq 1$ typical value for $p(accept \mid x_i)$
- Thus rejection sampling is usually impractical in high dimensions

Efficiency of Rejection Sampling

- Example:
  - assume $p(x)$ is Gaussian with covariance matrix: $\sigma_p^2 I$
  - assume $q(x)$ is Gaussian with covariance matrix: $\sigma_q^2 I$
  - clearly: $\sigma_q^2 \geq \sigma_p^2$
  - in D dimensions: $k = \left(\frac{\sigma_q}{\sigma_p}\right)^D$



- assume:
  - σ_q is 1% larger than σ_p, $D = 1000$
  - then $k = 1.01^{1000} \geq 20000$
  - and $p(accept) \leq \frac{1}{20000}$
- therefore: often impractical to find good proposal distribution $q(x)$ for high dimensions

Multivariate Sampling

- Multivariate: more than one dimension
- Idea: translate multivariate case into a univariate case:
- Enumerate all joint states $(x_1, x_2, \ldots, x_n)$ (assume discrete), i.e. give them each a unique i from 1 to the total (exponential) number of states
- Now we have to sample from univariate distributions again
- Problem: Exponential growth of states (with n)

Multivariate Sampling

- Another idea, use Bayes rule

$$p(x_1, x_2) = p(x_2 \mid x_1)p(x_1) \tag{8}$$

- Now first sample x_1, then x_2 both of which are univariate
- Now we have a one dimensional distribution again
- Problem: Need to know the conditional distributions

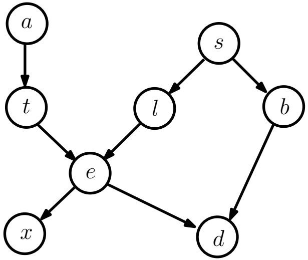
Ancestral Sampling

Ancestral Sampling

- For Belief Networks (remember) $p(x) = \prod_i p(x_i \mid \mathsf{pa}(x_i))$
- So the sampling algorithm should be clear

$$p(a,t,e,x,l,s,b,d) = p(a)p(s)p(t|a)p(l|s)p(b|s)p(e|t,l)p(x|e)p(d|e,b)$$

- Forward sampling: from parents to children
- sampling from each distribution $(p(a), p(t \mid a), \ldots)$
  may be (in itself / as a subproblem) difficult



Perfect Sampling

- Each instance drawn using forward sampling is independent!
- This is called perfect sampling
- In contrast to MCMC methods, where samples are dependent
- Remark: there is also a perfect sampling technique for MCMC, but that is applicable only to some special cases [Propp&Wilson, 1996]

Problem of Ancestral Sampling

- Problem: Evidence!
  - when a subset of the variables is observed
- Example, we have the following distribution
$$p(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3 \mid x_1, x_2)$$
- and have observed x_3.
- We want to sample from

$$p(x_1, x_2, \mid x_3) = \frac{p(x_1)p(x_2)p(x_3 \mid x_1, x_2)}{\sum_{x_1,x_2} p(x_1)p(x_2)p(x_3 \mid x_1, x_2)} \tag{9}$$

- Observing x_3 makes x_1, x_2 dependent
- Sample and discard inconsistent ones (in-efficient)

Importance Sampling

Importance Sampling

Approach:

- approximate expectation directly
  (but does not enable to draw samples from $p(z)$ directly)
- setting: $p(z)$ can be evaluated (up to a normalization constant)
- goal:

$$\mathbb{E}[f] = \int f(z)p(z)dz$$

Naïve method: grid-sampling

- discretize z-space into a uniform grid
- evaluate the integrand as a sum of the form:

$$\mathbb{E}[f] \simeq \sum_{l=1}^{L} f(z^l)p(z^l)$$

- but: number of terms grows exponentially with number of dimensions

Importance Sampling

Idea:

- use a proposal distribution $q(z)$ from which it is easy to draw samples
- express expectation in the form of a finite sum over samples $\{z^l\}$ drawn from $q(z)$:

$$\begin{aligned}\mathbb{E}[f] &= \int f(z)p(z)dz = \int f(z)\frac{p(z)}{q(z)}q(z)dz \\ &\simeq \frac{1}{L}\sum_{l=1}^{L}\frac{p(z^l)}{q(z^l)}f(z^l)\end{aligned}$$

- with importance weights: $r^l = \frac{p(z^l)}{q(z^l)}$

Importance Sampling

Typical setting:

- $p(z)$ can be only evaluated up to a normalization constant (unkown):
$$p(z) = \tilde{p}(z)/Z_p$$
- $q(z)$ can be also treated in a similar way:
$$q(z) = \tilde{q}(z)/Z_q$$
- then:

$$\begin{aligned}
\mathbb{E}[f] &= \int f(z)p(z)dz = \frac{Z_q}{Z_p}\int f(z)\frac{\tilde{p}(z)}{\tilde{q}(z)}q(z)dz \\
&\simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L}\tilde{r}^l f(z^l)
\end{aligned}$$

- with: $\tilde{r}^l = \frac{\tilde{p}(z^l)}{\tilde{q}(z^l)}$

Importance Sampling

Ratio of normalization constants can be evaluated :

$$\frac{Z_p}{Z_q} = \frac{1}{Z_q}\int \tilde{p}(z)dz = \int \frac{\tilde{p}(z)}{\tilde{q}(z)} q(z)dz \simeq \frac{1}{L}\sum_{l=1}^{L} \tilde{r}^l$$

- and therefore:

$$\mathbb{E}[f] \simeq \frac{Z_q}{Z_p}\frac{1}{L}\sum_{l=1}^{L} \tilde{r}^l f(z^l) = \sum_{l=1}^{L} w^l f(z^l)$$

- with:

$$w^l = \frac{\tilde{r}^l}{\sum_m \tilde{r}^m} = \frac{\frac{\tilde{p}(z^l)}{\tilde{q}(z^l)}}{\sum_m \frac{\tilde{p}(z^m)}{\tilde{q}(z^m)}}$$

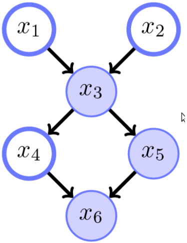
Importance Sampling

Observations:

- success of importance sampling depends crucially on how well the sampling distribution $q(z)$ matches the desired distribution $p(z)$
- often, $p(z)f(z)$ is strongly varying and has significant proportion of its mass concentrated over small regions of z-space
- as a result weights $\tilde{r}^l$ may be dominated by a few weights having large values
- practical issues: if none of the samples falls in the regions where $p(z)f(z)$ are large ...
  - the results may be arbitrarily wrong
  - and no diagnostic indication !
    (because there is no large variance in $\tilde{r}^l$ then)

Gibbs Sampling

Gibbs Sampling



- Sample from this distribution $p(x)$
- Idea: Sample sequence $x^0, x^1, x^2, \ldots$ by updating one variable at a time
- Eg. update x_4 by conditioning on the set of shaded variables Markov blanket

$$p(x_4 \mid x_1, x_2, x_3, x_5, x_6) = p(x_4 \mid x_3, x_5, x_6)$$

Gibbs Sampling: General Recipe

- Update x_i

$$p(x_i \mid x_{\setminus i}) = \frac{1}{Z} p(x_i \mid pa(x_i)) \prod_{j \in \mathsf{ch}(i)} p(x_j \mid \mathsf{pa}(x_j)) \tag{10}$$

- and the normalisation constant is

$$Z = \sum_{x_i} p(x_i \mid pa(x_i)) \prod_{j \in \mathsf{ch}(i)} p(x_j \mid \mathsf{pa}(x_j)) \tag{11}$$

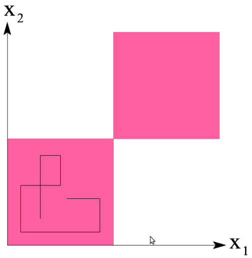
Gibbs Sampling: Remarks

- Think of Gibbs sampling as

$$x^{l+1} \sim q(\cdot \mid x^l) \tag{12}$$

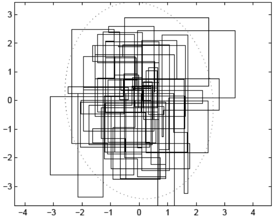
- Problem: States are highly dependent $(x^1, x^2, \ldots)$
- Need a long time to run Gibbs sampling to *forget* the initial state, this is called burn in phase
- Dealing with evidence is easy: simply clamp the variables to the values.
- Widely adopted technique for approximate inference (BUGS package `www.mrc-bsu.cam.ac.uk/bugs`)

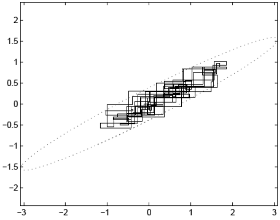
Gibbs Sampling: Remarks



- In this example the samples stay in the lower left quadrant
- Some technical requirements to Gibbs sampling
- The Markov Chain $q(x^{l+1} \mid x^{l})$ needs to be able to traverse the entire state-space (no matter where we start)
  - This property is called irreducible
  - Then $p(x)$ is the stationary distribution of $q(x' \mid x)$

Gibbs Sampling: Remarks


(a)


(b)

- Gibbs sampling is more efficient if states are not correlated
  - Left: Almost isotropic Gaussian
  - Right: correlated Gaussian
- The Markov chain has a higher mixing coefficient
  - i.e. it converges faster to the stationary distribution

Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC)

- Sample from a multi-variate distribution

$$p(x) = \frac{1}{Z} p^*(x) \tag{13}$$

  with Z intractable to calculate

- Idea: Sample from some $q(x^{l+1} \mid x^l)$ with a stationary distribution

$$q_\infty(x') = \int_x q(x' \mid x) q_\infty(x) \tag{14}$$

- Given $p(x)$ find $q(x' \mid x)$ such that $q_\infty(x) = p(x)$
- Gibbs sampling is one instance (that is why it is working)

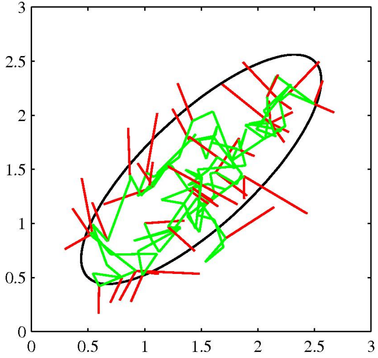
Metropolis sampling

- Special case of MCMC method (proposal distribution) with the following proposal distribution
  - symmetric: $q(x' \mid x) = q(x \mid x')$
- Sample x' and accept with probability

$$A(x', x) = \min\left(1, \frac{p^*(x')}{p^*(x)}\right) \in [0, 1] \tag{15}$$

  - If new state x' is more probable always accept
  - If new state is less probable accept with $\frac{p^*(x')}{p^*(x)}$

Example: 2D Gaussian



- 150 proposal steps, 43 are rejected (red)

Metropolis-Hastings sampling (1953)

- Slightly more general MCMC method when the proposal distribution is *not* symmetric
- Sample x' and accept with probability

$$A(x', x) = \min\left(1, \frac{\tilde{q}(x \mid x')p^*(x')}{\tilde{q}(x' \mid x)p^*(x)}\right) \tag{16}$$

- Note: when the proposal distribution is symmetric, Metropolis-Hastings reduces to standard Metropolis sampling

Is this sampling from the correct distribution?

- In the following we show that Metropolis-Hastings samples from the desired distribution $p(x)$
- Consider the following transition

$$q(x' \mid x) = \tilde{q}(x' \mid x) f(x', x) + \delta(x, x') \left(1 - \int_{x''} \tilde{q}(x'' \mid x) f(x'', x)\right)$$

  with proposal distribution $\tilde{q}$
- This is a distribution

$$\int_{x'} q(x' \mid x) = \int_{x'} \tilde{q}(x' \mid x) f(x', x) + \left(1 - \int_{x''} \tilde{q}(x'' \mid x) f(x'', x)\right) = 1$$

- Now find $f(x', x)$ such that stationary distribution is $p(x)$.

Continuing...

- We want $f(x', x)$ such that
$$p(x') = \int_x q(x' \mid x) p(x)$$
- using:
$$q(x' \mid x) = \tilde{q}(x' \mid x) f(x', x) + \delta(x, x') \left(1 - \int_{x''} \tilde{q}(x'' \mid x) f(x'', x)\right)$$
- we get:
$$\begin{aligned} p(x') &= \int_x \tilde{q}(x' \mid x) f(x', x) p(x) \\ &\quad + p(x') \left(1 - \int_{x''} \tilde{q}(x'' \mid x') f(x'', x')\right) \end{aligned}$$
- In order for this to hold we need to require
$$\int_x \tilde{q}(x' \mid x) f(x', x) p(x) = \int_{x''} \tilde{q}(x'' \mid x') f(x'', x') p(x')$$

Continuing...

- This holds for the Metropolis-Hastings acceptance rule

$$
\begin{aligned}
A(x', x) = f(x', x) &= \min\left(1, \frac{\tilde{q}(x \mid x')p^*(x')}{\tilde{q}(x' \mid x)p^*(x)}\right) \\
&= \min\left(1, \frac{\tilde{q}(x \mid x')p(x')}{\tilde{q}(x' \mid x)p(x)}\right)
\end{aligned}
$$

- we need to require (from previous slide):

$$\int_x \tilde{q}(x' \mid x) f(x', x) p(x) = \int_{x''} \tilde{q}(x'' \mid x') f(x'', x') p(x')$$

- which is satisfied because of the (detailed balance) property:

$$
\begin{aligned}
f(x', x)\tilde{q}(x' \mid x)p(x) &= \min(\tilde{q}(x' \mid x)p(x), \tilde{q}(x \mid x')p(x')) \\
&= \min(\tilde{q}(x \mid x')p(x'), \tilde{q}(x' \mid x)p(x)) \\
&= f(x, x')\tilde{q}(x \mid x')p(x')
\end{aligned}
$$

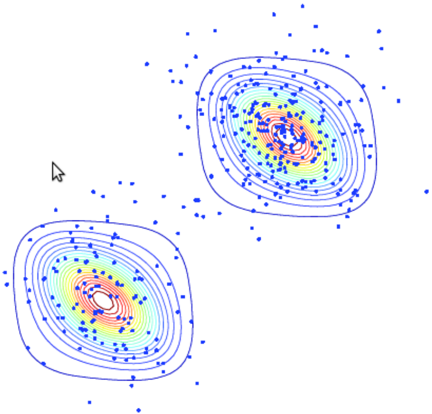
Continuing...

- A common proposal distribution is given by

$$\tilde{q}(x' \mid x) = \mathcal{N}(x' \mid x, \sigma^2 I)$$

- which is symmetric $\tilde{q}(x' \mid x) = \tilde{q}(x \mid x')$

Example: multi-modal distribution



- $\tilde{q}$ needs to bridge the gap (be irreducible)

Sampling

- Much much more to learn about sampling
- Widely used: Gibbs Sampling, Metropolis Hastings
- Usually requires experience and careful adpation to your specific problem