Probabilistic Machine Learning

Monte Carlo inference

Alexandre Graell i Amat
alexandre.graell@chalmers.se
https://sites.google.com/site/agraellamat

November 28 and 30, 2023

CHALMERS

Bayesian (probablistic) machine learning

In this course we consider problems of the form:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- $\mathcal{D}$: Observed data
- $\boldsymbol{\theta}$: parameters of some model explaining the data

Goal: Find $p(\boldsymbol{\theta}|\mathcal{D})$.

- Can be found exactly in some cases (conjugate priors)
- Computation complexity can be alleviated when $p(\mathcal{D}, \boldsymbol{\theta})$ defined by specific classes of probabilistic graphical models (BNs, MRFs, FGs)

And when computing $p(\boldsymbol{\theta}|\mathcal{D})$ is intractable?

A simple(?) example

$$x_1 \rightarrow x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \rightarrow x_6$$

$$\begin{aligned}
p(x_1) &= \mathcal{U}(x_1; [a_1, b_1]) \\
p(x_2|x_1) &= \mathcal{N}(x_1; x_1, \sigma_2^2) \\
p(x_3|x_2) &= \mathcal{N}(x_3; [x_2, \sigma_3^2]) \\
p(x_4|x_3) &= \mathcal{U}(x_4; [x_3 - a_4, x_3 + a_4]) \\
p(x_5|x_4) &= \mathcal{U}(x_5; [x_4 - a_5, x_4 + a_5]) \\
p(x_6|x_5) &= \mathcal{N}(x_6; x_5, \sigma_6^2)
\end{aligned}$$

$$\begin{aligned}
p(x_1|x_6) &= \frac{p(x_1, x_6)}{p(x_6)} \\
&= \int \cdots \int \frac{p(x_1, x_2, x_3, x_4, x_5, x_6)}{p(x_6)} \mathsf{d}x_2 \mathsf{d}x_3 \mathsf{d}x_4 \mathsf{d}x_5 \\
&= \int \cdots \int \frac{p(x_1)p(x_2|x_1)p(x_3|x_2)p(x_4|x_3)p(x_5|x_4)p(x_6|x_5)}{p(x_6)} \mathsf{d}x_2 \mathsf{d}x_3 \mathsf{d}x_4 \mathsf{d}x_5
\end{aligned}$$

Approximate inference

Need to resort to approximations:

Stochastic methods:

- Monte Carlo approximation (numerical sampling)

Deterministic approximate inference methods:

- Variational inference
- Expectation propagation

Monte Carlo inference

Idea: Generate samples $\boldsymbol{\theta}^{(\tau)}$ from posterior, $\boldsymbol{\theta}^{(\tau)} \sim p(\boldsymbol{\theta}|\mathcal{D})$, and use them to compute any quantity of interest, e.g., $p(\theta_1|\mathcal{D})$.

- Can achieve any desired level of accuracy by generating enough samples

Main issue: How do we efficiently generate samples from a probability distribution, particularly in high dimensions?

We will use Bishop's notation:

$p(\boldsymbol{z})$: probability density

(in the learning case, $\boldsymbol{z} = \boldsymbol{\theta}$ and $p(\boldsymbol{z}) = p(\boldsymbol{\theta}|\mathcal{D})$)

We will focus on evaluating expectations

Monte Carlo inference

Why expectations?

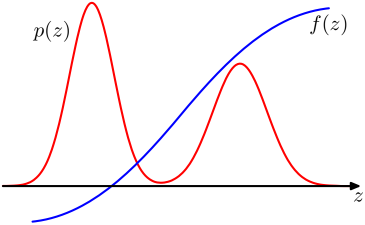
Example: Making predictions

$$
\begin{aligned}
p(t|\mathcal{D}) &= \int p(t|\boldsymbol{\theta}, \mathcal{D}) p(\boldsymbol{\theta}|\mathcal{D}) \mathrm{d}\boldsymbol{\theta} \\
&= \mathbb{E}_{\boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\mathcal{D})} \left[p(t|\boldsymbol{\theta}, \mathcal{D}) \right]
\end{aligned}
$$

Monte Carlo inference

Goal: Finding the expectation of a function $f(\boldsymbol{z})$ with respect to a probability distribution $p(\boldsymbol{z})$.

$$\mathbb{E}[f(\boldsymbol{z})] = \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$$



Idea: Replacing ensemble averages with empirical averages over randomly generated samples.

Monte Carlo methods

Basic formulation:

1. M i.i.d. samples $\mathbf{z}^{(m)} \sim p(\boldsymbol{z})$ are generated from $p(\boldsymbol{z})$
2. $\mathbb{E}[\mathbf{z}]$ approximated by the empirical average

$$\mathbb{E}[\mathbf{z}] \approx \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{z}^{(m)} = (\bar{z}_1, \ldots, \bar{z}_K)^{\mathsf{T}}$$

with

$$\bar{z}_j = \frac{1}{M} \sum_{m=1}^{M} z_j^{(m)}, \quad j = 1, \ldots, K, \qquad \mathsf{z}_j^{(m)} \sim p(\mathsf{z}_j)$$

3. $\mathbb{E}[f(\mathbf{z})]$ approximated by

$$\mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}[f(\mathbf{z})] = \int f(\boldsymbol{z}) p(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \approx \frac{1}{M} \sum_{m=1}^{M} f(\boldsymbol{z}^{(m)})$$

How do we sample from $p(\mathbf{z})$?

Sampling from a Bayesian network: Ancestral sampling

Assume:

$$p(\boldsymbol{z}) = \prod_{k=1}^{K} p(z_k | x_{\mathcal{P}(z_k)})$$

(ordered variables $\{\mathsf{z}_1, \ldots, \mathsf{z}_K\}$, with no arrow from any node to any lower numbered node)

Goal: Draw samples from $p(z_1, \ldots, z_K)$

Ancestral sampling:

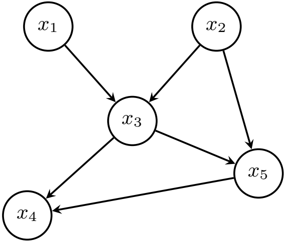
1. Draw sample for $\mathsf{z}_1 \sim p(z_1)$
2. Draw sample for $\mathsf{z}_2 \sim p(z_2 | z_{\mathcal{P}(2)})$

   $\vdots$

K. Draw sample for $\mathsf{z}_K \sim p(z_K | z_{\mathcal{P}(K)})$

We have obtained a sample from the joint distribution.

Sampling from a Bayesian network: Ancestral sampling



Sampling:

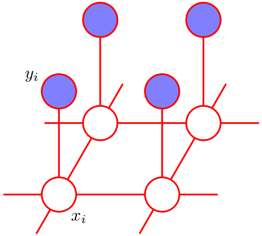
$$\mathsf{x}_1 \sim p(x_1)$$
$$\mathsf{x}_2 \sim p(x_2)$$
$$\mathsf{x}_3 \sim p(x_3|x_1, x_2)$$
$$\mathsf{x}_5 \sim p(x_5|x_2, x_3)$$
$$\mathsf{x}_4 \sim p(x_4|x_3, x_5)$$

We obtain a sample of

$$p(x_1, x_2, x_3, x_4, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_5|x_2, x_3)p(x_4|x_3, x_5)$$

Sampling in Markov random fields

Example: Ising model



Factorization:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

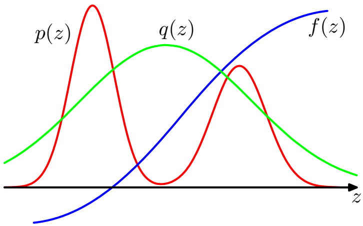
- We would like to derive $p(\boldsymbol{x}|\boldsymbol{y})$ or sample from it

Ancestral sampling not possible!

Importance sampling

Importance sampling: Approximate expectations with respect to an intractable distribution $p(\boldsymbol{z})$.

Idea: For distributions $p(\boldsymbol{z})$ from which it is difficult to sample (but we can evaluate), resort to a simpler distribution $q(\boldsymbol{z})$ (proposal distribution) from which sampling is easy.



Importance sampling

$$\mathbb{E}[f(\boldsymbol{z})] = \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z}$$

Observation:

Expectation can be expressed as an ensemble average over RV $\mathbf{z} \sim q(\boldsymbol{z})$,

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{z})] &= \int f(\boldsymbol{z})p(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \int f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\
&= \mathbb{E}_{\mathbf{z}\sim q(\boldsymbol{z})}\left[f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right]
\end{aligned}$$

if support of $q(\boldsymbol{z})$ contains that of $p(\boldsymbol{z})$

Importance sampling

$$\mathbb{E}[f(\boldsymbol{z})] = \mathbb{E}_{\mathbf{z}\sim q(\boldsymbol{z})}\left[f(\boldsymbol{z})\frac{p(\boldsymbol{z})}{q(\boldsymbol{z})}\right]$$

Importance sampling:

1. Generate M i.i.d. samples $\mathbf{z}^{(m)} \sim q(\boldsymbol{z})$
2. Compute the empirical approximation

$$\mathbb{E}[f(\boldsymbol{z})] \approx \frac{1}{M}\sum_{m=1}^{M}\frac{p(\boldsymbol{z}^{(m)})}{q(\boldsymbol{z}^{(m)})}f(\boldsymbol{z}^{(m)})$$

We express expectation in the form of a finite sum over samples $\{\boldsymbol{z}^{(m)}\}$ drawn from $q(\boldsymbol{z})$.

$\omega_m = p(\boldsymbol{z}^{(m)})/q(\boldsymbol{z}^{(m)})$: Importance weights (correct bias introduced by sampling from wrong distribution)

Importance sampling: Example

Want to compute the marginal

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) \mathrm{d}\boldsymbol{z}$$

Can be rewritten as

$$p(\boldsymbol{x}) = \int p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) \mathrm{d}\boldsymbol{z} = \mathbb{E}_{\mathbf{z} \sim p(\boldsymbol{z})}[p(\boldsymbol{x}|\boldsymbol{z})]$$

Importance sampling:

We express the marginal as an ensemble average over RV $\mathbf{z} \sim q(\boldsymbol{z})$:

$$\begin{aligned} p(\boldsymbol{x}) &= \int p(\boldsymbol{z}) p(\boldsymbol{x}|\boldsymbol{z}) \frac{q(\boldsymbol{z})}{q(\boldsymbol{z})} \mathrm{d}\boldsymbol{z} \\ &= \int p(\boldsymbol{x}|\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} q(\boldsymbol{z}) \mathrm{d}\boldsymbol{z} \\ &= \mathbb{E}_{\mathbf{z} \sim q(\boldsymbol{z})} \left[p(\boldsymbol{x}|\boldsymbol{z}) \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} \right] \end{aligned}$$

Importance sampling

Sometimes $p(\boldsymbol{z})$ can only be evaluated up to a normalization constant,

$$p(\boldsymbol{z}) = \frac{\tilde{p}(\boldsymbol{z})}{Z}$$

with $\tilde{p}(\boldsymbol{z})$ easy to evaluate but Z unknown

Importance sampling:

$$\begin{aligned}
\mathbb{E}[f(\boldsymbol{z})] &= \frac{1}{Z}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[f(\boldsymbol{z})\frac{\tilde{p}(\boldsymbol{z})}{q(\boldsymbol{z})}\right] \\
&\approx \frac{1}{Z}\frac{1}{M}\sum_{m=1}^{M}\frac{\tilde{p}(\boldsymbol{z}^{(m)})}{q(\mathbf{z}^{(m)})}f(\boldsymbol{z}^{(m)}) \\
&= \frac{1}{Z}\frac{1}{M}\sum_{m=1}^{M}\tilde{\omega}_m f(\boldsymbol{z}^{(m)})
\end{aligned}$$

Importance sampling

$$\mathbb{E}[f(\boldsymbol{z})] \approx \frac{1}{Z}\frac{1}{M}\sum_{m=1}^{M}\tilde{\omega}_m f(\boldsymbol{z}^{(m)})$$

Constant Z can be approximated as:

$$\begin{aligned} Z &= \int \tilde{p}(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \int \frac{\tilde{p}(\boldsymbol{z})}{q(\boldsymbol{z})}q(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\ &= \mathbb{E}_{\mathbf{z}\sim q(\boldsymbol{z})}\left[\frac{\tilde{p}(\boldsymbol{z})}{q(\boldsymbol{z})}\right] \approx \frac{1}{M}\sum_{i=1}^{M}\frac{\tilde{p}(\boldsymbol{z}^{(m)})}{q(\boldsymbol{z}^{(m)})} \\ &= \frac{1}{M}\sum_{i=1}^{M}\tilde{\omega}_m \end{aligned}$$

Importance sampling

A few remarks:

- How well importance sampling works depends on how well $q(\boldsymbol{z})$ matches $p(\boldsymbol{z})$
- Requires evaluation of $p(\boldsymbol{z})$ (but not sampling from it)
- Weights more regions where $p(\boldsymbol{z})$ and $|f(\boldsymbol{z})|$ are large
- Method can be very efficient (need less samples) than sampling from $p(\boldsymbol{z})$.

Example: Want to estimate the probability of a rare event $\mathcal{E}$.

- Define $f(\boldsymbol{z}) = 1\{\boldsymbol{z} \in \mathcal{E}\}$, for some set $\mathcal{E}$

Better to sample from $q(\boldsymbol{z}) \propto f(\boldsymbol{z})p(\boldsymbol{z})$ than from $p(\boldsymbol{z})$!

Markov chain Monte Carlo

Pitfall: Importance sampling may perform poorly in high dimensional spaces.

Alternative: Markov chain Monte Carlo

Markov chains

Markov chain: A sequence of RVs $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(M)}$ form a first-order Markov chain if

$$p(\boldsymbol{z}^{(i+1)}|\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(i)}) = p(\boldsymbol{z}^{(i+1)}|\boldsymbol{z}^{(i)})$$

Hence,

$$p(\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(M)}) = p(\boldsymbol{z}^{(1)}) \prod_{m=1}^{M-1} p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)})$$

- Can be specified by $p(\boldsymbol{z}^{(1)})$ and transition probabilities

$$T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) = p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)})$$

Homogeneous Markov chain: Transition probabilities are the same for all m (independent of time), $T_m(\boldsymbol{z}^{(m)}, \boldsymbol{z}^{(m+1)}) = T(\boldsymbol{z}', \boldsymbol{z})$.

Markov chains

Marginal probability:

$$p(\boldsymbol{z}^{(m+1)}) = \sum_{\boldsymbol{z}^{(m)}} p(\boldsymbol{z}^{(m+1)}|\boldsymbol{z}^{(m)})p(\boldsymbol{z}^{(m)})$$

Invariant stationary distribution: A distribution is invariant with respect to a Markov chain if each step in the chain leaves the distribution invariant.

- Let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_M)$ be a probability distribution. $\boldsymbol{\pi}$ is stationary if

$$\boldsymbol{\pi} = \boldsymbol{\pi}\boldsymbol{P}$$

- For a homogeneous Markov chain with transition probabilities $T(\boldsymbol{z}', \boldsymbol{z})$, $p^\star(\boldsymbol{z})$ is stationary if

$$p^\star(\boldsymbol{z}) = \sum_{\boldsymbol{z}'} T(\boldsymbol{z}', \boldsymbol{z})p^\star(\boldsymbol{z}')$$

Markov chain Monte Carlo

Goal: Sample from $p(\boldsymbol{z})$

Idea: Construct a Markov chain whose stationary distribution is target posterior density $p(\boldsymbol{z})$, then use Markov Chain to sample from its stationary distribution.

Idea (2): For a given $p(\boldsymbol{z})$, find a transition $p(\boldsymbol{z}'|\boldsymbol{z})$ which has $p(\boldsymbol{z})$ as its stationary distribution, i.e., for $m \to \infty$, $p(\boldsymbol{z}^{(m)})$ converges to $p(\boldsymbol{z})$ (irrespective of choice of $p(\boldsymbol{z}^{(1)})$ (ergodicity)).

- Can draw samples from Markov chain by ancestral sampling and take these as samples from $p(\boldsymbol{z})$:
  1. Initialization: Set $\boldsymbol{z}^{(1)}$
  2. At each time τ, draw sample $\boldsymbol{z}^{(\tau+1)}$ from $\mathbf{z}^{(\tau+1)} \sim p(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)})$

After a large τ all the values of $\mathbf{z}^{(\tau)}$ may be viewed as samples from $p(\boldsymbol{z})$.

Markov chain Monte Carlo

For every $p(\boldsymbol{z})$, more than one $p(\boldsymbol{z}'|\boldsymbol{z})$ with $p(\boldsymbol{z})$ as stationary distribution $\longrightarrow$ different MCMC sampling methods

- Gibbs sampling
- Metropolis-Hastings sampling
- Slice sampling
- Hamiltonian Monte Carlo

Gibbs sampling

Idea: Sample each variable in turn, conditioned on values of all other variables, i.e., given joint sample $\boldsymbol{z}^{(\tau)}$, generate new sample $\boldsymbol{z}^{(\tau+1)}$ by sampling each component in turn.

RVs $\mathsf{z}_1, \ldots, \mathsf{z}_M$ with joint distribution $p(\boldsymbol{z}) = (z_1, \ldots, z_M)$,

$$p(\boldsymbol{z}) = p(z_i | z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M) p(z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M)$$

Suppose we can easily sample from

$$p(z_i | z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_M) \triangleq p(z_i | \boldsymbol{z}_{\setminus i})$$

Gibbs sampling: At each step τ we replace value of one variable z_i by a value drawn from $p(z_i | \boldsymbol{z}_{\setminus i}^{(\tau)})$.

Gibbs sampling: Algorithm

Initialization: $\{z_i : i = 1, \ldots, M\}$ to some initial values $\{z_i^{(1)}\}$

For $\tau = 1, \ldots, T$ repeat:

1. Sample $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$
2. Sample $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, z_3^{(\tau)}, \ldots, z_M^{(\tau)})$

   $\vdots$

M. Sample $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, z_2^{(\tau+1)}, \ldots, z_{M-1}^{(\tau+1)})$

After procedure reaches stationarity, marginal density of any subset of variables can be approximated by a density estimate applied to sample values.

Need to choose initial state $z_2^{(1)}, \ldots, z_M^{(1)}$. As $T \to \infty$ effect of initialization vanishes ... but affects convergence.

Gibbs sampling

Gibbs sampling samples from required distribution:

- $p(\boldsymbol{z})$ invariant of each of Gibbs sampling steps $\longrightarrow$ of whole Markov chain

  Follows from:

  1. When sampling from $p(z_i|\boldsymbol{z}\backslash i)$, marginal $p(\boldsymbol{z}\backslash i)$ invariant
  2. We sample from correct distribution $p(z_i|\boldsymbol{z}\backslash i)$

- Must be ergodic (sufficient condition: conditional distributions not zero)

Gibbs sampling

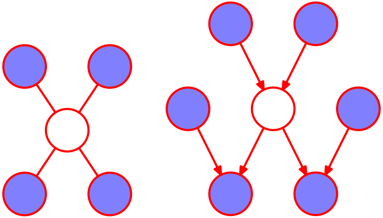
Observations:

- Gibbs sampling (generally) straightforward to implement
- Drawback: Samples are strongly dependent (strong dependencies between successive samples)
- Provided marginal of sampling distribution is correct, still a valid sampler
- Applicability depends on ability to sample from $p(z_i|\boldsymbol{z}\backslash i)$
- No need to know explicit form of $p(z_i|\boldsymbol{z}\backslash i)$, but need to be able to sample from them

Gibbs sampling

For graphical models:

Conditional distributions for individual nodes (variables) depend only on variables in Markov blanket $\longrightarrow$ to sample z_i only need to know values of neighbors



Gibbs sampling as a Markov chain

Facts:

- At sampling stage τ, we have a sample of joint variables, $\boldsymbol{z}^{(\tau)}$
- Based on $\boldsymbol{z}^{(\tau)}$, we produce new joint sample $\boldsymbol{z}^{(\tau+1)}$

Can write Gibbs sampling as a procedure that draws from

$$\mathbf{z}^{(\tau+1)} \sim q(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)})$$

for some $q(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)})$

If we update variable z_i, chosen at random from distribution $q(i)$, Gibbs sampling corresponds to drawing samples using Markov transition

$$q(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)}) = \sum_i q(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)}, i)q(i)$$

$$q(\boldsymbol{z}^{(\tau+1)}|\boldsymbol{z}^{(\tau)}, i) = p(z_i^{(\tau+1)}|\boldsymbol{z}_{\setminus i}^{(\tau)}) \prod_{j \neq i} \delta\left(z_j^{(\tau+1)}, z_j^{(\tau)}\right)$$

Want to show stationary distribution of $q(\boldsymbol{z}^{\star}|\boldsymbol{z})$ is $p(\boldsymbol{z})$ irrespective of $p(\boldsymbol{z}^{(1)})$.

Gibbs sampling as a Markov chain

Need to prove:

$$\int_{\boldsymbol{z}'} q(\boldsymbol{z}|\boldsymbol{z}')p(\boldsymbol{z}') = p(\boldsymbol{z})$$

We proceed:

$$\begin{aligned}
\int_{\boldsymbol{z}'} q(\boldsymbol{z}|\boldsymbol{z}')p(\boldsymbol{z}') &= \sum_i q(i) \int_{\boldsymbol{z}'} q(\boldsymbol{z}|\boldsymbol{z}', i)p(\boldsymbol{z}') \\
&= \sum_i q(i) \int_{\boldsymbol{z}'} \prod_{j \neq i} \delta\left(z_j, z_j'\right) p(z_i|\boldsymbol{z}'_{\setminus i})p(z_i', \boldsymbol{z}'_{\setminus i}) \\
&= \sum_i q(i) \int_{z_i'} p(z_i|\boldsymbol{z}_{\setminus i})p(z_i', \boldsymbol{z}_{\setminus i}) \\
&= \sum_i q(i)p(z_i|\boldsymbol{z}_{\setminus i})p(\boldsymbol{z}_{\setminus i}) \\
&= \sum_i q(i)p(z_i, \boldsymbol{z}_{\setminus i}) \\
&= p(\boldsymbol{z}) \sum_i q(i) = p(\boldsymbol{z})
\end{aligned}$$

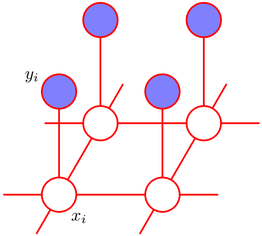
Gibbs sampling as a Markov chain

We have proven stationary distribution of $q(\boldsymbol{z}^\star|\boldsymbol{z})$ is $p(\boldsymbol{z})$ irrespective of $p(\boldsymbol{z}^{(1)})$.

- If we draw samples according to $q(\boldsymbol{z}|\boldsymbol{z}')$, in the limit we will tend to draw (dependent) samples from $p(\boldsymbol{z})$

Gibbs sampling generates a sequence of correlated samples $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \ldots$ from an easy-to-sample Markov chain $\mathbf{z}^{(1)}$ — $\mathbf{z}^{(2)}$ — ... with desired distribution $p(\boldsymbol{z})$ as stationary distribution.

Gibbs sampling for the Ising model



$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Z} \prod_{i,j} \psi_{i,j}(x_i, x_j) \prod_i \psi_i(x_i, y_i)$$

- $\mathsf{x}_i, \mathsf{y}_i \in \{+1, -1\}$ (Ising model)
- $\psi_i(x_i, y_i) = e^{\eta x_i y_i}$ and $\psi_{i,j}(x_i, x_j) = e^{\beta x_i x_j}$

Gibbs sampling for the Ising model

Goal:

$$\begin{aligned}\hat{\boldsymbol{x}} &= \arg\max_{\boldsymbol{x}} p(\boldsymbol{x}|\boldsymbol{y}) \\ &= \arg\max_{\boldsymbol{x}} p(\boldsymbol{x},\boldsymbol{y})\end{aligned}$$

Not feasible directly!

Idea: Sample from $p(\boldsymbol{x},\boldsymbol{y})$, then count the number of $+1$ and -1 for each x_i and make a decision $\longrightarrow$ Gibbs sampling!

We can write:

$$\begin{aligned}p(\boldsymbol{x},\boldsymbol{y}) &= p(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_M,\boldsymbol{y})p(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_M,\boldsymbol{y}) \\ &= p(x_i|\boldsymbol{x}_{\setminus i},\boldsymbol{y})p(\boldsymbol{x}_{\setminus i},\boldsymbol{y})\end{aligned}$$

Due to the graphical model:

$$p(x_i|\boldsymbol{x}_{\setminus i},\boldsymbol{y}) = p(x_i|\mathcal{N}(x_i),y_i)$$

Gibbs sampling: Algorithm

Initialization: $\{x_i : i = 1, \ldots, M\}$ to some initial values $\{x_i^{(1)}\}$, e.g., $x_i^{(1)} = +1$ and $x_i^{(1)} = -1$ with probability $1/2$.

For $\tau = 1, \ldots, T$ repeat:

1. Sample $x_1^{(\tau+1)} \sim p(x_1 | x_2^{(\tau)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)}, y_1)$
2. Sample $x_2^{(\tau+1)} \sim p(x_2 | x_1^{(\tau+1)}, x_3^{(\tau)}, \ldots, x_M^{(\tau)}, y_2)$

   $\vdots$

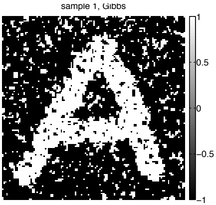
M. Sample $x_M^{(\tau+1)} \sim p(x_M | x_1^{(\tau+1)}, x_2^{(\tau+1)}, \ldots, x_{M-1}^{(\tau+1)}, y_M)$

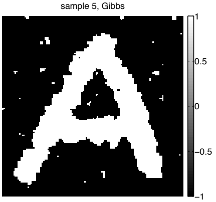
Gibbs sampling for the Ising model

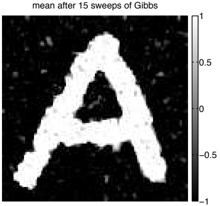
And the conditional probabilities $p(x_i|\boldsymbol{x}_{\setminus i}, \boldsymbol{y})$?

$$
\begin{aligned}
&p(x_i = +1|\boldsymbol{x}_{\setminus i}, \boldsymbol{y}) \\
&\qquad = \frac{\prod_{j\in\mathcal{N}(i)} \psi_{i,j}(+1, x_j)\psi(+1, y_i)}{\prod_{j\in\mathcal{N}(i)} \psi_{i,j}(+1, x_j)\psi(+1, y_i) + \prod_{j\in\mathcal{N}(i)} \psi_{i,j}(-1, x_j)\psi(-1, y_i)} \\
&\qquad = \frac{\exp\left(\beta\left(\sum_{j\in\mathcal{N}(i)} x_j\right) + \eta y_i\right)}{\exp\left(\beta\left(\sum_{j\in\mathcal{N}(i)} x_j\right) + \eta y_i\right) + \exp\left(-\beta\left(\sum_{j\in\mathcal{N}(i)} x_j\right) - \eta y_i\right)} \\
&\qquad = \sigma\left(2\left(\eta y_i + \beta \sum_{j\in\mathcal{N}(i)} x_j\right)\right)
\end{aligned}
$$

Gibbs sampling applied to the Ising model







- $\beta = \eta = 1$
- $y_i = x_i + n_i$, with $\mathsf{n}_i \sim \mathcal{N}(0, \sigma^2)$, $\sigma = 2$

left: sample from posterior after one sweep
center: sample from posterior after 5 sweeps
right: posterior mean, computed averaging over 15 sweeps

Back to Bayesian inference

Goal: Draw samples from joint posterior of parameters $\boldsymbol{w}$ given data $\mathcal{D}$, $p(\boldsymbol{w}|\mathcal{D})$

Gibbs sampling helpful if easy to sample from conditional distribution of each parameter given all other parameters and $\mathcal{D}$.

Reading

"Pattern recognition and machine learning,"

Chapter 11 (Intro, 11.1.4, 11.2, 11.3)