Approximate Inference

Guillaume Obozinski

Swiss Data Science Center

African Masters of Machine Intelligence, 2018-2019, AIMS, Kigali

Outline

Exact sampling with ancestral sampling

How do we sample from $$p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{\pi_i})\ ?$$

Exact sampling with ancestral sampling

How do we sample from $\quad p(x_1, \ldots, x_d) = \prod_{i=1}^{d} p(x_i \mid x_{\pi_i})$?

Algorithm 2 Ancestral sampling

1: **for** $i = 1$ to d **do**
2: $\quad z_i \leftarrow$ draw z_i from $\mathbb{P}(X_i = .|X_{\pi_i} = z_{\pi_i})$
3: **end for**
$\quad$ **return** $(z_1, \ldots, z_d)$

Gibbs Sampling

Let $X = (X_1, \ldots, X_d)$, (where X_i is associated with node i in an undirected graph) and define $X_{-i} := (X_j)_{j \neq i}$.

Gibbs Sampling

Let $X = (X_1, \ldots, X_d)$, (where X_i is associated with node i in an undirected graph) and define $X_{-i} := (X_j)_{j \neq i}$.

Gibbs algorithm

Iterate:

1. Select a node i
2. Obtain $x_i^{(t)}$ by sampling from $\mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)})$
3. Let $x_{-i}^{(t)} \leftarrow x_{-i}^{(t-1)}$

Gibbs Sampling

Let $X = (X_1, \ldots, X_d)$, (where X_i is associated with node i in an undirected graph) and define $X_{-i} := (X_j)_{j \neq i}$.

Gibbs algorithm

Iterate:

1. Select a node i
2. Obtain $x_i^{(t)}$ by sampling from $\mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)})$
3. Let $x_{-i}^{(t)} \leftarrow x_{-i}^{(t-1)}$

The node i can be selected at random (random scan Gibbs) or by cycling through the nodes (cyclic scan Gibbs).

Gibbs Sampling

Let $X = (X_1, \dots, X_d)$, (where X_i is associated with node i in an undirected graph) and define $X_{-i} := (X_j)_{j \neq i}$.

Gibbs algorithm

Iterate:

1. Select a node i
2. Obtain $x_i^{(t)}$ by sampling from $\mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)})$
3. Let $x_{-i}^{(t)} \leftarrow x_{-i}^{(t-1)}$

The node i can be selected at random (random scan Gibbs) or by cycling through the nodes (cyclic scan Gibbs).

Theorem

If $\mathbb{P}(X = x) > 0$ for all x, then the distribution of the generated random variable $X^{(t)}$ converges asymptotically to the distribution of X, i.e.,

$$\mathbb{P}(X^{(t)} = x) := \mathbb{P}(X_1^{(t)} = x_1, \dots, X_d^{(t)} = x_d) \underset{t \to \infty}{\rightarrow} \mathbb{P}(X_1 = x_1, \dots, X_d = x_d)$$

Using Gibbs to approximate $\mathbb{E}[f(X)]$

If the $(X^{(t)})_{1\leq t\leq T}$ were i.i.d. copies of X, then by the law of large numbers (LLN), we would have

$$\frac{1}{T}\sum_{t=1}^{T} f(X^{(t)}) \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}[f(X)]$$

[1] Note that this MC is due to the Gibbs sampling algorithm and has nothing to do with the graphical model!!

Using Gibbs to approximate $\mathbb{E}[f(X)]$

If the $(X^{(t)})_{1\leq t\leq T}$ were i.i.d. copies of X, then by the law of large numbers (LLN), we would have

$$\frac{1}{T}\sum_{t=1}^{T} f(X^{(t)}) \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}[f(X)]$$

But

- $\mathbb{P}(X^{(t)}=x) \neq \mathbb{P}(X=x)$ (although for $t > T_0$, $\mathbb{P}(X^{(t)}=x) \approx \mathbb{P}(X=x)$)
- $X^{(1)}, \ldots, X^{(t)}$ are not independent (they form a Markov chain[1])

[1] Note that this MC is due to the Gibbs sampling algorithm and has nothing to do with the graphical model!!

Using Gibbs to approximate $\mathbb{E}[f(X)]$

If the $(X^{(t)})_{1\leq t\leq T}$ were i.i.d. copies of X, then by the law of large numbers (LLN), we would have

$$\frac{1}{T}\sum_{t=1}^{T} f(X^{(t)}) \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}[f(X)]$$

But

- $\mathbb{P}(X^{(t)}=x) \neq \mathbb{P}(X=x)$ (although for $t > T_0$, $\mathbb{P}(X^{(t)}=x) \approx \mathbb{P}(X=x)$)
- $X^{(1)}, \ldots, X^{(t)}$ are not independent (they form a Markov chain[1])

However the LLN *for Markov chains* allows us to show that

$$\frac{1}{T}\sum_{t=1}^{T} f(X^{(t)}) \quad \xrightarrow{\text{a.s.}} \quad \mathbb{E}[f(X)]$$

[1] Note that this MC is due to the Gibbs sampling algorithm and has nothing to do with the graphical model!!

Burn-in

- In spite of the LLN for Markov chains, the samples produced at the beginning of the Gibbs algorithm are too far from having the correct distribution. So it is better to throw them away.

Burn-in

- In spite of the LLN for Markov chains, the samples produced at the beginning of the Gibbs algorithm are too far from having the correct distribution. So it is better to throw them away.
- After a certain amount of time T_0, then we can use the approximation

$$\mathbb{E}[f(X)] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} f(X^{(t)})$$

Burn-in

- In spite of the LLN for Markov chains, the samples produced at the beginning of the Gibbs algorithm are too far from having the correct distribution. So it is better to throw them away.
- After a certain amount of time T_0, then we can use the approximation

$$\mathbb{E}[f(X)] \approx \frac{1}{T - T_0} \sum_{t=T_0+1}^{T} f(X^{(t)})$$

- $\{1, \ldots, T_0\}$ is called the *burn-in* period

Application to the Ising model: setup

Remember that we have

$$p(x^{(k)};\eta) = \exp\Big(\sum_{i\in V}\eta_i\, x_i^{(k)} + \sum_{\{i,j\}\in E}\eta_{ij}\, x_i^{(k)}x_j^{(k)} - A(\eta)\Big)$$

Application to the Ising model: setup

Remember that we have

$$p(x^{(k)}; \eta) = \exp\Big(\sum_{i \in V} \eta_i \, x_i^{(k)} + \sum_{\{i,j\} \in E} \eta_{ij} \, x_i^{(k)} x_j^{(k)} - A(\eta) \Big)$$

$$\frac{1}{n} \nabla \ell(\eta) = \frac{1}{n} \sum_{k=1}^{n} \nabla \log p(x^{(k)}; \eta) = \bar{\phi} - \mu(\eta)$$

Application to the Ising model: setup

Remember that we have

$$p(x^{(k)};\eta) = \exp\Big(\sum_{i\in V}\eta_i\, x_i^{(k)} + \sum_{\{i,j\}\in E}\eta_{ij}\, x_i^{(k)}x_j^{(k)} - A(\eta)\Big)$$

$$\frac{1}{n}\nabla\ell(\eta) = \frac{1}{n}\sum_{k=1}^{n}\nabla\log p(x^{(k)};\eta) = \bar{\phi} - \mu(\eta)$$

with $\bar{\phi} = \frac{1}{n}\sum_{k=1}^{n}\phi(x^{(k)})$ and $\mu(\eta) = \nabla A(\eta) = \mathbb{E}_{\eta}[\phi(X)]$,

Application to the Ising model: setup

Remember that we have

$$p(x^{(k)};\eta) = \exp\left(\sum_{i\in V}\eta_i\, x_i^{(k)} + \sum_{\{i,j\}\in E}\eta_{ij}\, x_i^{(k)} x_j^{(k)} - A(\eta)\right)$$

$$\frac{1}{n}\nabla\ell(\eta) = \frac{1}{n}\sum_{k=1}^{n}\nabla\log p(x^{(k)};\eta) = \bar{\phi} - \mu(\eta)$$

with $\bar{\phi} = \frac{1}{n}\sum_{k=1}^{n}\phi(x^{(k)})$ and $\mu(\eta) = \nabla A(\eta) = \mathbb{E}_\eta[\phi(X)]$, where

$$\phi(x) = \begin{pmatrix} (x_i)_{i\in V} \\ (x_i x_j)_{(i,j)\in E}\end{pmatrix} \quad \text{and} \quad \mu(\eta) = \begin{pmatrix} (\mu_i)_{i\in V} \\ (\mu_{ij})_{(i,j)\in E}\end{pmatrix}$$

$$\text{with} \quad \begin{cases} \mu_i = \mathbb{E}_\eta[X_i] = \mathbb{P}_\eta(X_i = 1), \\ \mu_{ij} = \mathbb{E}_\eta[X_i X_j] = \mathbb{P}_\eta(X_i = 1, X_j = 1). \end{cases}$$

Application to the Ising model: setup

Remember that we have

$$p(x^{(k)};\eta) = \exp\Big(\sum_{i\in V}\eta_i\, x_i^{(k)} + \sum_{\{i,j\}\in E}\eta_{ij}\, x_i^{(k)}x_j^{(k)} - A(\eta)\Big)$$

$$\frac{1}{n}\nabla\ell(\eta) = \frac{1}{n}\sum_{k=1}^{n}\nabla\log p(x^{(k)};\eta) = \bar{\phi} - \mu(\eta)$$

with $\bar{\phi} = \frac{1}{n}\sum_{k=1}^{n}\phi(x^{(k)})$ and $\mu(\eta) = \nabla A(\eta) = \mathbb{E}_\eta[\phi(X)]$, where

$$\phi(x) = \begin{pmatrix}(x_i)_{i\in V}\\(x_ix_j)_{(i,j)\in E}\end{pmatrix} \quad\text{and}\quad \mu(\eta) = \begin{pmatrix}(\mu_i)_{i\in V}\\(\mu_{ij})_{(i,j)\in E}\end{pmatrix}$$

$$\text{with}\quad \begin{cases}\mu_i = \mathbb{E}_\eta[X_i] = \mathbb{P}_\eta(X_i=1),\\ \mu_{ij} = \mathbb{E}_\eta[X_iX_j] = \mathbb{P}_\eta(X_i=1, X_j=1).\end{cases}$$

Can we use Gibbs sampling to approximate μ_i and μ_{ij} ?

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model
- $x^{(1)*}, \ldots, x^{(T)*}$ be the sequence generated by Gibbs sampling.

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model
- $x^{(1)*}, \ldots, x^{(T)*}$ be the sequence generated by Gibbs sampling.

Then if $$\bar{x}_i := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)}, \quad \overline{x_i x_j} := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)},$$

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model
- $x^{(1)*}, \ldots, x^{(T)*}$ be the sequence generated by Gibbs sampling.

Then if $\quad \bar{x}_i := \frac{1}{n} \sum_{k=1}^{n} x_i^{(k)}, \quad \overline{x_i x_j} := \frac{1}{n} \sum_{k=1}^{n} x_i^{(k)} x_j^{(k)},$

$$\tilde{\mu}_i := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} x_i^{(t)*}, \quad \tilde{\mu}_{ij} := \frac{1}{T-T_0} \sum_{t=T_0+1}^{T} x_i^{(t)*} x_j^{(t)*},$$

we have

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model
- $x^{(1)*}, \ldots, x^{(T)*}$ be the sequence generated by Gibbs sampling.

Then if $\quad \bar{x}_i := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)}, \quad \overline{x_i x_j} := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)},$

$$\tilde{\mu}_i := \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} x_i^{(t)*}, \quad \tilde{\mu}_{ij} := \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} x_i^{(t)*} x_j^{(t)*}, \qquad \text{we have}$$

$$\bar{\phi}(x) = \begin{pmatrix} (\bar{x}_i)_{i\in V} \\ (\overline{x_i x_j})_{(i,j)\in E} \end{pmatrix} \quad \text{and} \quad \mu(\eta) = \begin{pmatrix} (\mu_i)_{i\in V} \\ (\mu_{ij})_{(i,j)\in E} \end{pmatrix} \text{ with } \begin{cases} \mu_i \approx \tilde{\mu}_i, \\ \mu_{ij} \approx \tilde{\mu}_{ij}. \end{cases}$$

Application to the Ising model: approximating μ_i and μ_{ij}

Let

- $x^{(1)}, \ldots, x^{(n)}$ be the i.i.d. training data used to learn the model
- $x^{(1)*}, \ldots, x^{(T)*}$ be the sequence generated by Gibbs sampling.

Then if $\quad \bar{x}_i := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)}, \quad \overline{x_i x_j} := \frac{1}{n}\sum_{k=1}^{n} x_i^{(k)} x_j^{(k)},$

$$\tilde{\mu}_i := \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} x_i^{(t)*}, \quad \tilde{\mu}_{ij} := \frac{1}{T-T_0}\sum_{t=T_0+1}^{T} x_i^{(t)*} x_j^{(t)*}, \qquad \text{we have}$$

$$\bar{\phi}(x) = \begin{pmatrix} (\bar{x}_i)_{i\in V} \\ (\overline{x_i x_j})_{(i,j)\in E} \end{pmatrix} \quad \text{and} \quad \mu(\eta) = \begin{pmatrix} (\mu_i)_{i\in V} \\ (\mu_{ij})_{(i,j)\in E} \end{pmatrix} \text{ with } \begin{cases} \mu_i \approx \tilde{\mu}_i, \\ \mu_{ij} \approx \tilde{\mu}_{ij}. \end{cases}$$

And so we can approximate the gradient of the average log-likelihood by

$$\frac{1}{n}\nabla \ell(\eta) = \bar{\phi} - \mu(\eta) \approx \begin{pmatrix} (\bar{x}_i - \tilde{\mu}_i)_{i\in V} \\ (\overline{x_i x_j} - \tilde{\mu}_{ij})_{(i,j)\in E} \end{pmatrix}.$$

Gibbs sampling for a Gibbs model

Gibbs algorithm

Iterate:

1. Select a node i
2. Obtain $x_i^{(t)}$ by sampling from $\mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)})$
3. Let $x_{-i}^{(t)} \leftarrow x_{-i}^{(t-1)}$

Gibbs sampling for a Gibbs model

Gibbs algorithm

Iterate:

1. Select a node i
2. Obtain $x_i^{(t)}$ by sampling from $\mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)})$
3. Let $x_{-i}^{(t)} \leftarrow x_{-i}^{(t-1)}$

If the distribution of X factorizes w.r.t. to an undirected graph G, why is Gibbs sampling easy to do?

$\rightarrow$ Because $\quad \mathbb{P}(X_i = \cdot \mid X_{-i} = x_{-i}^{(t-1)}) = \mathbb{P}(X_i = \cdot \mid X_M = x_M^{(t-1)})$

where M is the **Markov blanket** of node i.

Conditional probabilities in the Ising model

Denote $j \sim i$ if $j \in \mathcal{N}(i)$.

$$\begin{aligned}
\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) &= \frac{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i}) + \mathbb{P}(X_i = 0, X_{-i} = x_{-i})} \\
&= \left(1 + \frac{\mathbb{P}(X_i = 0, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}\right)^{-1} = (1 + r)^{-1}
\end{aligned}$$

Conditional probabilities in the Ising model

Denote $j \sim i$ if $j \in \mathcal{N}(i)$.

$$\begin{aligned}
\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) &= \frac{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i}) + \mathbb{P}(X_i = 0, X_{-i} = x_{-i})} \\
&= \left(1 + \frac{\mathbb{P}(X_i = 0, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}\right)^{-1} = (1+r)^{-1}
\end{aligned}$$

$$r = \frac{\exp\left(\eta_i \cdot 0 + \sum_{j\sim i} \eta_{ij} x_j \cdot 0\right) \exp\left(\sum_{j \neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i \neq j'} \eta_{jj} x_j x_{j'}\right)}{\exp\left(\eta_i \cdot 1 + \sum_{j\sim i} \eta_{ij} x_j \cdot 1\right) \exp\left(\sum_{j \neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i \neq j'} \eta_{jj} x_j x_{j'}\right)}$$

Conditional probabilities in the Ising model

Denote $j \sim i$ if $j \in \mathcal{N}(i)$.

$$\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) = \frac{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i}) + \mathbb{P}(X_i = 0, X_{-i} = x_{-i})}$$

$$= \left(1 + \frac{\mathbb{P}(X_i = 0, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}\right)^{-1} = (1+r)^{-1}$$

$$r = \frac{\exp\left(\eta_i \cdot 0 + \sum_{j\sim i} \eta_{ij} x_j \cdot 0\right) \exp\left(\sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i\neq j'} \eta_{jj} x_j x_{j'}\right)}{\exp\left(\eta_i \cdot 1 + \sum_{j\sim i} \eta_{ij} x_j \cdot 1\right) \exp\left(\sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i\neq j'} \eta_{jj} x_j x_{j'}\right)}$$

So $r = \exp\left(-\eta_i - \sum_{j\sim i} \eta_{ij} x_j\right)$ and

Conditional probabilities in the Ising model

Denote $j \sim i$ if $j \in \mathcal{N}(i)$.

$$\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) = \frac{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i}) + \mathbb{P}(X_i = 0, X_{-i} = x_{-i})}$$

$$= \left(1 + \frac{\mathbb{P}(X_i = 0, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}\right)^{-1} = (1+r)^{-1}$$

$$r = \frac{\exp\left(\eta_i \cdot 0 + \sum_{j\sim i} \eta_{ij} x_j \cdot 0\right) \exp\left(\sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i\neq j'} \eta_{jj'} x_j x_{j'}\right)}{\exp\left(\eta_i \cdot 1 + \sum_{j\sim i} \eta_{ij} x_j \cdot 1\right) \exp\left(\sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E, j\neq i\neq j'} \eta_{jj'} x_j x_{j'}\right)}$$

So $r = \exp\left(-\eta_i - \sum_{j\sim i} \eta_{ij} x_j\right)$ and

$$\boxed{\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) = \left(1 + \exp\left(-\eta_i - \sum_{j\sim i} \eta_{ij} x_j\right)\right)^{-1}}$$

Conditional probabilities in the Ising model

Denote $j \sim i$ if $j \in \mathcal{N}(i)$.

$$\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) = \frac{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i}) + \mathbb{P}(X_i = 0, X_{-i} = x_{-i})}$$

$$= \left(1 + \frac{\mathbb{P}(X_i = 0, X_{-i} = x_{-i})}{\mathbb{P}(X_i = 1, X_{-i} = x_{-i})}\right)^{-1} = (1 + r)^{-1}$$

$$r = \frac{\exp\left(\eta_i \cdot 0 + \sum_{j \sim i} \eta_{ij} x_j \cdot 0\right) \exp\left(\sum_{j \neq i} \eta_j x_j + \sum_{\{j,j'\} \in E, j \neq i \neq j'} \eta_{jj} x_j x_{j'}\right)}{\exp\left(\eta_i \cdot 1 + \sum_{j \sim i} \eta_{ij} x_j \cdot 1\right) \exp\left(\sum_{j \neq i} \eta_j x_j + \sum_{\{j,j'\} \in E, j \neq i \neq j'} \eta_{jj} x_j x_{j'}\right)}$$

So $r = \exp\left(-\eta_i - \sum_{j \sim i} \eta_{ij} x_j\right)$ and

$$\boxed{\mathbb{P}(X_i = 1 \mid X_{-i} = x_{-i}) = \left(1 + \exp\left(-\eta_i - \sum_{j \sim i} \eta_{ij} x_j\right)\right)^{-1}}$$

Note that this conditional probability has the same form as a logistic regression (but the parameters are obtained very differently)

Outline

Stochastic simulation *vs* Variational methods

To do *approximate inference*

- We have seen that we can approximate the computation of the moments using the LLN by sampling approximately form the model. In particular we have seen *Gibbs sampling*, which is one of the basic techniques from *stochastic simulation* also known as MCMC methods (Markov Chain Monte-Carlo). This methods count as well the *Metropolis-Hasting* algorithm, and many others.

[2] Similar to the maximization w.r.t. q in the E step of EM

Stochastic simulation *vs* Variational methods

To do *approximate inference*

- We have seen that we can approximate the computation of the moments using the LLN by sampling approximately form the model. In particular we have seen *Gibbs sampling*, which is one of the basic techniques from *stochastic simulation* also known as MCMC methods (Markov Chain Monte-Carlo). This methods count as well the *Metropolis-Hasting* algorithm, and many others.
- Variational methods turn the inference problem into an optimization problem[2]. Some of the most standard methods are
  - Mean field
  - Bethe variational formulations and tree-reweighted formulations
  - Expectation propagation

[2] Similar to the maximization w.r.t. q in the E step of EM

Stochastic simulation *vs* Variational methods

To do *approximate inference*

- We have seen that we can approximate the computation of the moments using the LLN by sampling approximately form the model. In particular we have seen *Gibbs sampling*, which is one of the basic techniques from *stochastic simulation* also known as MCMC methods (Markov Chain Monte-Carlo). This methods count as well the *Metropolis-Hasting* algorithm, and many others.
- Variational methods turn the inference problem into an optimization problem[2]. Some of the most standard methods are
  - Mean field
  - Bethe variational formulations and tree-reweighted formulations
  - Expectation propagation

In this lecture, we will see how *mean field* applies to the Ising model.

[2] Similar to the maximization w.r.t. q in the E step of EM

A KL divergence for p_η in exponential family form

To keep things as simple as possible we consider a discrete random variable X. Let $p(x;\eta)$ be a distribution from the exponential family whose form is

$$p(x;\eta) = \exp\Big(\langle \eta, \phi(x)\rangle - A(\eta)\Big)$$

(We assumed here $h(x) = 1$ for all x). For any distribution q, we have

$$\begin{aligned}
\mathrm{KL}(q\|p) &= -\sum_{x\in\mathcal{X}} q(x)\log\frac{p_\eta(x)}{q(x)} \\
&= \mathbb{E}_q\big[\log p_\eta(X)\big] + H(q) \\
&= \mathbb{E}_q\big[\langle \eta, \phi(X)\rangle - A(\eta)\big] + H(q) \\
&= \langle \eta, \mu_q\rangle - A(\eta) + H(q)
\end{aligned}$$

$$\text{So}\quad \begin{aligned}
A(\eta) &= \langle \eta, \mu_q\rangle + H(q) - \mathrm{KL}(q\|p_\eta) \\
&= \langle \eta, \mu_\eta\rangle + H(p_\eta) \geq \langle \eta, \mu_q\rangle + H(q)
\end{aligned}$$

Reformulating inference as a variational problem

From the previous (in)equalities we have that

$$p_\eta = \arg\max_q \langle \eta, \mu_q \rangle + H(q)$$

For a given moment parameter μ we can define its entropy as

$$\tilde{H}(\mu) = \max_q H(q) \quad \text{s.t.} \quad \mathbb{E}_q\big[\phi(X)\big] = \mu.$$

We then have

$$\boxed{\mu(\eta) = \arg\max_{\mu \in \mathcal{M}} \langle \eta, \mu \rangle + \tilde{H}(\mu)}$$

with $\mathcal{M}$ the set of allowable moment parameters, called the *marginal polytope*.
We will not show this is in this course, but it turns out that $\mathcal{M}$ is a convex set and that $\tilde{H}(\mu)$ is a concave function, however when inference is NP-hard both $\mathcal{M}$ and $\tilde{H}$ are NP-hard to compute.

Principle in variational inference

So we have

$$\boxed{\mu(\eta) = \arg\max_{\mu \in \mathcal{M}} \langle \eta, \mu \rangle + \tilde{H}(\mu)}$$

The main idea in VI is to modify the set of distributions q considered, or the set of μ considered or the set $\mathcal{M}$ to yield an optimization problem which is easier to solve.
In the Mean Field, the idea is to constraint q to be such that

$$q(x_1, \ldots, x_d) = \prod_{j=1}^{d} q_j(x_j).$$

Mean field for the Ising model

Let q be a distribution on $(X_1, \dots, X_d)$ that makes them all independent and q_j the marginal on X_j. Since X_j is binary, q_j is entirely characterized by $\mathbb{E}_q[X_j] := \mu_{q_j} := \mu_j$.

$$\langle \eta, \mu_q \rangle + H(q)$$

$$= \sum_{i \in V} \eta_i \, \mathbb{E}_q[X_i] + \sum_{\{i,j\} \in E} \eta_{ij} \, \mathbb{E}_q[X_i X_j] + H(q)$$

Mean field for the Ising model

Let q be a distribution on $(X_1, \dots, X_d)$ that makes them all independent and q_j the marginal on X_j. Since X_j is binary, q_j is entirely characterized by $\mathbb{E}_q[X_j] := \mu_{q_j} := \mu_j$.

$$\begin{aligned}
&\langle \eta, \mu_q \rangle + H(q) \\
&= \sum_{i \in V} \eta_i \, \mathbb{E}_q[X_i] + \sum_{\{i,j\} \in E} \eta_{ij} \, \mathbb{E}_q[X_i X_j] + H(q) \\
&= \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j + \sum_{i=1}^{d} H(q_i)
\end{aligned}$$

Mean field for the Ising model

Let q be a distribution on $(X_1, \ldots, X_d)$ that makes them all independent and q_j the marginal on X_j. Since X_j is binary, q_j is entirely characterized by $\mathbb{E}_q[X_j] := \mu_{q_j} := \mu_j$.

$$\begin{aligned}
&\langle \eta, \mu_q \rangle + H(q) \\
&= \sum_{i \in V} \eta_i \, \mathbb{E}_q[X_i] + \sum_{\{i,j\} \in E} \eta_{ij} \, \mathbb{E}_q[X_i X_j] + H(q) \\
&= \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j + \sum_{i=1}^{d} H(q_i) \\
&= \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \big[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \big].
\end{aligned}$$

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i\in V} \eta_i \, \mu_i + \sum_{\{i,j\}\in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1-\mu_i)\log(1-\mu_i)\right]$$

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \right]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0,1]$ but the entropy term already enforces $\mu_i \in [0,1]$.

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \right]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0,1]$ but the entropy term already enforces $\mu_i \in [0,1]$.
- the objective is non-concave since we replaced μ_{ij} by $\mu_i \mu_j$, but it is concave for each μ_i if $(\mu_j)_{j \neq i}$ is fixed.

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1 - \mu_i) \log(1 - \mu_i) \right]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0, 1]$ but the entropy term already enforces $\mu_i \in [0, 1]$.
- the objective is non-concave since we replaced μ_{ij} by $\mu_i \mu_j$, but it is concave for each μ_i if $(\mu_j)_{j \neq i}$ is fixed.
- We can compute the partial derivatives of the objective:

$$\frac{\partial \mathrm{obj}}{\partial \mu_i} = \eta_i + \sum_{j \in \mathcal{N}(i)} \eta_{ij} \mu_j - \log \frac{\mu_i}{1 - \mu_i}.$$

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \right]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0,1]$ but the entropy term already enforces $\mu_i \in [0,1]$.
- the objective is non-concave since we replaced μ_{ij} by $\mu_i \mu_j$, but it is concave for each μ_i if $(\mu_j)_{j \neq i}$ is fixed.
- We can compute the partial derivatives of the objective:

$$\frac{\partial \mathrm{obj}}{\partial \mu_i} = \eta_i + \sum_{j \in \mathcal{N}(i)} \eta_{ij} \mu_j - \log \frac{\mu_i}{1 - \mu_i}.$$

- and do a partial maximization w.r.t. μ_i by setting this partial derivative to 0.

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i \in V} \eta_i \, \mu_i + \sum_{\{i,j\} \in E} \eta_{ij} \, \mu_i \mu_j - \sum_{i=1}^{d} \left[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i) \right]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0,1]$ but the entropy term already enforces $\mu_i \in [0,1]$.
- the objective is non-concave since we replaced μ_{ij} by $\mu_i \mu_j$, but it is concave for each μ_i if $(\mu_j)_{j \neq i}$ is fixed.
- We can compute the partial derivatives of the objective:

$$\frac{\partial \mathrm{obj}}{\partial \mu_i} = \eta_i + \sum_{j \in \mathcal{N}(i)} \eta_{ij} \mu_j - \log \frac{\mu_i}{1-\mu_i}.$$

- and do a partial maximization w.r.t. μ_i by setting this partial derivative to 0. This yields

$$\mu_i^{(t+1)} = \Big(1 + \exp \big(- \eta_i - \sum_{j \in \mathcal{N}(i)} \eta_{ij} \mu_j^{(t)} \big) \Big)^{-1}$$

Mean Field optimization problem for the Ising model

$$\max_{\mu} \sum_{i\in V} \eta_i \, \mu_i + \sum_{\{i,j\}\in E} \eta_{ij} \, \mu_i\mu_j - \sum_{i=1}^{d} \big[\mu_i \log \mu_i + (1-\mu_i) \log(1-\mu_i)\big]$$

Remarks:

- there would be constraints of the form $\mu_i \in [0,1]$ but the entropy term already enforces $\mu_i \in [0,1]$.
- the objective is non-concave since we replaced μ_{ij} by $\mu_i\mu_j$, but it is concave for each μ_i if $(\mu_j)_{j\neq i}$ is fixed.
- We can compute the partial derivatives of the objective:

$$\frac{\partial \mathrm{obj}}{\partial \mu_i} = \eta_i + \sum_{j\in\mathcal{N}(i)} \eta_{ij}\mu_j - \log \frac{\mu_i}{1-\mu_i}.$$

- and do a partial maximization w.r.t. μ_i by setting this partial derivative to 0. This yields

$$\mu_i^{(t+1)} = \Big(1 + \exp\big(-\eta_i - \sum_{j\in\mathcal{N}(i)} \eta_{ij}\mu_j^{(t)}\big)\Big)^{-1}$$

Comparing Gibbs updates and Mean Field updates

Gibbs updates

$$\text{Draw} \quad x_i^{(t+1)} \sim \mathrm{Ber}(\mu_i^{(t+1)}) \quad \text{with}$$

$$\mu_i^{(t+1)} := \left(1 + \exp\left(-\eta_i - \sum_{j \in \mathcal{N}(i)} \eta_{ij}\, x_j^{(t)}\right)\right)^{-1}$$

Comparing Gibbs updates and Mean Field updates

Gibbs updates

$$\text{Draw} \quad x_i^{(t+1)} \sim \mathrm{Ber}(\mu_i^{(t+1)}) \quad \text{with}$$

$$\mu_i^{(t+1)} := \Big(1 + \exp\big(-\eta_i - \sum_{j \in \mathcal{N}(i)} \eta_{ij}\, x_j^{(t)}\big)\Big)^{-1}$$

Mean Field

$$\mu_i^{(t+1)} = \Big(1 + \exp\big(-\eta_i - \sum_{j \in \mathcal{N}(i)} \eta_{ij}\, \mu_j^{(t)}\big)\Big)^{-1}$$