Approximate inference, Sampling & Variational inference **Fall 2015**

Cours 9 — November 25

Enseignant: Guillaume Obozinski *Scribe: Basile Clément, Nathan de Lara*

9.1 Approximate inference with MCMC

9.1.1 Gibbs sampling

Let us consider an undirected graph and its associated distribution p from which we want to sample (in order to do inference for example). It is assumed that:

- It is difficult to sample directly from p.
- It is easy to sample from $\mathbb{P}_p(X_i = .|X_{-i} = x_{-i})$

The idea consists in using the Markov property so that:

$$\mathbb{P}_p(X_i = .|X_{-i} = x_{-i}) = \mathbb{P}_p(X_i = .|X_{N_i} = x_{N_i}) \tag{9.1}$$

Where N_i is the Markov blanket of the node i. Based on this, Gibbs sampling is a process that converges in distribution to p.

The most classical version of the Gibbs sampling algorithm is *cyclic scan Gibbs sampling*.

Algorithm 1 Cyclic scan Gibbs sampling

```
initialize $t = 0$ and $\mathbf{x}^0$
while $t < T$ do
    for $i = 1..d$ do
        $x_i^t \sim \mathbb{P}_p(X_i = .|X_{-i} = x_{-i}^{t-1})$
        $x_j^t = x_j^{t-1} \ \forall j \neq i$
        $t = t + 1$
    end for
end while
return $\mathbf{x}^T$
```

Another version of the algorithm called *random scan Gibbs sampling* consists in picking the index i at random at each step t.

Algorithm 2 Random scan Gibbs sampling

```
initialize $t = 0$ and $\mathbf{x}^0$
while $t < T$ do
    Draw $i$ uniformly at random in $\{1, \dots, d\}$
    $x_i^t \sim \mathbb{P}_p(X_i = .|X_{-i} = x_{-i}^{t-1})$
    $x_j^t = x_j^{t-1} \;\; \forall j \neq i$
    $t = t + 1$
end while
return $\mathbf{x}^T$
```

9.1.2 Application to the Ising Model

Let us now consider the Ising model on a graph $G = (V, E)$. X is a random variable which takes values in $\{0, 1\}^d$ with a probability distribution that depends on some parameter η:

$$p_\eta(x) = \exp\left(\sum_i \eta_i x_i + \sum_{\{i,j\}\in E} \eta_{ij} x_i x_j - A(\eta)\right) \tag{9.2}$$

To apply the Gibbs sampling algorithm, we need to compute $\mathbb{P}(X_i = x_i|X_{-i} = x_{-i})$

We have

$$\mathbb{P}(X_i = x_i, X_{-i} = x_{-i}) = \frac{1}{Z(\eta)} \exp\Big(\eta_i x_i + \sum_{j\in N_i} \eta_{ij} x_i x_j + \sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E,\, j,j'\neq i} \eta_{jj'} x_j x_{j'}\Big)$$

and thus

$$\mathbb{P}(X_{-i} = x_{-i}) = \frac{1}{Z(\eta)} \sum_{z\in\{0,1\}} \exp\Big(\eta_i z + \sum_{j\in N_i} \eta_{ij} z x_j + \sum_{j\neq i} \eta_j x_j + \sum_{\{j,j'\}\in E,\, j,j'\neq i} \eta_{jj'} x_j x_{j'}\Big)$$

Taking the ratio of the two previous quantities, the two last terms cancel out and we get

$$\mathbb{P}(X_i = x_i|X_{-i} = x_{-i}) = \frac{\exp\Big(x_i\eta_i + \sum_{j\in N_j} x_i x_j \eta_{ij}\Big)}{1 + \exp\Big(\eta_i + \sum_{j\in N_j} x_j \eta_{ij}\Big)}$$

In particular:

$$\begin{aligned}
\mathbb{P}(X_i = x_i|X_{-i} = x_{-i}) &= \frac{\exp\Big(\eta_i + \sum_{j\in N_j} x_j \eta_{ij}\Big)}{1 + \exp\Big(\eta_i + \sum_{j\in N_j} x_j \eta_{ij}\Big)} \\
&= \left(1 + \exp\left(-(\eta_i + \sum_{j\in N_i} \eta_{ij} x_j)\right)\right)^{-1} \\
&= \sigma\Big(\eta_i + \sum_{j\in N_i} \eta_{ij} x_j\Big),
\end{aligned}$$

where σ is the logistic function $\sigma : z \mapsto (1+e^{-z})^{-1}$.

Without surprise, the conditional distribution $\mathbb{P}(X_i = x_i | X_{-i} = x_{-i})$ only depends on the variables that are neighbors of i in the graph and that form its Markov blanket, since we must have

$$\mathbb{P}(X_i = x_i | X_{-i} = x_{-i}) = \mathbb{P}(X_i = x_i \mid X_{N_i} = x_{N_i}).$$

Since the conditional distribution of X_i given all other variable is Bernoulli, it is easy to sample it, using a uniform random variable.

Proposition 1 *Random scan Gibbs sampling satisfies detailed balance for π the Gibbs distribution of interest (i.e. the distribution of the graphical model).*

Proof Let us consider one step of the random scan Gibbs sampling algorithm starting from π, the distribution of the graphical model. The idea is to prove the reversibility. We first prove the result for an index i fixed, that is we prove that the transition $q_{i,Gibbs}(x^{t+1} \mid x^t)$ that only resamples the ith coordinate of x^t is reversible for π. We write $p_\pi(x_i|x_{-i})$ the conditional distribution $p_\pi(x_i|x_{-i}) = \pi(x_i, x_{-i})/(\sum_{x'_{-i}} \pi(x_i, x'_{-i}))$ of the Gibbs distribution π. Using the Kronecker symbol δ defined by $\delta(x,y) = 1$ if $x = y$ and $\delta(x,y) = 0$ else we have:

$$\begin{aligned}
\pi(x^t)\, q_{i,Gibbs}(x^{t+1} \mid x^t) &= \pi(x^t)\, \delta(x^{t+1}_{-i}, x^t_{-i})\, p_\pi(x^{t+1}_i \mid x^t_i) \\
&= \pi(x^t_{-i})\, p_\pi(x^t_i | x^t_{-i})\, \delta(x^{t+1}_{-i}, x^t_{-i})\, p_\pi(x^{t+1}_i \mid x^t_{-i}) \\
&= \pi(x^{t+1}_{-i})\, p_\pi(x^t_i \mid x^{t+1}_{-i})\, \delta(x^t_{-i}, x^{t+1}_{-i})\, p_\pi(x^{t+1}_i \mid x^{t+1}_{-i}) \\
&= \pi(x^{t+1})\, q_{i,Gibbs}(x^t \mid x^{t+1}).
\end{aligned}$$

Detailed balance for $q_{i,Gibbs}$ is valid for any i. In the random scan case, the index i being chosen at random uniformly with probability $\frac{1}{d}$, the Gibbs transition is in fact:

$$\frac{1}{d} \sum_{i=1}^{d} q_{i,Gibbs}(x^{t+1} \mid x^t)$$

The result is then obtained by taking the average over i in the previous derivation. Thus π is a stationary distribution of the random scan Gibbs transition. ■

Proposition 2 *If the Gibbs transition (e.g. random, cycle, etc.) is regular, then the MC defined by the Gibbs sampling algorithm converges in distribution to π, the Gibbs distribution.*

Exercise 1 *Extend Gibbs method to Potts model.*

Exercise 2 *Prove that the Gibbs transition is a special case of Metropolis-Hastings proposal that is always accepted.*

9.2 Variational inference

9.2.1 Overview

The goal is to do approximate inference without using sampling. Indeed, algorithms such as Metropolis-Hastings or Gibbs sampling can be very slow to converge; besides, in practice, it is very difficult to find a good stopping criterion. People working on MCMC methods try to find clever tricks to speed up the process, hence the motivation for variational methods.

Let us consider a distribution on $\mathcal{X}$ finite (but usually very large) and Q an exponential family with $q_\eta(x) = \exp(\eta^T \phi(x) - A(\eta))$. Let us assume that the distribution of interest p, that is for example the distribution of our graphical model that we are working with, is in Q. The goal is to compute $\mathbb{E}_p[\phi(x)]$.

Computing this expectation corresponds to probabilistic inference in general. For example, for Potts model, using the notation $[K] := \{1, \ldots, K\}$, we have

$$\phi(x) = \begin{pmatrix} (x_{ik})_{i \in V, k \in [K]} \\ (X_{ik} X_{jl})_{ij \in E;\, k,l \in [K]} \end{pmatrix}$$

We recall that: $p = \text{argmin}_q D(q||p)$ where:

$$D(q||p) = \sum_{x \in \mathcal{X}} q(x) \log \frac{q(x)}{p(x)} = \mathbb{E}_q[-\log p(X)] - H(q)$$

Since p is in Q, it is associated with a parameter η:

$$\begin{aligned} \mathbb{E}_q[-\log p(X)] &= \mathbb{E}_q\left[-\eta^T \phi(X) + A(\eta)\right] \\ &= -\eta^T \underbrace{\mathbb{E}_q[\phi(X)]}_{\mu(q)} + A(\eta) \end{aligned}$$

where $\mu(q)$ is the moment parameter (see course on exponential families). Thus we have:

$$-D(p||q) = \eta^T \mu(q) + H(q) - A(\eta)$$

This quantity is always negative (≤ 0) thus, for all q, $A(\eta) \geq \eta^T \mu(q) + H(q)$. Maximizing with respect to q in the exponential family leads to:

$$\boxed{A(\eta) = \max_{q \in Q} \eta^T \mu(q) + H(q)} \tag{9.3}$$

and the unique value of q that attains the maximum is p.

Remark 9.2.1 *It is possible here to get rid of q and express things only in terms of the moment. It is indeed a way to parameterize the distribution q : for a μ realizable in the exponential family there is a single distribution q_μ. The maximization problem becomes:*

$$\max_{\mu \in \mathcal{M}} \eta^T \mu + \tilde{H}(\mu),$$

where $\tilde{H}(\mu) = H(q_\mu)$ and where $\mathcal{M}$ is called the marginal polytope and is the set of all possible moments[1]*. The maximum is only attained for $\mu^* = \mu(p) = \mathbb{E}_p[\phi(X)]$, which is exactly the expectation that needs to be computed.*

It turns out that it is possible to show that $\tilde{H}$ is always a concave function, so that the optimization problem above is a convex optimization problem.

It is interesting to note that we have thus turned the probabilistic inference problem, which, a priori, required to compute expectations, that is integrals, into an optimization problem, which is furthermore convex. Unfortunately this convex optimization problem is NP-hard to solve in general because it solves the NP-hard probabilistic inference problem, and it is not possible to escape the fact that the latter is NP-hard. This optimization problem is thus in general intractable and this is because of two reasons:

- *For a general graph the marginal polytope $\mathcal{M}$ has number of faces which is exponential in the tree width of the graph.*
- *The function $\tilde{H}(\mu)$ can be extremely complicated to write explicitly.*

9.2.2 Mean field

In order to approximate the optimization problem it is possible either to change the set of distribution Q, the moments M or to change the definition of the entropy $\tilde{H}$. The mean field technique consists in choosing q in a set that makes all variables independent:

For a graphical models on variables $x_1 \dots x_d$, let us consider:

$$Q_{\perp\!\!\!\perp} = \{q \mid q(x) = q_1(x_1) \dots q_d(x_d)\},$$

the collection of distributions that make the variables $X_1, \dots, X_d$ independents.

We consider the optimization problem (8.3), but in which we replace Q by Q_π

$$\max_{q \in Q_{\perp\!\!\!\perp}} \eta^T \mu(q) + H(q). \tag{9.4}$$

Note that in general $p \notin Q_\pi$ so that the solution cannot be exactly $\mu(p)$.

In order to write this optimization problem for a Potts model, we need to write explicitly $\eta^T \mu(q)$ and $H(q)$

[1] We have seen in the course on exponential families that the distribution of maximum entropy q under the moment constraint $\mathbb{E}_q[\phi(X)] = \mu$ is also, when it exists, the distribution of maximum likelihood in the exponential family associated with the sufficient statistic ϕ. This essentially – but not exactly – shows that for any moment μ there exists a member of the exponential family, say q, such that $\mu = \mu(q)$. In fact, to be rigorous one has to be careful about what happens at points of the boundary of the set $\mathcal{M}$: the correct statement is that for every μ in the interior of $\mathcal{M}$ there exists a distribution q in the exponential family such that $\mu(q) = \mu$. The points on the boundary of $\mathcal{M}$ are only corresponding to limits of distributions of the exponential family that can be degenerate, like the Bernoulli distribution with probability 1 (or 0) for example in the Bernoulli family case, which are themselves not in the family.

Moments in the mean field formulation

$$\begin{aligned}\eta^T \mu(q) &= \eta^T \mathbb{E}_q\left[\phi(X)\right] \\ &= \sum_{i\in V, k\in[K]} \eta_{ik}\, \mathbb{E}_q\left[X_{ik}\right] + \sum_{(i,j)\in E} \eta_{ijkl}\, \mathbb{E}_q\left[X_{ik}X_{ji}\right]\end{aligned}$$

We have

$$\mathbb{E}_q\left[X_{ik}\right] = \mathbb{E}_{q_i}\left[X_{ik}\right] = \mu_{ik}(q)$$

On the other hand, the independence of the variables lead to:

$$\mathbb{E}_q\left[X_{ik}X_{jl}\right] = \mathbb{E}_{q_i}\left[X_{ik}\right]\mathbb{E}_{q_j}\left[X_{jl}\right] = \mu_{ik}\,\mu_{jl}$$

Note that if we had not constrained q to make these variables independent, we would in general have a moment here of the form $\mathbb{E}_q\left[X_{ik}X_{jl}\right] = \mu_{ijkl}$. This is the main place where the mean field approximation departs from the exact variational formulation (8.3).

Entropy $H(q)$ in the mean field formulation

By independence of the variables: $H(q) = H(q_1) + \cdots + H(q_d)$. Recall that q_i is the distribution on a single node, and that X_i is a multinomial random variable:

$$H(q_i) = -\sum_{k=1}^{K} \mathbb{P}_{q_i}(X_{ik}=1)\log \mathbb{P}_{q_i}(X_{ik}=1) = -\sum_{k=1}^{K} \mu_{ik}\log\mu_{ik}$$

Mean field formulation for the Potts model

In the end, putting everything together the optimization problem (8.4) can be written as

$$\begin{aligned}&\max_{\mu} \sum_{i,k} \eta_{ik}\,\mu_{ik} + \sum_{i,j,k,l} \eta_{ijkl}\mu_{ik}\mu_{jl} - \sum_{i,k}\mu_{ik}\log\mu_{ik} \\ &\text{s.t. } \forall i,k,\ \mu_{ik} \geq 0 \\ &\qquad \forall i,\ \sum_{k=1}^{K}\mu_{ik} = 1.\end{aligned}$$

The problem is simple to express, however we cannot longer expect that it will solve our original problem (8.3), because by restricting to the set $Q_{\perp\!\!\!\perp}$, we have restrained the forms that the moment parameters $\mu_{ijkl} := \mathbb{E}[X_{ik}X_{jl}]$ can take. In particular since p is not in $Q_{\perp\!\!\!\perp}$ in general, the optimal solution of the mean field formulation does not retrieve the correct moment parameter $\mu(p)$. The approximation will be reasonable if $\mu(p)$ is not too far from the sets of moments that are achievable by moments of distributions in $Q_{\perp\!\!\!\perp}$, since the moments of p are approximated by the moments of the closest independent distribution. Note however that the mean field approximation is much more subtle than ignoring the binary potentials in the model, which would be a too naive way of finding an “approximation" with an independent distribution.

One difficulty though is that the objective function is no longer concave, because of the products $\mu_{ik}\mu_{jl}$ which arise because of the independence assumption from the mean field approximation. Coordinate descent on each of the μ_i (not the μ_{ik}) is an algorithm of choice to solve this kind of problem. To present the algorithm we consider the case of the Ising model, which is a special case of the Potts model with 2 states for each variable.

Mean field formulation for the Ising model

When working with the Ising model is simple to reduce the number of variables by using the fact that if $\mu_{i2} = 1 - \mu_{i1}$, we therefore write μ_i for μ_{i1} and the mean field optimization problem becomes

$$\begin{aligned} &\max_{\mu} \sum_i \eta_i \, \mu_i + \sum_{i,j} \eta_{ij} \, \mu_i \mu_j - \sum_i \big(\mu_i \log \mu_i + (1-\mu_i)\log(1-\mu_i)\big) \\ &\text{s.t.} \quad \mu_i \in [0,1]. \end{aligned}$$

The stationary points for each coordinate correspond to the zeros of the partial derivatives:

$$\frac{df}{d\mu_i} = \eta_i + \sum_{j \in N_i} \eta_{ij}\mu_j - \log\frac{\mu_i}{1-\mu_i}$$

So that

$$\begin{aligned} \frac{df}{d\mu_i} = 0 \;&\Leftrightarrow\; \log \mu_i/(1-\mu_i) = \eta_i + \sum_{j \in N_i} \eta_{ij}\mu_j \\ &\Leftrightarrow\; \mu_i^* = \sigma(\eta_i + \sum_{j \in N_i} \eta_{ij}\mu_j), \end{aligned}$$

where σ is the logistic function $\sigma : z \mapsto (1+e^{-z})^{-1}$.

Note that in Gibbs sampling $x_i^{t+1} = 1$ with probability $\sigma(\eta_i + \sum_{j \in N_i} \eta_{ij} x_j)$. This is called mean field because the sampling is replaced by an approximation where it is assumed that the sample value is equal to its expectation, which for the physicist correspond to the mean field in the ferromagnetic Ising model.

Finally, lets insist that the mean field formulation is only one of the formulations for variational inference, there are several other ones, among which structured mean field, expectation propagation, loopy belief propagation (which can be reinterpreted as as solving a variational formulation as well), tree-reweighted variational inference, etc.