ML Probabilistic Modeling 2

by ambedkar@IISc

- Mote-Carlo Methods
- Typical Sets
- MCMC
- Gibbs Sampling
- MRFs with Latent Variables

Inference for probabilistic models

Definition

Probabilistic inference deals with providing answers to relevant questions once we have the model

There are two approaches to inferential problems

- ***Markov Chain Monte-Carlo***: This approach answers questions by generating samples from the distribution. (**This is not easy!**)
- ***Variational Approximation***: This formulates inference as an optimization problem

Sampling by Monte-Carlo Methods:History

Stanislaw Ulam and Von Neumann developed several Monte-Carlo methods during World War-2.

Reference: Introduction to MCMC for Machine Learning by Andrieu, Frieton, Doucet, and Jordan

Aim of Monte-Carlo methods

1. Given a probability distribution $\mathcal{P}(x)$, generate samples $\{X^{(n)}\}_{n=1}^{N}$
2. Given a Borel measurable function

$$\phi : \mathbb{R}^N \to \mathbb{R}$$

Estimate

$$E_p\phi(x) = \int \mathcal{P}(x)\phi(x)dx$$

$(1) \implies (2)$ If we solve the sampling problem *i.e.*, if we can generate N samples $\{X^{(n)}\}_{n=1}^{N}$ from $\mathcal{P}(x)$, then we can approximate $E_p\phi(x)$ with

$$\hat{\phi} = \frac{1}{N}\sum_{n=1}^{N}\phi(X^{(n)})$$

Monte-Carlo Estimates

Suppose $\{X^{(n)}\}_{n=1}^{N}$ are iid random samples with distribution $\mathcal{P}(x)$ then $\hat{\phi} = \frac{1}{N}\sum_{n=1}^{N}\phi(X^{(n)})$ is an estimate for $E_p\phi(x)$

Note that $\hat{\phi}$ is a random variable.

Let us calculate the mean of $\hat{\phi}$

$$E_p\hat{\phi} = E_p\frac{1}{N}\sum_{n=1}^{N}\phi\left(X^{(n)}\right) = \frac{1}{N}\sum_{n=1}^{N}E_p\phi\left(X^{(n)}\right) = E_p\phi(x)$$

$\Longrightarrow$ **Estimator $\hat{\phi}$ is unbiased.**

Monte Carlo Estimator cont..

Let us calculate the variance of the estimator $\hat{\phi}$

$Var\hat{\phi} = \frac{\sigma^2}{N}$ where σ^2 is the variance of X

Result

Let $X_1, ..., X_n$ be iid random variable and let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$Var\bar{X} = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}$$

As N increases covariance of $\hat{\phi}$ decreases.

- **Calculating normalizing factor in Baye's Rule**
  To obtain posterior $\mathcal{P}(X|Y)$ one of the challenging tasks is to calculate the normalizing factor or evidence.

$$\mathcal{P}(x \mid y) = \frac{\mathcal{P}(y \mid x)\mathcal{P}(x)}{\int_x \mathcal{P}\left(y \mid x'\right)\mathcal{P}\left(x'\right)dx'}$$

- **Marginalization**
  Given a joint posterior of X and Z, we may need to calculate marginal posterior

$$\mathcal{P}(x|y) = \int_{\mathcal{X}} \mathcal{P}(x, z|y)dz$$

Monte Carlo Principle

Let $\mathcal{P}$ be a probability distribution defined on $\mathcal{X}$ and X is the corresponding random variable. the principle of Monte-Carlo simulation is to draw iid set of samples,

$$\{x^{(n)}\}_{x=1}^{N} \text{ from } \mathcal{P}(X)$$

and the samples can be used to approximate the target density with the empirical point mass function.

$$\mathcal{P}_N(x) = \frac{1}{N} \sum_{n=1}^{N} \underbrace{\delta_{x^{(n)}}(x)}_{\text{Dirac Delta function}}$$

Monte Carlo Principle :Estimating the Expectation

Let $f : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function and

$$
\begin{aligned}
E_p f(x) &= \int_{\mathcal{X}} f(x)\mathcal{P}(x)dx \\
&\approx^{MC} \frac{1}{N}\sum_{n=1}^{N} f(X^{(n)}) \\
&= \hat{f} \\
&= E_{\mathcal{P}}^{N} f
\end{aligned}
$$

Law of Large Numbers

- **Weak law of large numbers**

  Let $X_1, ..X_N$ are iid with mean μ and finite variance. Then

$$\frac{1}{N}\sum_{n=1}^{N} X_n \xrightarrow{P} \mu$$

$$\text{i.e., } \mathcal{P}\left(\left|\frac{1}{N}\sum_{n=1}^{N} X_n - \mu\right| \geq \epsilon\right) \to 0 \text{ as } N \to \infty$$

- **Strong law of large numbers**

  Let $X_1, ..X_N$ are iid with common mean μ and finite fourth moment then

$$\mathcal{P}(\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{n} X_n = \mu) = 1$$

$$\lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{n} X_n =^{a.s.} \mu$$

Monte Carlo Principle :Estimating the Expectation

Now by strong law of large numbers

$$\hat{f} = E_p^N f = \frac{1}{N} \sum_{n=1}^{N} f(X^{(n)}) \xrightarrow[N \to \infty]{as} E_p f$$

Let us assume that variance f w.r.t $\mathcal{P}$

$$\sigma_f^2 = E_{\mathcal{P}(x)}[f^2(x) - (Ef(x))^2]$$

is finite

We have the variance of the estimates

$$Var(\hat{f}) = \frac{\sigma_f^2}{N} \to 0 \text{ as } N \to \infty$$

Monte Carlo Principle :Estimating the Expectation cont ..

Now, what about the error?

Given N we are interested in the following quantity

$$|\hat{f}_N - Ef(x)|$$

where $E^{(N)}f(x) = \hat{(}f)$

Theorem

Central Limit Theorem

Let $X_1, .., X_N$ be a sequence of iid random variables .Let $EX_i = \mu$ and $VarX_i = \sigma^2, i = 1, 2, ...$ .Let $S_N = \sum_{n=1}^{N} X_n$.Then

$$\frac{S_N - N\mu}{\sigma\sqrt{N}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, 1)$$

$$\hat{\mu}_N = \frac{S_N}{N} \text{ then } \sqrt{N}\left(\hat{\mu}_N - \mu\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma^2\right)$$

Monte Carlo Principle :Estimating the Expectation cont ..

Now by the central limit theorem

$$\sqrt{N}[\hat{f}_N - Ef(x)] \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2)$$

Challenges involved in sampling

When we work with MRFs the assumption is that the underlying distribution is of the form

$$\mathcal{P}(x) = \frac{\mathcal{P}^*(x)}{z}$$

(for example when we use softmax to get the same form)

Two challenges

- Calculating Normalizing constant
- High dimensionality

Challenges involved in sampling

- Calculating Z is very difficult as

$$Z = \sum_{X \in \mathcal{X}} \mathcal{P}^*(x)$$

  If x is a D-dimensional vector then $\mathcal{X}$ is typically as big as exponential in D. That is we need to perform **exponential** number of summations.

- As the dimensionality increases the problem becomes even more difficult.

Uniform Sampling

Is uniform sampling good enough? **Let us see !**

$\mathcal{P}$ is a probability distribution over space $\mathcal{X}$ and we have to calculate $E_p f(X)$

Definition

Monte Carlo Estimate *: Draw samples* $\{x^{(n)}\}$ *from* $\mathcal{P}$ *and then approximate the expectation with*

$$\hat{f}_{\mathcal{P}}(N) = E_{\mathcal{P}}^{(N)} f(X) = \frac{1}{N} \sum_{n=1}^{N} f(x^{(n)})$$

Now instead of sampling from P let us draw samples from the uniform distribution U *and approximate the expectation with*

$$\hat{f}_U(N) = E_U^{(N)} f(x) = \sum_{n=1}^{N} f(x^{(n)}) \mathcal{P}(x^{(n)})$$

Uniform Sampling cont ..

Algorithm [1]

1. Draw samples $\{x^{(n)}\}_{n=1}^{N}$ uniformly from $\mathcal{X}$(This is easy)
2. Evaluate $\mathcal{P}^*(x^{(n)})$ for $n = 1, 2, ..N$
3. set $Z_N = \sum_{n=1}^{N} P^*(x^{(n)})$
4. Return

$$\hat{f}_U(N) = E_U^{(N)} f(X) =_{(n=1)}^{N} f(x^{(n)}) \frac{\mathcal{P}^*(x^{(n)})}{Z_N}$$

as our estimation of $E_p f(x)$

[1] Assumption $\mathcal{P}(x) = \frac{\mathcal{P}^*(x)}{Z}$

Typical Sets

Some Facts :

1. Typical set is a small region T in $\mathcal{X}$, where a high dimensional probability distribution is concentrated.

2. Volume of T is given by $|T| \approx 2^{H(x)}$ where

$$H(x) = -\sum_{x \in \mathcal{X}} P(x) ln(\mathcal{P}(x))$$

which is the Shannon entropy of X.

Is this intuitive ?

Analysis of Uniform Sampling :Typical Sets

- Total size of sample space $|\mathcal{X}| = 2^N$
  Size of typical set $|T| = 2^{H(x)}$

- Probability that a sample cpomes from typical set is $\frac{2^H}{2^N}$
  i.e., the number of samples required to hit typical set is
  $N_{min} = 2^{N-H}$

- **Case1 :** If $\mathcal{P}(x)$ is uniform then $H(x)$ attains maximum
  Hence, $H = log2^N = N$*i.e.*, $R_{min} = 1$

Analysis of Uniform Sampling :Typical Sets cont ..

- **Case 2:** Suppose entropy of X is $\frac{N}{2}$(FACT: most common systems have this scenario)
- Minimum number of samples to hit typical set is $R_{min} \approx 2^{N-\frac{N}{2}} = 2^{\frac{N}{2}}$
- For $N = 1000$, $R_{min} = 10^{500}$
  R_{min} is a huge number , roughly square of the number of particles in the universe.

Uniform sampling is not useful!

Marcov Chain

Marcov Chain is time discrete stochastic process $\{x_n : n \in \mathbb{N}\}$, that satisfies Markov property. i.e.

$$P(x_{n+1} = j | x_n = i, x_{n-1} = i_{n-1}, ..., x_0 = i_0) = P(x_{n+1} = j | x_n = i)$$

Transition Matrix of Marcov chain at n^{th} step is

$$P^{(n)} = [p_{ij}^{(n)}], \text{ where } p_{ij}^{(n)} = p(x_{n+1} = j | x_n = i)$$

Homogeneous Marcov Chain A Marcov chain $x_n : x \in N$ is homogeneous if

$$P^{(n)} = P, \text{ for all } n \in \mathbb{N}$$

Marcov Chain Cont..

Assumption: Let's assume all Marcov Chains are homogeneous.

Note:- Let μ_0 be the distribution of x_0, i.e. $\mu_0(i) = P(x_0 = i)$, then

the distribution of x_n is, $\mu_n = \mu^T P^n$

Stationary distribution: A distribution π on $\mathcal{X}$ is called stationary distribution for MC $\{x_n : n \in \mathbb{N}\}$ with transition matrix P, if it satisfies

$$\pi^T = \pi^T P$$

If MC reaches to stationary distribution $\mu_n = \pi$, then

$$\mu_{n+m} = \pi, \text{ for all } m \geq 1$$

Marcov Chain Cont..

Detailed Balance Condition

A sufficient (not necessary) condition for a distribution π to be stationary w.r.t. Marcov Chain described by transition probability matrix $[P_{ij}]_{i,j\in\mathcal{X}}$ is

$$\pi(i)P_{ij} = \pi(j)P_{ji}$$

Irreducible Marcov Chain

A marcov chain $\{x_n : n \in \mathbb{N}\}$ is called irreducible, if one can get from any state in $\mathcal{X}$ to any other state in finite number of transition, i.e.

$$\forall i, j \in \mathcal{X}, \exists n > 0 \text{ such that } P(x_n = j | x_0 = i) > 0$$

Marcov Chain Cont..

Let $\{x_n : n \in \mathbb{N}\}$ be a Marcov chain with finite state space, then

if MC is irreducible $\rightarrow$ MC has a unique stationary distribution.

Aperiodic Marcov chain: MC is said to be aperiodic, if for all $i \in \mathcal{X}$

$$gcd\{n \in \mathbb{N} : P(x_n = i | x_0 = i) > 0\} = 1$$

Marcov Chain Cont..

Condition for a MC to converge to a stationary distribution

Let $\{x_n : n \in \mathbb{N}\}$ be a MC on finite state space $\mathcal{X}$. Then Marcov chain will converge to it's stationary distribution if the following properties are satisfied :

- Marcov chain is irreducible.
- Marcov chain is aperiodic.

i.e. if π is a stationary distribution, then

$$\lim_{x\to\infty} d(\mu_0^T P^n, \pi^T) = 0$$

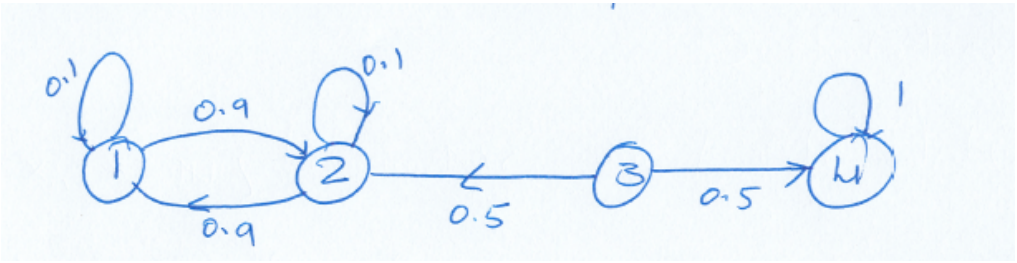
The distance d is defined as

$$d(\alpha, \beta) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\alpha(x) - \beta(x)|$$

where α and β are distributions on $\mathcal{X}$

MC : Example (Irreducibility)

Irreducibility imposes the rule that if, it is possible to go from one state to other with probability greater than 0.

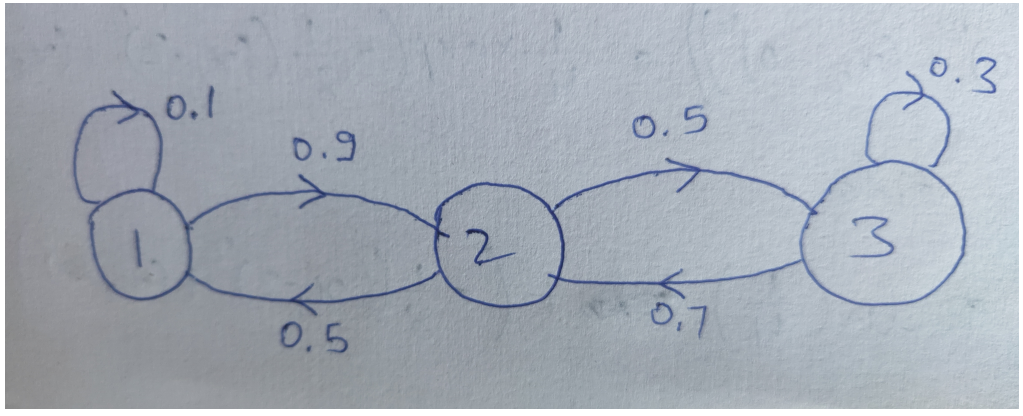


State transition example

In the given example, if we start from 1 or 2, we can never reach 4. Also if we start from 4, we can never reach 1 and 2. Hence, this MC is not irreducible.

MC: Example (Aperiodic)

Periodic State: A state i in said to be periodic, if MC can return to that state only within the steps strictly greater than 1.



State transition example

Here state 2 is periodic.

MC is said to be aperiodic, if no state in MC is periodic. Hence, given MC is not aperiodic.

Idea of Markov Chain Monte Carlo

Aim: To draw samples from a distribution Q defined on a finite space $\mathcal{X}$.

Strategy:

1. Construct an irreducible and aperiodic MC with a stationary distribution $\pi = Q$.

2. If n is large enough, a sample X_t from the constructed chain is then approximately a sample from π.

One special case of this strategy is Gibbs Sampling.

Gibbs Sampling: Idea

- Gibbs sampling belongs to a class of Metropolis-Hastings algorithms.
- This is an MCMC algorithm for producing samples from the joint probability distribution of multiple random variables.
- The main idea is to each variable (in the joint distribution) subsequently based on the conditional distribution gives the state of others.

Gibbs Sampling Cont..

Consider an MRF $X = (X_1, X_2, ..., X_D)$ w.r.t graph $G = (V, E)$, where $V = 1, 2, 3, .., D$.

- Each random variable X_d takes values from a finite set X (i.e X is takes values from $\mathcal{X}^D$).
- The joint distribution of X is, $\pi(x) = \frac{1}{Z} e^{\varepsilon(x)}$
- Assume that MRF $X^{(t)}$ starts at the state $\mathcal{X}^D$ and changes state with time.
  Now consider $\{X^{(t)} : t \in \mathbb{N}\}$ as MC, whose state space is $\mathcal{X}^D$.
- Note: $X^{(t)} = (X_1^{(t)}, ..., X_D^{(t)})$

Gibbs Sampling: Algorithm

- Start at configuration $x^{(0)} = (x_1^{(0)}, x_2^{(0)}, ..., x_D^{(0)}) \in \mathcal{X}^D$
- Repeat until convergence for t = 1,2,...
  1. set $x \leftarrow x^{t-1}$
  2. for each variable x_d, where d=1,2,..D
     1. Sample $x'_d \sim P(x_d|x_{-d})$
     2. Update $x \leftarrow (x_1, ..., x'_d, ..., x_D)$
  3. Set $x^t \leftarrow x$

Gibbs Sampling Cont...

Transition Probabilities:

Let $x, y \in \mathcal{X}^D$ are two states of MRF X. Then,

$$P_{x,y} = \begin{cases} Q(d)\pi\left(y_d \mid x_{-d}\right), & \text{if } \exists d \in [\mathrm{D}] = \mathrm{V} \text{ such that } x_{-d} = y_{-d} \\ 0, & \text{otherwise} \end{cases}$$

(i.e. MC can jump from one state x to another state y, only if they differ in one coordinate)

$$P_{xx} = \sum_{d \in V} Q(d)\pi(x_d | x_{-d})$$

Gibbs Sampling: Claims

Claim: MC is irreducible.
Since π is strictly positive and so the conditional probability distribution of single variable.

- $\Rightarrow$ Every $x_d^{(t)}$, d=1,2,...,D can take every state in $\mathcal{X}$ in single transition step.
  Hence every $(X_1^{(t)}, ..., X_D^{(t)})$ in every state $(X_1^{(t)}, ..., X_D^{(t)})$ can reach any other state in $\mathcal{X}^D$ in finitely many steps.

- $\Rightarrow$ MC is irreducible.

Gibbs Sampling: Claims (cont..)

- MC is aperiodic.
  $\because P_{xx} > 0,\ \forall x \in \mathcal{X}^D \Rightarrow$MC is aperiodic.
  $P_{xx} = \sum_{d \in V} Q(d)\pi(x_d|x_{-d})$
- MC converges to a stationary distribution.
  (Since MC is irreducible and aperiodic, it converges to a stationary distribution.)
- To show that joint distribution π of MRF is the stationary distribution of MC defined by their transition probabilities.
  (Recall balanced condition)
  $\pi(x)P_{xy} = \pi(y)P_{yx}$
  If x==y, it is easy
  If x and y differ in more than one random variable then
  $P_{xy} = P_{yx} = 0$

Gibbs Sampling: Claims (Cont..)

Assume that x and y only in the state exactly one variable X_i, i.e. $y_i = x_j$ for $i \neq j$, i,j=1,2,...,D

$$x_i = y_i$$

Now,

$$\begin{aligned}
\pi(x)P_{xy} &= \pi(x)Q(d)\pi(y_d|x_{-d}) \\
&= \pi(x_d, x_{-d})Q(d)\frac{\pi(y_d, x_{-d})}{\pi(x_{-d})} \\
&= \pi(y_d, x_{-d})Q(d)\frac{\pi(y_d, x_{-d})}{\pi(x_{-d})} \\
&= \pi(y)Q(d)\pi(x_d|x_{-d}) \\
&= \pi(y)P_{yx}
\end{aligned}$$

Learning in Undirected Graphical Models: Parameterizing MRF

Consider MRF $X = (X_1, ..., X_D)$ w.r.t. a graph $G = (V, E)$ is of the form $P(x_1, ..., x_D) = \frac{1}{\not{Z}(\theta)} \prod_{c \in \varepsilon} \Psi_c(x_c : \theta)$.
We can parameterize P as follows.

$$
\begin{aligned}
P(x_1, ..., x_D) &= \frac{1}{\not{Z}(\theta)} \exp(\sum_{c \in \varepsilon} \log \Psi_c(x_c : \theta)) \\
&= \frac{1}{\not{Z}(\theta)} \exp(\sum_{c \in \varepsilon} \sum_{x'_c} 1\{x'_c = x_c\} \log \Psi_c(x'_c : \theta)) \\
&= \frac{1}{\not{Z}(\theta)} \exp(\theta^T \phi(x))
\end{aligned}
$$

This is called exponential distribution.

Maximum Likelihood Estimation of MRF

Given Data $\mathcal{D} = \{x^{(1)}, ..., x^{(n)}\}$ i.i.d. samples, the Likelihood function is $\mathcal{L} : \theta \to \mathbb{R}$ i.e. $(\mathcal{L}(\theta) = P(\mathcal{D} \mid \theta))$. We have,

$$\begin{aligned}
\mathcal{L}(\theta \mid \mathcal{D}) &= \ln \prod_{x \in \mathcal{D}} P(x \mid \theta) \\
&= \sum_{x \in \mathcal{D}} \ln P(x \mid \theta) \\
&= \sum_{x \in \mathcal{D}} [\theta^T \phi(x) - \ln \mathcal{Z}(\theta)] \\
&= \sum_{x \in \mathcal{D}} \theta^T \phi(x) - |\mathcal{D}| \log \mathcal{Z}(\theta) \\
\bigtriangledown_\theta \mathcal{L}(\theta|\mathcal{D}) &= \sum_{x \in \mathcal{D}} \phi(x) - |\mathcal{D}| E_{x \sim P} \phi(x)
\end{aligned}$$

i.e. Maximum Likelihood estimation using gradient descent boils

MRFs with Latent Variables

Aim: To model unknown probability distribution Q.
Consider MRF x, over a graph G = (V,E). Suppose $|D| = D$.
We split x into,

- Visible: $V_1, ..., V_m$ (observed)
- Latent: $H_1, ..., H_n$ (hidden)

We have m+n = D = $|V|$
Why latent variables ?
Latent variables allow us to describe complex distribution over visible variables by means of simple conditional distribution.

MRFs with Latent Variables (Cont..)

The gibbs distribution of MRF describes the distribution of $X = (V, H)$.
We are interested in modeling V

$$P(v) = \sum_h P(v, h) = \frac{1}{\not{Z}} \sum_h \mathrm{e}^{-E(v,h)}$$

where $\not{Z} = \sum_{v,h} \mathrm{e}^{-E(v,h)}$

Roles of different variables:

- Visible variables: Correspond to observed features. Ex. Network.
- Latent variables: Introduces dependencies between observed variables. Ex. communication in network.

Log likelihood gradient of MRF with Latent Variables

Let $v = (v_1, ..., v_m)$ be single training example.

$$\begin{aligned}
\ln \mathcal{L}(\theta \mid v) &= \ln P(v \mid \theta) \\
&= \ln \frac{1}{\cancel{Z}} \sum_h e^{-E(v,h)} \\
&= \ln \sum_h e^{-E(v,h)} - \ln \sum_{v,h} e^{-E(v,h)}
\end{aligned}$$

$$\frac{\partial \ln \mathcal{L}(\theta \mid v)}{\partial \theta} = \frac{\partial}{\partial \theta} \left(\ln \sum_h e^{-E(v,h)} \right) - \frac{\partial}{\partial \theta} \left(\ln \sum_{v,\hbar} e^{-E(v,h)} \right)$$

Log likelihood gradient of MRF with Latent Variables

$$\frac{\partial \ln \mathcal{L}(\theta \mid v)}{\partial \theta} = \frac{1}{\sum_h e^{-E(v,h)}} \sum_{\hbar} e^{-E(v,h)} \frac{\partial E(v,h)}{\partial \theta}$$

$$+ \frac{1}{\sum_{v,h} e^{-E(v,a)}} \sum_{\sigma,h} e^{-[(v,h)} \frac{\partial E(v,h)}{\partial \theta}$$

$$\left[\text{We have } P(h \mid v) = \frac{P(v,h)}{P(v)} = \frac{e^{-E(v,h)}}{\sum_h e^{-E(v,h)}} \right]$$

$$= \sum_h P(h \mid v) \frac{\partial E(v,h)}{\partial \theta} + \sum_{v,h} P(v,h) \frac{\partial E(v,h)}{\partial \theta}$$

$$= E_{P(v,h)} \frac{\partial E(v,h)}{\partial \theta} - E_{P(h|v)} \frac{\partial E(v,h)}{\partial \theta}$$