Intractable Likelihood Functions

Michael Gutmann

Probabilistic Modelling and Reasoning (INFR11134)
School of Informatics, University of Edinburgh

Spring Semester 2020

Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}$$

Assume that $\mathbf{x}, \mathbf{y}, \mathbf{z}$ each are $d = 500$ dimensional, and that each element of the vectors can take $K = 10$ values.

- Topic 1: Representation We discussed reasonable weak assumptions to efficiently represent $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$.
- Topic 2: Exact inference We have seen that the same assumptions allow us, under certain conditions, to efficiently compute the posterior probability or derived quantities.

Recap

$$p(\mathbf{x}|\mathbf{y}_o) = \frac{\sum_{\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}{\sum_{\mathbf{x},\mathbf{z}} p(\mathbf{x},\mathbf{y}_o,\mathbf{z})}$$

- Topic 3: Learning How can we learn the non-negative numbers $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$ from data?
  - Probabilistic, statistical, and Bayesian models
  - Learning by parameter estimation and learning by Bayesian inference
  - Basic models to illustrate the concepts.
  - Models for factor and independent component analysis, and their estimation by maximising the likelihood.
- Issue 4: For some models, exact inference and learning is too costly even after fully exploiting the factorisation (independence assumptions) that were made to efficiently represent $p(\mathbf{x}, \mathbf{y}, \mathbf{z})$.

  Topic 4: Approximate inference and learning

Recap

Examples we have seen where inference and learning is too costly:

- Computing marginals when we cannot exploit the factorisation.
- During variable elimination, we may generate new factors that depend on many variables so that subsequent steps are costly.
- Even if we can compute $p(\mathbf{x}|\mathbf{y}_o)$, if $\mathbf{x}$ is high-dimensional, we will generally not be able to compute expectations such as

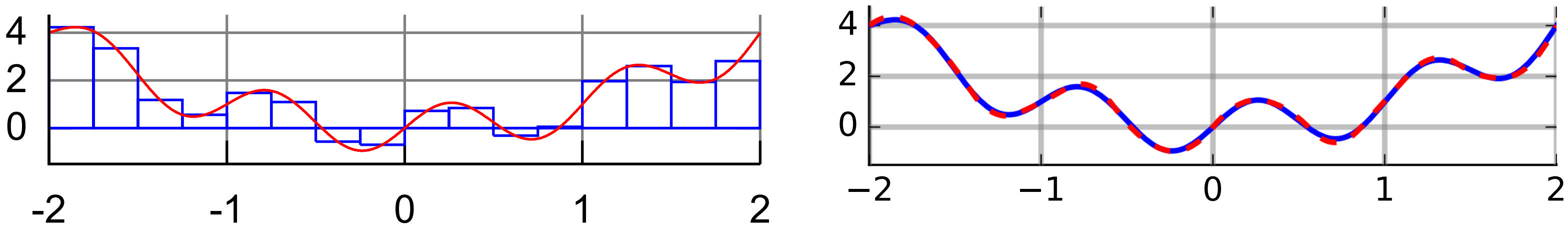
$$\mathbb{E}\left[g(\mathbf{x}) \mid \mathbf{y}_o\right] = \int g(\mathbf{x}) p(\mathbf{x}|\mathbf{y}_o) \mathrm{d}\mathbf{x}$$

  for some function g.
- Solving optimisation problems such as $\operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can be computationally costly.
- Here: focus on computational issues when evaluating $\ell(\boldsymbol{\theta})$ that are caused by high-dimensional integrals (sums).

Computing integrals

$$\int_{\mathbf{x} \in S} f(\mathbf{x}) \mathrm{d}\mathbf{x} \qquad S \subseteq \mathbb{R}^d$$

- In some cases, closed form solutions possibles.
- If $\mathbf{x}$ is low-dimensional ($d \leq 2$ or ≤ 3), highly accurate numerical methods exist (with e.g. Simpson's rule),



see https://en.wikipedia.org/wiki/Numerical_integration.

- Curse of dimensionality: Solutions feasible in low dimensions become quickly computationally prohibitive as the dimension d increases.
- We then say that evaluating the integral (sum) is computationally "intractable".

Program

1. Intractable likelihoods due to unobserved variables

2. Intractable likelihoods due to intractable partition functions

3. Combined case of unobserved variables and intractable partition functions

Program

1. Intractable likelihoods due to unobserved variables
   - Unobserved variables
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem

2. Intractable likelihoods due to intractable partition functions

3. Combined case of unobserved variables and intractable partition functions

Unobserved variables

- Observed data $\mathcal{D}$ correspond to observations of some random variables.
- Our model may contain random variables for which we do not have observations, i.e. "unobserved variables".
- Conceptually, we can distinguish between
  - hidden/latent variables: random variables that are important for the model description but for which we (normally) never observe data (see e.g. HMM, factor analysis)
  - variables for which data are missing: these are random variables that are (normally) observed but for which $\mathcal{D}$ does not contain observations for some reason
    (e.g. some people refuse to answer in polls, malfunction of the measurement device, etc.)

The likelihood in presence of unobserved variables

- Likelihood function is (proportional to the) probability that the model generates data like the observed one for parameter $\boldsymbol{\theta}$
- We thus need to know the distribution of the variables for which we have data (e.g. the “visibles” $\mathbf{v}$)
- If the model is defined in terms of the visibles and unobserved variables $\mathbf{u}$, we have to marginalise out the unobserved variables (sum rule) to obtain the distribution of the visibles

$$p(\mathbf{v};\boldsymbol{\theta}) = \int_{\mathbf{u}} p(\mathbf{u},\mathbf{v};\boldsymbol{\theta})\mathrm{d}\mathbf{u}$$

(replace with sum in case of discrete variables)

- Likelihood function is implicitly defined via an integral

$$L(\boldsymbol{\theta}) = p(\mathcal{D};\boldsymbol{\theta}) = \int_{\mathbf{u}} p(\mathbf{u},\mathcal{D};\boldsymbol{\theta})\mathrm{d}\mathbf{u},$$

which is generally intractable.

Evaluating the likelihood by solving an inference problem

- The problem of computing the integral

$$p(\mathbf{v};\boldsymbol{\theta}) = \int_{\mathbf{u}} p(\mathbf{u},\mathbf{v};\boldsymbol{\theta})\mathrm{d}\mathbf{u}$$

  corresponds to a marginal inference problem.
- Even if an analytical solution is not possible, we can sometimes exploit the properties of the model (independencies!) to numerically compute the marginal efficiently (e.g. by message passing).
- For each likelihood evaluation, we then have to solve a marginal inference problem.
- Example: In HMMs the likelihood of $\boldsymbol{\theta}$ can be computed using the alpha recursion (see e.g. Barber Section 23.2). Note that this only provides the value of $L(\boldsymbol{\theta})$ at a specific value of $\boldsymbol{\theta}$, and not the whole function.

Evaluating the gradient by solving an inference problem

- The likelihood is often maximised by gradient ascent

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$$

  where ϵ denotes the step-size.
- The gradient $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can be expressed as

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{u} \sim p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta} \right]$$

  where the expectation is taken with respect to $p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})$.
  (not obvious; we will prove it below)

Evaluating the gradient by solving an inference problem

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{u} \sim p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta} \right]$$

Interpretation:

- $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})$ is the gradient of the log-likelihood if we had observed the data $(\mathbf{u}, \mathcal{D})$ (gradient after "filling-in" data).
- $p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta})$ indicates which values of $\mathbf{u}$ are plausible given $\mathcal{D}$ (and when using parameter value $\boldsymbol{\theta}$).
- $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ is a weighted average of gradients for filled-in data where the weight indicates the plausibility of the values that are used to fill-in the missing data.

Proof

The key to the proof of

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{u} \sim p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta} \right]$$

is that $f'(x) = \log f(x)' f(x)$ for some function $f(x)$.

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) &= \nabla_{\boldsymbol{\theta}} \log \int_{\mathbf{u}} p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \\
&= \frac{1}{\int_{\mathbf{u}} p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u}} \int_{\mathbf{u}} \nabla_{\boldsymbol{\theta}} p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \\
&= \frac{\int_{\mathbf{u}} \nabla_{\boldsymbol{\theta}} p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u}}{p(\mathcal{D}; \boldsymbol{\theta})} \\
&= \frac{\int_{\mathbf{u}} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \right] p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u}}{p(\mathcal{D}; \boldsymbol{\theta})} \\
&= \int_{\mathbf{u}} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \right] p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \\
&= \mathbb{E}_{\mathbf{u} \sim p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta} \right]
\end{aligned}$$

where we have used that $p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta}) = p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) / p(\mathcal{D}; \boldsymbol{\theta})$.

How helpful is the connection to inference?

- The (log) likelihood and its gradient can be computed by solving an inference problem.
- This is helpful if the inference problems can be solved relatively efficiently.
- Allows one to use approximate inference methods (e.g. sampling) for likelihood-based learning.

Program

1. Intractable likelihoods due to unobserved variables
    - Unobserved variables
    - The likelihood function is implicitly defined via an integral
    - The gradient of the log-likelihood can be computed by solving an inference problem

2. Intractable likelihoods due to intractable partition functions

3. Combined case of unobserved variables and intractable partition functions

Program

1. Intractable likelihoods due to unobserved variables
2. Intractable likelihoods due to intractable partition functions
   - Unnormalised models and the partition function
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
3. Combined case of unobserved variables and intractable partition functions

Unnormalised statistical models

- Unnormalised statistical models: statistical models where some elements $\tilde{p}(\mathbf{x};\boldsymbol{\theta})$ do not integrate/sum to one

$$\int \tilde{p}(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x} = Z(\boldsymbol{\theta}) \neq 1$$

- Partition function $Z(\boldsymbol{\theta})$ can be used to normalise unnormalised models via

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{\tilde{p}(\mathbf{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})}$$

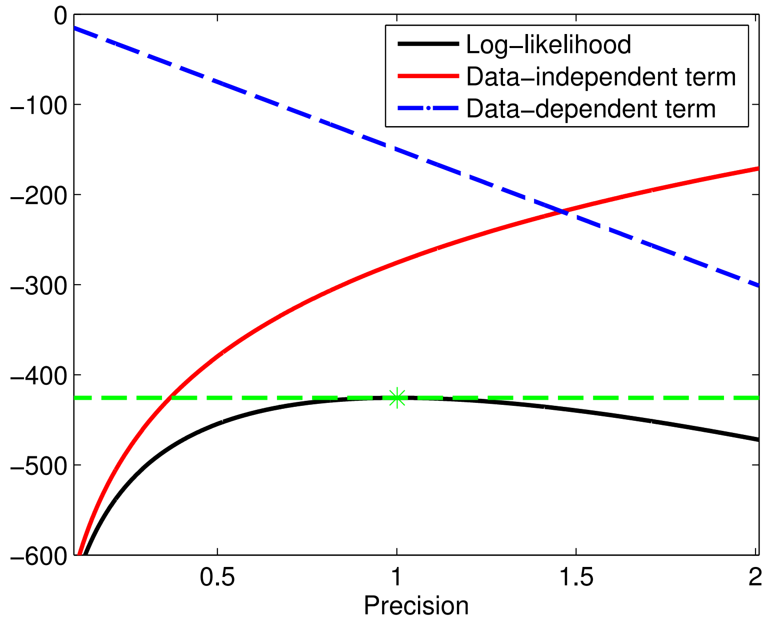
- But $Z(\boldsymbol{\theta})$ is only implicitly defined via an integral: to evaluate Z at $\boldsymbol{\theta}$, we have so compute an integral.

The partition function is part of the likelihood function

- Consider $p(x;\theta) = \frac{\tilde{p}(x;\theta)}{Z(\theta)} = \frac{\exp\left(-\theta\frac{x^2}{2}\right)}{\sqrt{2\pi/\theta}}$
- Log-likelihood function for precision $\theta \geq 0$

$$\ell(\theta) = -n \log\sqrt{\frac{2\pi}{\theta}} - \theta\sum_{i=1}^{n}\frac{x_i^2}{2}$$

- Data-dependent and independent terms balance each other.
- Ignoring $Z(\theta)$ leads to a meaningless solution.
- Errors in approximations of $Z(\theta)$ lead to errors in MLE.



The partition function is part of the likelihood function

- Assume you want to learn the parameters for an unnormalised statistical model $\tilde{p}(\mathbf{x};\boldsymbol{\theta})$ by maximising the likelihood.
- For the likelihood function, we need the normalised statistical model $p(\mathbf{x};\boldsymbol{\theta})$

$$p(\mathbf{x};\boldsymbol{\theta}) = \frac{\tilde{p}(\mathbf{x};\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \qquad Z(\boldsymbol{\theta}) = \int \tilde{p}(\mathbf{x};\boldsymbol{\theta})\mathrm{d}\mathbf{x}$$

- Partition function enters the log-likelihood function

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \log p(\mathbf{x}_i;\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \log \tilde{p}(\mathbf{x}_i;\boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta})\end{aligned}$$

- If the partition function is expensive to evaluate, evaluating and maximising the likelihood function is expensive.

The partition function in Bayesian inference

- Since the likelihood function is needed in Bayesian inference, intractable partition functions are also an issue here.
- The posterior is

$$\begin{aligned} p(\boldsymbol{\theta}; \mathcal{D}) &\propto L(\boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &\propto \frac{\tilde{p}(\mathcal{D}; \boldsymbol{\theta})}{Z(\boldsymbol{\theta})} p(\boldsymbol{\theta}) \end{aligned}$$

- Requires the partition function.
- If the partition function is expensive to evaluate, likelihood-based learning (MLE or Bayesian inference) is expensive.

Evaluating $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ by solving an inference problem

- When we interpreted MLE as moment matching, we found that (see slide 51 of *Basics of Model-Based Learning*)

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \boldsymbol{\theta}) - n \int \mathbf{m}(\mathbf{x}; \boldsymbol{\theta}) p(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}$$

$$\propto \frac{1}{n} \sum_{i=1}^{n} \mathbf{m}(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbb{E}\left[\mathbf{m}(\mathbf{x}; \boldsymbol{\theta})\right]$$

  where the expectation is taken with respect to $p(\mathbf{x}; \boldsymbol{\theta})$ and

$$\mathbf{m}(\mathbf{x}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{x}; \boldsymbol{\theta})$$

- Gradient ascent on $\ell(\boldsymbol{\theta})$ is possible if the expected value $\mathbb{E}\left[\mathbf{m}(\mathbf{x}; \boldsymbol{\theta})\right]$ can be computed.
- Problem of computing the partition function becomes problem of computing the expected value with respect to $p(\mathbf{x}; \boldsymbol{\theta})$.

Program

1. Intractable likelihoods due to unobserved variables

2. Intractable likelihoods due to intractable partition functions
   - Unnormalised models and the partition function
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem

3. Combined case of unobserved variables and intractable partition functions

Program

1. Intractable likelihoods due to unobserved variables
2. Intractable likelihoods due to intractable partition functions
3. Combined case of unobserved variables and intractable partition functions
   - Restricted Boltzmann machine example
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving two inference problems

Unnormalised models with unobserved variables

In some cases, we both have unobserved variables and intractable partition functions.

Example: Restricted Boltzmann machines (see Tutorial 2)

- Unnormalised statistical model (binary $v_i, h_i \in \{0, 1\}$)

$$p(\mathbf{v}, \mathbf{h}; \mathbf{W}, \mathbf{a}, \mathbf{b}) \propto \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)$$

- Partition function (see solutions to Tutorial 2)

$$Z(\mathbf{W}, \mathbf{a}, \mathbf{b}) = \sum_{\mathbf{v},\mathbf{h}} \exp\left(\mathbf{v}^\top \mathbf{W}\mathbf{h} + \mathbf{a}^\top \mathbf{v} + \mathbf{b}^\top \mathbf{h}\right)$$

$$= \sum_{\mathbf{v}} \exp\left(\sum_i a_i v_i\right) \prod_{j=1}^{\dim(\mathbf{h})} \left[1 + \exp\left(\sum_i v_i W_{ij} + b_j\right)\right]$$

- Becomes quickly very expensive to compute as the number of visibles increases.

Unobserved variables and intractable partition functions

- Assume we have data $\mathcal{D}$ about the visibles $\mathbf{v}$ and the statistical model is specified as

$$p(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \propto \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \qquad \int_{\mathbf{u},\mathbf{v}} \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v} = Z(\boldsymbol{\theta}) \neq 1$$

- Log-likelihood features two generally intractable integrals

$$\ell(\boldsymbol{\theta}) = \log \left[\int_{\mathbf{u}} \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \right] - \log \left[\int_{\mathbf{u},\mathbf{v}} \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v} \right]$$

Unobserved variables and intractable partition functions

- The gradient $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ is given by the difference of two expectations

  $$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{u}|\mathcal{D};\boldsymbol{\theta})}\left[\mathbf{m}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mid \mathcal{D}; \boldsymbol{\theta}\right] - \mathbb{E}_{p(\mathbf{u},\mathbf{v};\boldsymbol{\theta})}\left[\mathbf{m}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}); \boldsymbol{\theta}\right]$$

  where

  $$\mathbf{m}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})$$

  (not obvious; we will prove it below)
- The first expectation is with respect to $p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta})$.
- The second expectation is with respect to $p(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})$.
- Gradient ascent on $\ell(\boldsymbol{\theta})$ is possible if the two expectations can be computed.
- As before, we need to solve inference problems as part of the learning process.

Proof

For the second term due to the log partition function, the same calculations as before give

$$\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) = \int \left[\nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})\right] p(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \mathrm{d}\mathbf{v}$$

(replace $\mathbf{x}$ with $(\mathbf{u}, \mathbf{v})$ in the derivations on slide 50 of *Basics of Model-Based Learning*)

This is an expectation of the "moments" $\mathbf{m}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})$

$$\mathbf{m}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta}) = \left[\nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})\right]$$

with respect to $p(\mathbf{u}, \mathbf{v}; \boldsymbol{\theta})$.

Proof

For the first term, the same steps as for the case of normalised models with unobserved variables give

$$\nabla_{\boldsymbol{\theta}} \log \int_{\mathbf{u}} \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} = \frac{\int_{\mathbf{u}} \left[\nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})\right] \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u}}{\tilde{p}(\mathcal{D}; \boldsymbol{\theta})}$$

And since

$$\frac{\tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})}{\tilde{p}(\mathcal{D}; \boldsymbol{\theta})} = \frac{\tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})/Z(\boldsymbol{\theta})}{\tilde{p}(\mathcal{D}; \boldsymbol{\theta})/Z(\boldsymbol{\theta})} = \frac{p(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})}{p(\mathcal{D}; \boldsymbol{\theta})} = p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta})$$

we have

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \log \int_{\mathbf{u}} \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} &= \int_{\mathbf{u}} \left[\nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta})\right] p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \\ &= \int_{\mathbf{u}} \mathbf{m}(\mathbf{u}, \mathcal{D}; \boldsymbol{\theta}) p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta}) \mathrm{d}\mathbf{u} \end{aligned}$$

which is the posterior expectation of the "moments" when evaluated at $\mathcal{D}$, and where the expectation is taken with respect to the posterior $p(\mathbf{u}|\mathcal{D}; \boldsymbol{\theta})$.

Program recap

1. Intractable likelihoods due to unobserved variables
   - Unobserved variables
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
2. Intractable likelihoods due to intractable partition functions
   - Unnormalised models and the partition function
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
3. Combined case of unobserved variables and intractable partition functions
   - Restricted Boltzmann machine example
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving two inference problems