Intractable Likelihood Functions

Michael U. Gutmann

Probabilistic Machine Learning (INFR11298)
School of Informatics, The University of Edinburgh

Autumn Semester 2026

Recap

- Topic 1: Representation Directed and undirected graphical models; factorisation and independencies.
- Topic 2: Exact inference Variable elimination and message passing; inference for hidden Markov models.
- Topic 3: Actions and decision making Interventions and graph surgery; decision theory and loss functions.
- Topic 4: Learning Statistical and Bayesian models; parameter estimation and Bayesian inference; factor and independent component analysis by maximum likelihood.
- Issue 5: For some models, exact inference and learning is too costly even after fully exploiting the assumptions made.

  $\Rightarrow$ Topic 5: Approximate inference and learning

Recap

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

- Computing marginals when exploiting a factorisation is impossible.
- During variable elimination, we may generate new factors that depend on many variables.
- 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

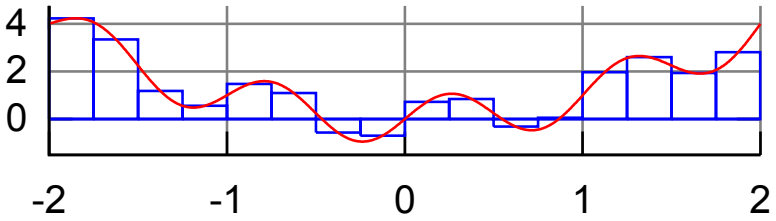
$$\mathbb{E}\left[g(\mathbf{x}) \mid \mathbf{y}_o\right] = \int g(\mathbf{x}) p(\mathbf{x}|\mathbf{y}_o) \mathrm{d}\mathbf{x}$$

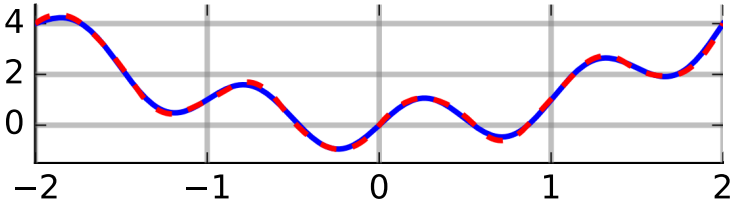
  for some function g.
- Solving optimisation problems such as $\operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ can be computationally costly.
- Here: focus on computational issues of evaluating $\ell(\boldsymbol{\theta})$ that arise from high-dimensional integrals (sums).

Computing integrals

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

- In some cases, closed form solutions possible.
- 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.

  For example, with 100 grid points per dimension, we need 100^d function evaluations: 10^{20} for $d = 10$.
- 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
   - Example: overcomplete ICA

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, the "visibles" $\mathbf{v}$.
- Our model may contain random variables for which we do not have observations, i.e. "unobserved variables" $\mathbf{h}$.
- 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

- Model $p(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$; iid data $\mathcal{D} = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, where each $\mathbf{v}_i$ has its own unobserved $\mathbf{h}_i$.
- The likelihood requires the distribution of the visibles: marginalise out $\mathbf{h}$ (sum rule)

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

(replace with sum in case of discrete variables)

- Log-likelihood: a sum of n terms, each defined via an integral

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log p(\mathbf{v}_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}_i$$

- Each term is the solution to a marginal inference problem. Generally intractable, but model structure (factorisation) sometimes helps, e.g. in case of factor trees.

A key identity

- For a non-negative function $g(\mathbf{x}; \boldsymbol{\theta})$:

$$\nabla_{\boldsymbol{\theta}} \log \left[\int g(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \right] = \mathbb{E}_{\mathbf{x} \sim \pi(\mathbf{x}; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log g(\mathbf{x}; \boldsymbol{\theta}) \right]$$

  where $\pi(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{x}; \boldsymbol{\theta}) / \int g(\mathbf{x}'; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}'$ is the normalised version of g.

- In words: the gradient of the log of an integral is an expectation with respect to the *normalised* integrand.

- Proof: using $\nabla g = (\nabla \log g) g$,

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \log \int g(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} &= \frac{\int \nabla_{\boldsymbol{\theta}} g(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}}{\int g(\mathbf{x}'; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}'} \\ &= \int \left[\nabla_{\boldsymbol{\theta}} \log g(\mathbf{x}; \boldsymbol{\theta}) \right] \frac{g(\mathbf{x}; \boldsymbol{\theta})}{\int g(\mathbf{x}'; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}'} \mathrm{d}\mathbf{x} \end{aligned}$$

- We will use this identity throughout the lecture.

Evaluating the gradient by solving an inference problem

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}_i$$

- For learning, we often do not need the value of $\ell(\boldsymbol{\theta})$ itself but want to maximise it, e.g. by gradient ascent $\boldsymbol{\theta}' = \boldsymbol{\theta} + \epsilon \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ with step-size ϵ. This requires the gradient, not $\ell(\boldsymbol{\theta})$.
- To obtain the gradient, we apply the key identity with $\mathbf{x} = \mathbf{h}_i$ to each term $\log \int p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}_i$ of $\ell(\boldsymbol{\theta})$. The normalised integrand is the posterior:

$$\frac{p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta})}{\int p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}_i} = \frac{p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta})}{p(\mathbf{v}_i; \boldsymbol{\theta})} = p(\mathbf{h}_i | \mathbf{v}_i; \boldsymbol{\theta})$$

- Hence

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}_i | \mathbf{v}_i; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta}) \right]$$

- The problem is shifted to computing an expectation with respect to a posterior (an inference problem) for each $\mathbf{v}_i$.

Evaluating the gradient by solving an inference problem

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}_i|\mathbf{v}_i;\boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta})\right]$$

Interpretation:

- $\nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i, \mathbf{v}_i; \boldsymbol{\theta})$ is the gradient of the log-likelihood if we had also observed $\mathbf{h}_i$ (gradient after "filling-in" data).
- $p(\mathbf{h}_i|\mathbf{v}_i; \boldsymbol{\theta})$ indicates which values of $\mathbf{h}_i$ are plausible given $\mathbf{v}_i$ (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 how plausible the filled-in values are.

Example: overcomplete ICA with Laplace sources

- ICA: $\mathbf{v} = \mathbf{A}\mathbf{h}$ with independent Laplace latents $p(h_j) = \frac{1}{\sqrt{2}} e^{-\sqrt{2}|h_j|}$.
- Overcomplete ($H = \dim(\mathbf{h}) > \dim(\mathbf{v}) = D$) with noise:

$$\mathbf{v} = \mathbf{A}\mathbf{h} + \mathbf{n}, \qquad \mathbf{n} \sim \mathcal{N}(\mathbf{n}; \mathbf{0}, \sigma^2 \mathbf{I}), \qquad \mathbf{A} \in \mathbb{R}^{D \times H}$$

- $\mathbf{A}$ is not invertible; we have to marginalise out $\mathbf{h}$:

$$p(\mathbf{v}; \mathbf{A}) = \int \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h}, \sigma^2 \mathbf{I}) \prod_{j=1}^{H} p(h_j) \mathrm{d}\mathbf{h}$$

- Integral does not have a closed form.

Can we estimate $p(\mathbf{v}; \mathbf{A})$ as sample average over the prior?

- The integral is an expectation with respect to the prior, so we can use Monte Carlo with samples $\mathbf{h}^{(s)} \sim p(\mathbf{h})$:

$$p(\mathbf{v}; \mathbf{A}) = \mathbb{E}_{\mathbf{h} \sim p(\mathbf{h})} \left[\mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h}, \sigma^2 \mathbf{I}) \right] \approx \frac{1}{S} \sum_{s=1}^{S} \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h}^{(s)}, \sigma^2 \mathbf{I})$$

  Easy to sample, and the estimate is unbiased (see later).
- But $\mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h}, \sigma^2 \mathbf{I})$ is non-negligible only if $\mathbf{y} = \mathbf{A}\mathbf{h} \in \mathbb{R}^D$ is within about σ of $\mathbf{v}$. For small σ, the prior probability of this is

$$P\left(\|\mathbf{y} - \mathbf{v}\| \lesssim \sigma \right) \approx p_{\mathbf{y}}(\mathbf{v}) \cdot \text{vol}(\text{ball of radius } \sigma) \propto \sigma^D$$

  (argument is not examinable)
- For N samples to land in the ball we need on the order of $S = N\sigma^{-D}$ samples, e.g. $N \times 10^{100}$ if $\sigma = 0.1$, $D = 100$, which is expensive.
- Better: use importance sampling (see later), or approaches that approximate the gradient of the log-likelihood.

Example: overcomplete ICA with Laplace sources

- Filled-in log-likelihood and gradient (if we had observed $\mathbf{h}$):

$$\log p(\mathbf{h}, \mathbf{v}; \mathbf{A}) = -\frac{1}{2\sigma^2}\|\mathbf{v} - \mathbf{A}\mathbf{h}\|^2 + \sum_j \log p(h_j) + \text{const}$$

$$\nabla_{\mathbf{A}} \log p(\mathbf{h}, \mathbf{v}; \mathbf{A}) = \frac{1}{\sigma^2}(\mathbf{v} - \mathbf{A}\mathbf{h})\mathbf{h}^\top = \frac{1}{\sigma^2}(\mathbf{v}\mathbf{h}^\top - \mathbf{A}\mathbf{h}\mathbf{h}^\top)$$

  (gradient of a least-squares regression of $\mathbf{v}$ on $\mathbf{h}$)

- Gradient of the log-likelihood for iid data $\mathbf{v}_1, \ldots, \mathbf{v}_n$:

$$\nabla_{\mathbf{A}} \ell(\mathbf{A}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\mathbf{v}_i \mathbb{E}_{p(\mathbf{h}_i|\mathbf{v}_i;\mathbf{A})}[\mathbf{h}_i^\top] - \mathbf{A}\mathbb{E}_{p(\mathbf{h}_i|\mathbf{v}_i;\mathbf{A})}[\mathbf{h}_i\mathbf{h}_i^\top] \right)$$

- Requires the posterior mean and second moment of $\mathbf{h}_i$ for every $\mathbf{v}_i$.
- Posterior $p(\mathbf{h}|\mathbf{v}; \mathbf{A}) \propto \mathcal{N}(\mathbf{v}; \mathbf{A}\mathbf{h}, \sigma^2\mathbf{I}) \prod_j p(h_j)$ is not Gaussian; its mean and second moment are not available in closed form.

Approximating the posterior $p(\mathbf{h}|\mathbf{v}; \mathbf{A})$

- The posterior is proportional to the joint distribution $p(\mathbf{h}, \mathbf{v}; \mathbf{A})$ that is easy to evaluate, which leads to options:
- Point estimate (MAP): replace the posterior by its mode $\hat{\mathbf{h}} = \operatorname{argmax}_{\mathbf{h}} \log p(\mathbf{h}, \mathbf{v}; \mathbf{A})$. Use $\mathbb{E}[\mathbf{h}] \approx \hat{\mathbf{h}}$ and $\mathbb{E}[\mathbf{h}\mathbf{h}^\top] \approx \hat{\mathbf{h}}\hat{\mathbf{h}}^\top$. Cheap, but ignores the uncertainty about $\mathbf{h}$.
- Laplace approximation: Gaussian centred at $\hat{\mathbf{h}}$, with covariance given by the curvature of $\log p(\mathbf{h}, \mathbf{v}; \mathbf{A})$ at $\hat{\mathbf{h}}$. Needs a smoothed prior here, since $|h_j|$ has a kink at 0. (sidenote, not examinable)
- Sampling: draw samples $\mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(S)}$ from $p(\mathbf{h}|\mathbf{v}; \mathbf{A})$, e.g. by Markov chain Monte Carlo, which only needs $p(\mathbf{h}, \mathbf{v}; \mathbf{A})$. Replace the expectations by sample averages.
- Variational approximation: fit a tractable distribution $q(\mathbf{h}|\mathbf{v})$ (e.g. Gaussian) by minimising $\mathrm{KL}(q(\mathbf{h}|\mathbf{v}) \| p(\mathbf{h}|\mathbf{v}; \mathbf{A}))$, and use $q(\mathbf{h}|\mathbf{v})$ instead of $p(\mathbf{h}|\mathbf{v}; \mathbf{A})$ when computing the posterior expectations.
- Details in later lectures.

How helpful is the connection to inference?

- The (log) likelihood can be computed by solving an inference problem.
- Its gradient also requires solving an inference problem, with an additional expectation step.
- Practical when inference is fast/tractable and the required expectations are computable.
- Allows one to use approximate inference methods (e.g. sampling or variational methods) 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
   - Example: overcomplete ICA
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
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
   - Example: Boltzmann machine for binary images

3. Combined case of unobserved variables and intractable partition functions

Unnormalised (energy-based) statistical models

- Unnormalised statistical models: $\tilde{p}(\mathbf{v}; \boldsymbol{\theta}) \geq 0$ that do not integrate/sum to one

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

- The partition function $Z(\boldsymbol{\theta})$ is only implicitly defined via an integral: to evaluate Z at $\boldsymbol{\theta}$, we have to compute an integral.
- $Z(\boldsymbol{\theta})$ enters the log-likelihood for iid data $\mathbf{v}_1, \ldots, \mathbf{v}_n$:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \tilde{p}(\mathbf{v}_i; \boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta})$$

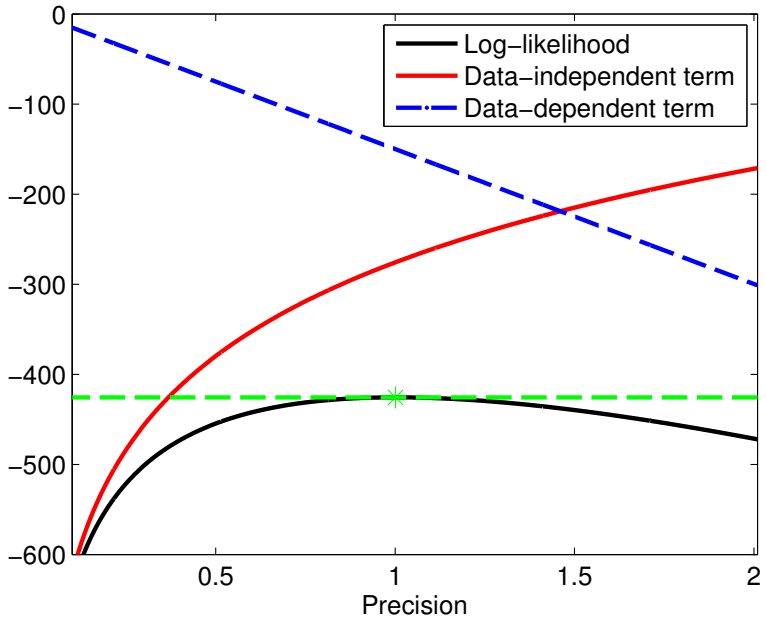
- If $Z(\boldsymbol{\theta})$ is expensive to evaluate, likelihood-based learning is expensive. This includes Bayesian inference: $p(\boldsymbol{\theta}|\mathcal{D}) \propto L(\boldsymbol{\theta})p(\boldsymbol{\theta})$ requires $Z(\boldsymbol{\theta})$ too.

Example: the partition function matters

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

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

- 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.



Evaluating the gradient by solving an inference problem

$$\nabla_{\boldsymbol{\theta}} \log \left[\int g(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \right] = \mathbb{E}_{\mathbf{x} \sim \pi(\mathbf{x}; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log g(\mathbf{x}; \boldsymbol{\theta}) \right]$$

- Apply the key identity to $\log Z(\boldsymbol{\theta}) = \log \int \tilde{p}(\mathbf{v}; \boldsymbol{\theta}) \mathrm{d}\mathbf{v}$.

$$\nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) = \mathbb{E}_{p(\mathbf{v}; \boldsymbol{\theta})} \left[\mathbf{m}(\mathbf{v}; \boldsymbol{\theta}) \right], \qquad \mathbf{m}(\mathbf{v}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{v}; \boldsymbol{\theta})$$

- Hence

$$\begin{aligned} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) &= \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{v}_i; \boldsymbol{\theta}) - n \nabla_{\boldsymbol{\theta}} \log Z(\boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \mathbf{m}(\mathbf{v}_i; \boldsymbol{\theta}) - n \, \mathbb{E}_{p(\mathbf{v}; \boldsymbol{\theta})} \left[\mathbf{m}(\mathbf{v}; \boldsymbol{\theta}) \right] \end{aligned}$$

- Compares the empirical average of the "features" $\mathbf{m}(\mathbf{v}; \boldsymbol{\theta})$ with their expectation under the model; $\boldsymbol{\theta}$ is updated to reduce the mismatch.
- The problem of computing $Z(\boldsymbol{\theta})$ becomes the problem of computing an expectation with respect to $p(\mathbf{v}; \boldsymbol{\theta})$.

Example: Boltzmann machine for binary images

- Binary image with d pixels $v_j \in \{0, 1\}$ (off/on).
- Unnormalised model with single-pixel and pairwise terms

$$\tilde{p}(\mathbf{v}; \boldsymbol{\theta}) = \exp\Big(-\frac{1}{2}\sum_{jk}\theta_{jk}v_j v_k + \sum_j \theta_j v_j\Big), \qquad \theta_{jk} = \theta_{kj},\ \theta_{jj} = 0$$

  $\theta_j > 0$ favours pixel j being on; $\theta_{jk} < 0$ favours pixels j and k being on together.
- Partition function

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{v}\in\{0,1\}^d} \exp\Big(-\frac{1}{2}\sum_{jk}\theta_{jk}v_j v_k + \sum_j \theta_j v_j\Big)$$

  is a sum over 2^d configurations.
- 28×28 image: $2^{784} \approx 10^{236}$ terms (the observable universe is estimated to have about 10^{80} atoms).

Example: Boltzmann machine for binary images

- Feature for θ_{jk}, $j \neq k$: $\dfrac{\partial}{\partial \theta_{jk}} \log \tilde{p}(\mathbf{v}; \boldsymbol{\theta}) = -v_j v_k$ (where we used that $\theta_{jk} = \theta_{kj}$)
- Gradient for images $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(n)}$, with $v_j^{(i)}$ denoting pixel j of image i:

$$\frac{1}{n}\frac{\partial \ell}{\partial \theta_{jk}} = -\underbrace{\frac{1}{n}\sum_{i=1}^{n} v_j^{(i)} v_k^{(i)}}_{\text{both on in data}} + \underbrace{\mathbb{E}_{p(\mathbf{v};\boldsymbol{\theta})}\left[v_j v_k\right]}_{\text{both on under model}}$$

- At the MLE, the model reproduces how often each pair of pixels is on in the data.
- The model term $\mathbb{E}_{p(\mathbf{v};\boldsymbol{\theta})}[v_j v_k] = \sum_{\mathbf{v}} v_j v_k \, \tilde{p}(\mathbf{v}; \boldsymbol{\theta}) / Z(\boldsymbol{\theta})$ involves two sums over all 2^d configurations of $\mathbf{v}$: the numerator and $Z(\boldsymbol{\theta})$. $\Rightarrow$ Approximate the expectation by sampling from the model, which does not require $Z(\boldsymbol{\theta})$ (see later).

Program

1. Intractable likelihoods due to unobserved variables

2. Intractable likelihoods due to intractable partition functions
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
   - Example: Boltzmann machine for binary images

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 exercises)

- Unnormalised statistical model (binary $v_j, h_k \in \{0, 1\}$)

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

- Partition function

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

- The sum over $\mathbf{h}$ can be done analytically, but a sum over $2^{\dim(\mathbf{v})}$ terms remains (see exercises). Very expensive as the number of visibles increases.

Unobserved variables and intractable partition functions

- Statistical model specified via an unnormalised $\tilde{p}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$:

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

- Log-likelihood for iid data $\mathbf{v}_1, \ldots, \mathbf{v}_n$ features two generally intractable integrals

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log \int \tilde{p}(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h} - n \log \int \tilde{p}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta}) \mathrm{d}\mathbf{h} \mathrm{d}\mathbf{v}$$

- An expression for $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$ follows from the key identity

$$\nabla_{\boldsymbol{\theta}} \log \left[\int g(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \right] = \mathbb{E}_{\mathbf{x} \sim \pi(\mathbf{x}; \boldsymbol{\theta})} \left[\nabla_{\boldsymbol{\theta}} \log g(\mathbf{x}; \boldsymbol{\theta}) \right]$$

  applied to both terms, with $\mathbf{x} = \mathbf{h}$ for the first and $\mathbf{x} = (\mathbf{h}, \mathbf{v})$ for the second.

Unobserved variables and intractable partition functions

- With $\mathbf{m}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \log \tilde{p}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \mathbb{E}_{p(\mathbf{h}|\mathbf{v}_i;\boldsymbol{\theta})} \left[\mathbf{m}(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta})\right] - n\, \mathbb{E}_{p(\mathbf{h},\mathbf{v};\boldsymbol{\theta})} \left[\mathbf{m}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})\right]$$

- First term, due to the unobserved variables: normalising $\tilde{p}(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta})$ over $\mathbf{h}$ gives the posterior $p(\mathbf{h}|\mathbf{v}_i; \boldsymbol{\theta})$.
- Second term, due to the partition function: normalising $\tilde{p}(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$ over $(\mathbf{h}, \mathbf{v})$ gives the model $p(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$.
- Gradient ascent on $\ell(\boldsymbol{\theta})$ is possible if the two expectations can be computed. Typically done by taking a sample average, and hence requires sampling from $p(\mathbf{h}|\mathbf{v}_i; \boldsymbol{\theta})$ and $p(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$.

Summary

- Unobserved variables and unnormalised models enable flexible modelling: complex dependencies can arise through latent variables, and flexible $\tilde{p}$ are easy to specify.
- But we pay a price when learning: the likelihood involves intractable integrals.
- Via the key identity, the gradient of the log-likelihood can be written in terms of expectations. These are generally intractable too, but can be approximated by sampling or variational methods (coming lectures).

Source	**Log-likelihood** $\ell(\boldsymbol{\theta})$	**Gradient:** $\mathbb{E}$ **w.r.t.**
Unobserved variables	$\sum_i \log \int p(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h}$	posterior $p(\mathbf{h}\|\mathbf{v}_i; \boldsymbol{\theta})$
Partition function	$\sum_i \log \tilde{p}(\mathbf{v}_i; \boldsymbol{\theta}) - n \log Z(\boldsymbol{\theta})$	model $p(\mathbf{v}; \boldsymbol{\theta})$
Both	$\sum_i \log \int \tilde{p}(\mathbf{h}, \mathbf{v}_i; \boldsymbol{\theta}) \mathrm{d}\mathbf{h} - n \log Z(\boldsymbol{\theta})$	$p(\mathbf{h}\|\mathbf{v}_i; \boldsymbol{\theta})$, $p(\mathbf{h}, \mathbf{v}; \boldsymbol{\theta})$

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
   - Example: overcomplete ICA
2. Intractable likelihoods due to intractable partition functions
   - The likelihood function is implicitly defined via an integral
   - The gradient of the log-likelihood can be computed by solving an inference problem
   - Example: Boltzmann machine for binary images
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