Lecture X

Some notes on inference
Sampling
Sampling

We may make approximations to densities in order to make the integrals tractable, but they are still *approximations*. In many cases we must have greater accuracy. Monte-Carlo methods are approaches to sampling which enable expectations and integrals to be evaluated with ease, without the requirements of making gross approximations.

The problems we wish to solve are therefore of the form,

- to generate samples from $f(w)$ and, more importantly, to evaluate the *partition function* or normalising constant, $Z = \int f(w)dw$,

- to estimate expectations under this distribution of the form

$$\langle g \rangle = \int f(w)g(w)dw$$
The need to sample efficiently

Consider we can evaluate $f(w)$ for any $w$. Can’t we just generate random (uniform) samples $w$, evaluate $f(w)$ and then add them all up to get $Z \approx \sum_r f(w_r)$? Sure—we can do this, but...

Imagine we have a 1-D ($d = 1$) system and we generate $w$ in a set of 100 uniformly spaced bins. This is nice and easy. Now let $d = 2$. To get the same resolution, we have to draw $100 \times 100$ samples. Still OK? Now imagine $d = 20$, we have to draw $10^{2\times20}$ samples. Imagine we have a fast computer, and can draw $10^6$ per second. We will have to wait $10^{34}$ seconds, or about $10^{16}$ universe lifetimes (the age of our universe is about $10^{18}$ s) – a rather long time. So there must be better ways!
Importance sampling

We can evaluate $f(w)$ but cannot sample from it directly.

Use a simpler function, $q(w)$, as a proposal.

Draw $R$ samples from $q(w)$, $\{w_r\}$.

Note that where $q > f$ we will be over-sampling etc.

Define importance weights $a_r$

$$a_r = \frac{f(w_r)}{q(w_r)}$$

$$\langle g \rangle = \frac{\sum_r a_r g(w_r)}{\sum_r a_r}$$
Problems

So what are the problems with importance sampling?

- getting a proposal which covers the true distribution - the better the fit the lower the variance in the estimates,

- sampling in high dimensions from proposal distributions

The former is less of an issue than the latter, as we may have some qualitative knowledge about the true density. The latter is a key issue, as it is fine to draw samples from \( q \) in a high-d space, but the weights will be dominated by a few very large \( a_r \) hence the estimator above will be of high variance (in effect we are averaging over only a small number of samples). This is overcome by getting \( q \) as close as possible to \( f \) – not a trivial task.
Rejection sampling

\[ cq(w) > f(w) \]
Metropolis-Hastings

Both importance and rejection sampling work well if \( q(w) \approx f(w) \). This is a pig to ensure, as we have a single, fixed proposal. The Metropolis algorithm (named after Nick Metropolis) instead uses a simple proposal \( q \) which depends on the current state, \( w^{(t)} \) say. The proposal, \( q(w; w^{(t)}) \), can be any simple density, e.g. a Gaussian located at \( w^{(t)} \). This movement of the proposal density means that it need not look anything like \( f(w) \).

As before, we consider we can evaluate \( f(w) \) for any \( w \) and draw some \( w' \) from \( q \) at time step \( (t) \). To decide whether we accept the sample we compute

\[
acc = \min \left\{ 1, \frac{f(w')}{f(w^{(t)})} \frac{q(w^{(t)}; w')}{q(w'; w^{(t)})} \right\}
\]

and then draw a random variable \( u \sim U[0, 1] \) and accept if \( u \leq acc \). If we accept then we let \( w^{(t+1)} = w' \) else \( w^{(t+1)} = w^{(t)} \).
**Simple line fit**

**A simple example:** Fitting a straight line through a set of data points.
Consider, as before, the set of $N$ data points, $(x, y)$. We will model this data using a generative model of the form

$$y = w_1 x + w_2 + \eta$$

in which $w_1, w_2$ are parameters and $\eta$ is a noise process which is assumed white (Gaussian) with a fixed variance (to make life easier!). What we would like to do is to infer

$$P(w_1, w_2 | X = \{x\})$$
So we consider a MH algorithm in which, as in the previous example, the proposal is simple and consists of a hybrid mixture of a low & high variance Gaussian – let this be $q(w_1, w_2)$.

MH states that the acceptance is

$$\text{acc} = \min \left\{ 1, \frac{f(w'_1, w'_2) \cdot q(w_1^{(t)}, w_2^{(t)}; w'_1, w'_2)}{f(w_1^{(t)}, w_2^{(t)}) \cdot q(w'_1, w'_2; w_1^{(t)}, w_2^{(t)})} \right\}$$

but due to the fact that $q$ is symmetric, the terms in $q$ cancel – great! Now re-writing $f(w_1, w_2) \equiv P(w_1, w_2|X)$ (i.e. this is what we want) and noting that

$$P(w_1, w_2|X) = P(X|w_1, w_2)P(w_1, w_2)/P(X)$$

so

$$\text{acc} = \min \left\{ 1, \frac{P(X|w_1', w_2')P(w_1', w_2')}{P(X|w_1^{(t)}, w_2^{(t)})P(w_1^{(t)}, w_2^{(t)})} \right\}$$

having cancelled the $P(X)$ terms, also as the priors cancel we get the even simpler

$$\text{acc} = \min \left\{ 1, \frac{P(X|w_1', w_2')}{P(X|w_1^{(t)}, w_2^{(t)})} \right\}$$
Simple line fit

Draws from posterior $p(w_1, w_2)$

$p(y|x)$
Approximate Bayes
MAP revisited

Consider $p(t|x, w)p(w)$ in the case of $L_2$ loss function between $y, t$ and a (generalised) linear model for $y(x, w) = w^T \Phi(x)$

$$L(x, w) = \frac{1}{2} \mathbb{E} \{(y_{x,w} - t)^2\} + \frac{\lambda}{2} w^T w$$

The simple penalised least-squares solution via the pseudo-inverse is

$$w = (\lambda I + \Phi^T \Phi)^{-1} \Phi^T t$$

From a Bayesian perspective we are mode finding - the MAP solution

$$p(t|x, w)p(w) = \mathcal{N}(t|y(x, w^*), \beta^{-1})\mathcal{N}(w|0, \alpha^{-1})$$

Where $\lambda = \alpha/\beta$
What we want

To integrate out the uncertainty over $w$

To perform full Bayesian integration is costly

We can make some useful approximations

Start by looking at the Laplace (quadratic) approximation
Laplace approximation

Consider a situation in which we are given some information, $x$, and require an inference of $t$ given $x$, conditioned on a model with parameters $w$. We may formulate the likelihood $p(t|x, w)$, and then perform marginalisation to remove the ‘nuisance’ parameters, $w$, i.e.

$$p(t|x) = \int_{\text{all } w} p(t|x, w)p(w)dw$$

How might we make approximations so as to keep the above analytic?
Laplace

One approach is to take the integrand to be well-approximated by a Gaussian, so re-writing the above as

\[
p(t|x) = \int \exp \left( \log f(w) \right) \, dw
\]

in which \( f(w) = p(t|x, w)P(w) \), so we may make a second-order Taylor expansion of \( \log f(w) \) around the mode, \( \hat{w} \),

\[
\log f(w) \approx \log f(\hat{w}) + \nabla \log f(w)|_{\hat{w}} + \frac{1}{2}(w - \hat{w})^T \nabla \nabla^T \log f(w)|_{\hat{w}} (w - \hat{w})
\]

as the gradient at \( \hat{w} \) is zero be definition of the mode, so we have

\[
p(t|x) \approx f(\hat{w}) \int \exp \left( \frac{1}{2}(w - \hat{w})^T \nabla \nabla^T \log f(w)|_{\hat{w}} (w - \hat{w}) \right) \, dw
\]
Evidence & predictive distribution

- This gives evidence of the form

\[ p(t|x) \approx p(t|x, \hat{w})p(\hat{w}) \frac{(2\pi)^{d/2}}{\sqrt{\text{det} \, H}} \]

\[ L(x) = \log(p(t|x, \hat{w}) + \log p(\hat{w}) + \frac{d}{2} \log(2\pi) - \frac{1}{2} \log \text{det} \, H \]

- MAP evidence
- Complexity penalty

- Predictive distribution is Normal

\[ p(t|x) = \mathcal{N} \left( t \,|\, y(x, \hat{w}), \frac{1}{\beta} + \phi(x)^T H^{-1} \phi(x) \right) \]
Simple interpretation

\[ p(t|x, \hat{w})p(\hat{w}) \]

Gaussian with mode at \( \hat{w} \)
and inverse covariance of
\[ \nabla\nabla^T \log f(w) = H \]
the Hessian matrix
A simple example

A straight line fit! Consider a set of $N$ data points, $(x, t)$. We will model this data using a generative model of the form

$$ t = w_1 x + w_2 + \eta $$

in which $w_1, w_2$ are parameters and $\eta$ is a noise process which is assumed white (Gaussian) with a fixed variance $1/\beta$ (to make life easier! Set up a prior over $w = (w_1, w_2)^T$ as

$$ p(w) = N(0, \alpha^{-1}) $$

hence

$$ f(w) = p(t|x, w)p(w) \propto \exp\left(-\frac{\beta}{2}(t - w^T x)^2\right) \exp\left(-\frac{\alpha}{2}w^T w\right) $$
Hessian component

The Hessian matrix is the second derivative matrix w.r.t. \( w \) of the error,

\[
L(w) = \sum_n \frac{\beta}{2} \left( (t_n - w^T x_n)^2 \right) + \frac{\alpha}{2} w^T w
\]

hence

\[
H = \nabla \nabla^T L(w) = \beta \sum_n x_n x_n^T + \alpha I
\]

where \( I \) is the identity matrix.
ML-II : practical Bayes

A ‘full’ Bayesian treatment would place priors over the hyper-parameters as well as parameters – this leads to a computationally costly algorithm.

In practice the approximation based on integration over parameters with re-estimation of hyper-parameters is very useful – this is referred to as the evidence approximation, or Maximum-likelihood II inference.

Leads to a very practical algorithm especially for generalised linear models.
Evidence approximation

- Re-estimate the hyper-parameters while we integrate the parameters

\[
\frac{\partial L}{\partial \{\alpha, \beta\}} = 0
\]

\[
\begin{align*}
\hat{\alpha} &= \frac{\gamma}{\hat{w}^T \hat{w}} \\
\frac{1}{\hat{\beta}} &= \frac{1}{N - \gamma} \mathbb{E}\{y(x, \hat{w}) - t\}^2
\end{align*}
\]

Applied as a set of coupled re-estimation formulae

\[
\gamma = \sum_i \frac{\lambda_i}{\hat{\alpha} + \lambda_i}
\]

\(\lambda_i\) are eigen-values of the Hessian \(H\)
Rapidity is useful: committees even better

- The preceding result (for GLMs) only requires successive application of *generalised pseudo-inverse* and a set of re-estimation formulae; *very rapid inference*
- Can exploit this and *Jensen’s inequality* which ensures that

\[ L(D, \mathbb{E}\{w\}) \leq \mathbb{E}\{L(D, w)\} \]

- So committees are a good thing

\[
\bar{t} = \mathbb{E}\{\bar{t}_c\} = \mathbb{E}\{\frac{1}{\beta_c}\} + \mathbb{E}\{\Phi_c^T \mathbf{H}_c^{-1} \Phi_c\} + \text{var}\{\bar{t}_c\}
\]
Committees

\[ \mathbb{E}\left\{ \frac{1}{\beta_c} \right\} + \mathbb{E}\left\{ \Phi_c^T H_c^{-1} \Phi_c \right\} \]

\[ \text{var}\{\overline{t_c}\} \]
Variational Bayes
The variational bound

$$\log p(D) = \log \frac{p(D, \theta)}{p(\theta|D)}$$

$$\log p(D) = \int q(\Theta) \log \frac{q(\Theta) p(D, \Theta)}{q(\Theta) p(\Theta|D)} d\Theta$$

$$= \int q(\Theta) \log \frac{p(D, \Theta)}{q(\Theta)} d\Theta + \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|D)} d\Theta$$

$$= \boxed{F[q(\Theta)]} + \boxed{KL[q(\Theta) \| p(\Theta|D)]}$$

- ve free energy  Kullback-Liebler divergence

$q(\Theta)$ is a tractable approximation to the posterior $p(\Theta|D)$
How does that help us?

As the KL divergence is strictly $\geq 0$ so $F[q]$ is a strict *lower bound* to $\log p(\mathcal{D})$

Inference consists of maximising $F[q]$ w.r.t. $q$
Updates

$$KL(q_{new} || p)$$

$$q_{old} \rightarrow q_{new}$$

$$F(q_{old})$$

$$F(q_{new})$$

$$\log p(D)_{new}$$

$$\log p(D)_{old}$$
VB as inflating a balloon...
VB as inflating a balloon...
VB as inflating a balloon...
Expectation-Maximisation

• Consider for the moment the regression model

\[ \hat{t}(x) = y(x|w, \alpha, \beta) \]

In which \( w \) are parameters and \( \alpha, \beta \) hyperparameters

Consider the \textit{complete-data likelihood}

\[ \log p(t, w|\alpha, \beta) = \log p(t|w, \beta)p(w|\alpha) \]

\[ q(w, \alpha, \beta) = q(w)q(\alpha)q(\beta) \]
EM-GLM

1. E-step: maximise $\mathbb{E}_{q(w)}[\log p(t, w | \alpha, \beta)]$ w.r.t. $q(w)$
   This gives $w_{new} = w_{MAP}$

2. M-step: maximise $\mathbb{E}$ w.r.t. $q(\alpha), q(\beta)$
   This gives the same updates to $\alpha, \beta$ as the ML-II scheme.

Although this gives a very efficient set of updates, the posterior densities in EM are delta functions

$$q(w) = \delta(w - w^*)$$

This means that the extremisation steps are easy, as we do not have to worry about functional derivatives. But it does make for a non-Bayesian solution.
And so to VB

• Unlike standard EM, in which we re-estimate parameters at ML (or MAP) values we work with full distributions over both parameters & hyper-parameters

• This means we need to extremise w.r.t. functions (i.e. distributions) rather than point values, BUT...

• If functional forms are in exponential family then analytic solutions

• Can re-cast as message passing using sufficient statistics

• If we factor the $q$ distribution then scales excellently
  – Mean-field approximations
Advantages

- Computationally tractable
- ‘Full’ Bayesian framework
  - Explicit handling of uncertainty at all levels
  - Model comparison & selection
    - Use free energy as a proxy for true model posteriors
- Posteriors conjugate to priors
  - Only need update distribution parameters to get posteriors
  - Fits iterative procedures
- Classic techniques, such as EM, as limit cases
Summary

• With little computational overhead approximate Bayes methods
  • Provide full predictive distributions
  • Make explicit use of uncertainty, not just ‘noise’
  • Be easily formulated as iterative methods – ideal for sequential processing of data
  • As parameters are marginalised out, we can perform explicit model comparison / selection

• Non-Bayesian methods are limit cases, where distributions are collapsed to delta functions