Integrals in Statistical Modelling

- **Parameter estimation**

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \int d\mathcal{Y}\; P(\mathcal{Y}|\theta) P(\mathcal{X}|\mathcal{Y}, \theta)$$

  (or using EM)

$$\theta^{\mathsf{new}} = \operatorname*{argmax}_{\theta} \int d\mathcal{Y}\; P(\mathcal{Y}|\mathcal{X}, \theta^{\mathsf{old}}) \log P(\mathcal{X}, \mathcal{Y}|\theta)$$

- **Prediction**

$$p(x|\mathcal{D}, m) = \int d\theta\; p(\theta|\mathcal{D}, m) p(x|\theta, \mathcal{D}, m)$$

- **Model selection or weighting** (by marginal likelihood)

$$p(\mathcal{D}|m) = \int d\theta\; p(\theta|m) p(\mathcal{D}|\theta, m)$$

These integrals are often intractable:

- **Analytic intractability**: integrals may not have closed form in non-linear, non-Gaussian models $\Rightarrow$ numerical integration.
- **Computational intractability**: Numerical integral (or sum if $\mathcal{Y}$ or θ are discrete) may be exponential in data or model size.

Simple Monte Carlo Sampling

Idea: Sample from $p(x)$, average values of $F(x)$.

Simple Monte Carlo:

$$\int F(x)p(x)dx \simeq \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)}),$$

where $x^{(t)}$ are (independent) samples drawn from $p(x)$.

$\left[\text{For example: } x^{(t)} = G^{-1}(u^{(t)}) \text{ with } u \sim \text{Uniform}[0,1] \text{ and } G(x) = \int_{-\infty}^{x} p(x')dx'\right]$

Attractions:

- unbiased
- variance goes as $1/T$, independent of dimension!

Problems:

- it may be difficult or impossible to obtain the samples directly from $p(x)$
- regions of high density $p(x)$ may not correspond to regions where $F(x)$ varies a lot (thus each evaluation might have very high variance).

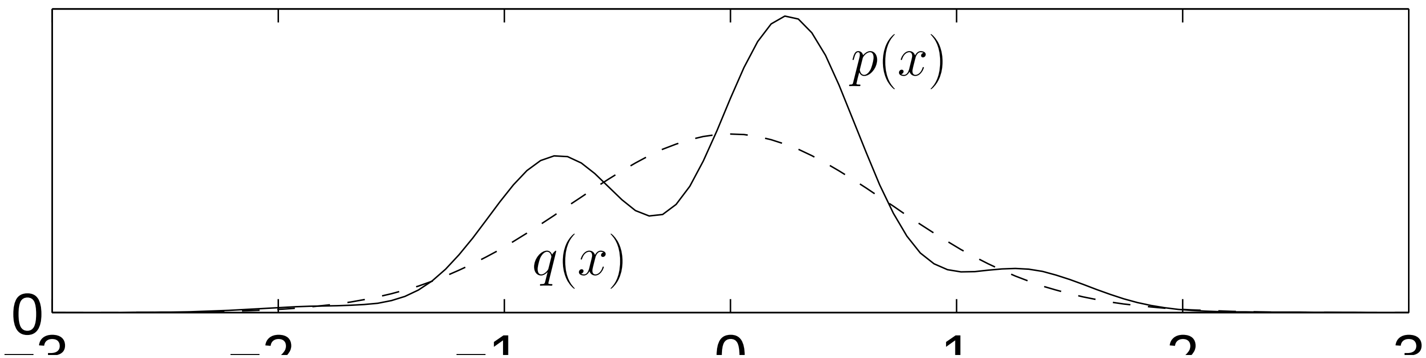
Importance Sampling

Idea: Sample from a different distribution $q(x)$ and weight those samples by $p(x)/q(x)$

Sample $x^{(t)}$ from $q(x)$:

$$\int F(x)p(x)dx = \int F(x)\frac{p(x)}{q(x)}q(x)dx \simeq \frac{1}{T}\sum_{t=1}^{T} F(x^{(t)})\frac{p(x^{(t)})}{q(x^{(t)})},$$

where $q(x)$ is non-zero wherever $p(x)$ is; weights $w^{(t)} \equiv p(x^{(t)})/q(x^{(t)})$



Attraction: unbiased; no need for upper bound (cf rejection sampling).

Problems: it may be difficult to find a suitable $q(x)$. Monte Carlo average may be dominated by few samples (high variance); or none of the high weight samples may be found!

Unnormalised densities

What if we have $f(x) \propto p(x)$, but the normaliser is unknown?

IS still works if we just normalise the weights:

$x^{(i)} \sim q$ and $w^{(i)} = f(x)/q(x) \Rightarrow$

$$\frac{\sum_i F(x^{(i)})w^{(i)}}{\sum_i w^{(i)}} \to \frac{\langle F(x)w(x)\rangle_q}{\langle w(x)\rangle_q} = \frac{\int dx\ F(x)\frac{f(x)}{q(x)}q(x)}{\int dx\ \frac{f(x)}{q(x)}q(x)} = \int dx\ F(x)\frac{f(x)}{Z_f}$$

Indeed $\sum_i w^{(i)} \to \int dx\ f(x)$ so IS provides a way to find the normaliser for f.

For example, if $f(\theta) = P(\theta)P(\mathcal{D}|\theta)$, then Z_f is the marginal likelihood or evidence for the model (sampling from f itself doesn't help us find this).

Unnormalised densities

What if we also have $g(x) \propto q(x)$ with intractable normaliser?

As long as we can sample from $g(x)/Z_g$ we can still find expectations:

$$\frac{\sum_i F(x^{(i)}) w^{(i)}}{\sum_i w^{(i)}} \rightarrow \frac{\langle F(x) w(x) \rangle_q}{\langle w(x) \rangle_q} = \frac{\displaystyle\int dx\ F(x) \frac{f(x)}{g(x)} \frac{g(x)}{Z_g}}{\displaystyle\int dx\ \frac{f(x)}{g(x)} \frac{g(x)}{Z_g}} = \int dx\ F(x) \frac{f(x)}{Z_f}$$

But now, $\displaystyle\sum_i w^{(i)} \rightarrow \frac{Z_f}{Z_g}$, so we can only recover the ratio of normalisers.

If $g(\theta) \propto P(\theta)$ and $f(\theta) = g(\theta) P(\mathcal{D}|\theta)$ [i.e., prior is non-normalised, but likelihood is a normalised conditional], then this ratio is still the evidence.

Analysis of Importance Sampling

Weights:

$$w^{(t)} \equiv \frac{p(x^{(t)})}{q(x^{(t)})}$$

Define a weighting *function* $w(x) = p(x)/q(x)$.

Importance sample is unbiased:

$$\mathsf{E}_q\left[w(x)F(x)\right] = \int q(x)w(x)F(x)dx = \int p(x)F(x)dx$$

$$\mathsf{E}_q\left[w(x)\right] = \int q(x)w(x)dx = 1$$

The weights have variance $\mathsf{Var}\left[w(x)\right] = \mathsf{E}_q\left[w(x)^2\right] - 1$, with:

$$\mathsf{E}_q\left[(w(x)^2)\right] = \int \frac{p(x)^2}{q(x)^2} q(x)dx = \int \frac{p(x)^2}{q(x)} dx$$

- How does variance effect the estimated integral?
- How does it relate to the *effective number of samples*?
- What happens if $p(x) = \mathcal{N}(0, \sigma_p^2)$ and $q(x) = \mathcal{N}(0, \sigma_q^2)$?

Improving proposals

So IS works well when the proposal density q is similar to the target f.

Idea: Move q closer using a Markov chain sampler for f.

Define the Markov chain transition probability to be $T_f(x', x)$. We can easily sample from:

$$\tilde{q}(x) = \int dx' q(x') T_f(x', x).$$

Can we use these samples for to compute importance-weighted integrals?

Unfortunately, computing the density $\tilde{q}(x)$ is intractable in general (even an unnormalised version).

Annealed Importance Sampling (AIS) adds two tricks to make this idea work.

Joint sampling

Idea 1: Consider samples of the pair:

$$(x_1, x) \sim \tilde{q}(x_1, x) = q(x_1)T_f(x_1, x)$$

We could use these as proposals for samples from $\tilde{f}(x_1, x) = f(x)T_f^{-1}(x, x_1)$, where $T_f^{-1}(x, x_1)$ is the reversed transition process satisfying

$$f(x)T_f^{-1}(x, x_1) = f(x_1)T_f(x_1, x)$$

Then, if we use weights $w^{(i)} = \dfrac{\tilde{f}(x_1^{(i)}, x^{(i)})}{\tilde{q}(x_1^{(i)}, x^{(i)})}$, we can evaluate expectations with respect to the joint. But if the function evaluated depends only on x (and not x_1), then this is the same as evaluating with respect to the marginal on x, which (by the above) is f.

BUT, this doesn't really help:

$$w = \frac{\tilde{f}(x_1, x)}{\tilde{q}(x_1, x)} = \frac{f(x)T_f^{-1}(x, x_1)}{q(x_1)T_f(x_1, x)} = \frac{f(x_1)T_f(x_1, x)}{q(x_1)T_f(x_1, x)} = \frac{f(x_1)}{q(x_1)}$$

Intermediate transitions

Idea 2: Use a Markov chain for a distribution q_1 "between" q and f.

$$\begin{aligned}(x_1, x) \sim \tilde{q}(x_1, x) &= q(x_1)T_1(x_1, x)\\ \tilde{f}(x_1, x) &= f(x)T_1^{-1}(x, x_1)\end{aligned}$$

with

$$q_1(x)T_1^{-1}(x, x_1) = q_1(x_1)T_1(x_1, x)$$

Then the weights are

$$w = \frac{\tilde{f}(x_1, x)}{\tilde{q}(x_1, x)} = \frac{f(x)T_1^{-1}(x, x_1)}{q(x_1)T_1(x_1, x)} = \frac{f(x)T_1(x_1, x)q_1(x_1)/q_1(x)}{q(x_1)T_1(x_1, x)} = \frac{f(x)}{q_1(x)}\frac{q_1(x_1)}{q(x_1)}$$

Each ratio f/q_1 and q_1/q should be better behaved than f/q because q_1 lies in between – we will analyse a specific case soon.

Annealed Importance Sampling

AIS uses a chain of n proposal distributions

$$q \longrightarrow q_{n-1} \longrightarrow q_{n-2} \longrightarrow \cdots \longrightarrow q_1$$

with MCMC transitions $T_i(x, x')$ corresponding to q_i.

A usual choice: $q_i = q^{1-\beta_i} f^{\beta_i}$ with $0 < \beta_{n-1} < \beta_{n-2} < \cdots < \beta_1 < 1$ (note unnormalised q_i).

We use this to generate a sample:

$$(x_{n-1}, x_{n-2}, \ldots, x_1, x) \sim \tilde{q} = q(x_{n-1}) T_{n-1}(x_{n-1}, x_{n-2}) \ldots T_1(x_1, x)$$

and weight relative to

$$\tilde{f} = f(x) T_1^{-1}(x, x_1) T_2^{-1}(x_1, x_2) \ldots T_{n-1}^{-1}(x_{n-2}, x_{n-1})$$

By similar algebra to before, this gives weights:

$$w(x_{n-1}, x_{n-2}, \ldots, x_1, x) = \frac{q_{n-1}(x_{n-1})}{q(x_{n-1})} \frac{q_{n-2}(x_{n-2})}{q_{n-1}(x_{n-2})} \cdots \frac{q_1(x_1)}{q_2(x_1)} \frac{f(x)}{q_1(x)}$$

Weight variance

For AIS with standard annealing schedule:

$$w(x_{n-1}, x_{n-2}, \ldots, x_1, x) = \prod_{k=1}^{n} \frac{q_{k-1}(x_{k-1})}{q_k(x_{k-1})}$$

where $q_n = q$; $q_0 = f$; $\beta_n = 0$ and $\beta_0 = 1$;

$$= \prod_{k=1}^{n} \frac{q^{1-\beta_{k-1}}(x_{k-1}) f^{\beta_{k-1}}(x_{k-1})}{q^{1-\beta_k}(x_{k-1}) f^{\beta_k}(x_{k-1})}$$

$$= \prod_{k=1}^{n} \frac{f^{\beta_{k-1}-\beta_k}(x_{k-1})}{q^{\beta_{k-1}-\beta_k}(x_{k-1})}$$

and, if the βs are evenly spaced by $1/n$

$$= \left(\prod_{k=1}^{n} \frac{f(x_{k-1})}{q(x_{k-1})} \right)^{1/n}$$

As $n \to \infty$, and provided the Markov chain "mixes" (weird, because non-stationary), this will approach log-normal with shrinking variance.

Some notes

- Trade-off between computation (Markov steps) and variance. Neal argues optimal point when Var $[\log w] = 1$.

- If T_i is properly normalised conditional, normaliser of target joint is just normaliser of f. So $\sum_i w^{(i)} \to Z_f$.

- Can extend chain using T_f. Weight (on all samples together) remains the same.

- See Neal, R. (1998). Annealed importance sampling. Technical Report 9805 (revised), Department of Statistics, University of Toronto.