Sampling and Monte Carlo Integration

Michael U. Gutmann

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

Spring Semester 2024

Recap

Learning and inference often involves intractable sums or integrals, e.g.

- Marginalisation

$$p(\mathbf{x}) = \int_{\mathbf{y}} p(\mathbf{x}, \mathbf{y}) \mathrm{d}\mathbf{y}$$

- Expectations

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

  for some function g.

- For unobserved variables, likelihood and gradient of the log lik

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

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

Recap

- For unnormalised models with intractable partition functions

$$L(\boldsymbol{\theta}) = \frac{\tilde{p}(\mathcal{D}; \boldsymbol{\theta})}{\int_{\mathbf{x}} \tilde{p}(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x}}$$

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \propto \mathbf{m}(\mathcal{D}; \boldsymbol{\theta}) - \mathbb{E}_{p(\mathbf{x};\boldsymbol{\theta})} [\mathbf{m}(\mathbf{x}; \boldsymbol{\theta})]$$

- Combined case of unnormalised models with intractable partition functions and unobserved variables.
- We have seen variational inference as an approach to deal with intractable marginalisations and likelihoods due to unobserved variables.
- Here: methods to approximate integrals and expectations using sampling.

Program

1. Monte Carlo integration
2. Sampling

Program

1. Monte Carlo integration
   - Approximating expectations by averages
   - Importance sampling
   - Effective sample size
2. Sampling

Averages with iid samples

- (From exercises): For Gaussians, the sample average is an estimate (MLE) of the mean (expectation) $\mathbb{E}[x]$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathbb{E}[x]$$

- Gaussianity not needed: assume x_i are iid observations of $x \sim p(x)$.

$$\mathbb{E}[x] = \int x p(x) \mathrm{d}x \approx \bar{x}_n \qquad \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Subscript n reminds us that we used n samples to compute the average.
- Approximating integrals by means of sample averages is called Monte Carlo integration.

Averages with iid samples

- Sample average is unbiased

$$\mathbb{E}\left[\bar{x}_n\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[x_i] \overset{*}{=} \frac{n}{n}\mathbb{E}[x] = \mathbb{E}[x]$$

  ($*$: "identically distributed" assumption is used, not independence)
- Variability

$$\mathbb{V}\left[\bar{x}_n\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} x_i\right] \overset{*}{=} \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[x_i] = \frac{1}{n}\mathbb{V}[x]$$

  ($*$: independence assumption used)
- Expected squared error decreases as $1/n$

$$\mathbb{E}\left[\left(\bar{x}_n - \mathbb{E}[x]\right)^2\right] = \mathbb{V}\left[\bar{x}_n\right] = \frac{1}{n}\mathbb{V}[x]$$

Averages with iid samples

- Weak law of large numbers:

$$\mathbb{P}\left(|\bar{x}_n - \mathbb{E}[x]| \geq \epsilon\right) \leq \frac{\mathbb{V}[x]}{n\epsilon^2}$$

- As $n \to \infty$, the probability for the sample average to deviate from the expected value goes to zero if the variance is finite.
- We say that sample average converges in probability to the expected value.
- Speed of convergence depends on the variance $\mathbb{V}[x]$.
- Different “laws of large numbers” exist that make different assumptions.

Chebyshev's inequality

- Weak law of large numbers is a direct consequence of Chebyshev's inequality
- Chebyshev's inequality: Let s be some random variable with mean $\mathbb{E}[s]$ and variance $\mathbb{V}[s]$.

$$\mathbb{P}\left(|s - \mathbb{E}[s]| \geq \epsilon\right) \leq \frac{\mathbb{V}[s]}{\epsilon^2}$$

- This means that for *all* random variables with finite mean and variance:
  - probability to deviate more than three standard deviation from the mean is less than $1/9 \approx 0.11$ (set $\epsilon = 3\sqrt{\mathbb{V}(s)}$)
  - Probability to deviate more than 6 standard deviations: $1/36 \approx 0.03$.

These are conservative values; for many distributions, the probabilities will be smaller.

Proofs (not examinable)

- Chebyshev's inequality follows from Markov's inequality.
- Markov's inequality: For a random variable $y \geq 0$

$$\mathbb{P}(y \geq t) \leq \frac{\mathbb{E}[y]}{t} \quad (t > 0)$$

- Chebyshev's inequality is obtained by setting $y = |s - \mathbb{E}[s]|$

$$\mathbb{P}\left(|s - \mathbb{E}[s]| \geq t\right) = \mathbb{P}\left((s - \mathbb{E}[s])^2 \geq t^2\right)$$
$$\leq \frac{\mathbb{E}\left[(s - \mathbb{E}[s])^2\right]}{t^2}.$$

Chebyshev's inequality follows with $t = \epsilon$, and because $\mathbb{E}[(s - \mathbb{E}[s]^2)]$ is the variance $\mathbb{V}[s]$ of s.

Proofs (not examinable)

Proof for Markov's inequality: Let t be an arbitrary positive number and y a one-dimensional non-negative random variable with pdf p.
We can decompose the expectation of y using t as split-point,

$$\mathbb{E}[y] = \int_0^\infty u p(u) \mathrm{d}u = \int_0^t u p(u) \mathrm{d}u + \int_t^\infty u p(u) \mathrm{d}u.$$

Since $u \geq t$ in the second term, we obtain the inequality

$$\mathbb{E}[y] \geq \int_0^t u p(u) \mathrm{d}u + \int_t^\infty t p(u) \mathrm{d}u.$$

The second term is t times the probability that $y \geq t$, so that

$$\begin{aligned} \mathbb{E}[y] &\geq \int_0^t u p(u) \mathrm{d}u + t\mathbb{P}(y \geq t) \\ &\geq t\mathbb{P}(y \geq t), \end{aligned}$$

where the second line holds because the first term in the first line is non-negative. This gives Markov's inequality

$$\mathbb{P}(y \geq t) \leq \frac{\mathbb{E}(y)}{t} \quad (t > 0)$$

Averages with correlated samples

- When computing the variance of the sample average

$$\mathbb{V}\left[\bar{x}_n\right] = \frac{\mathbb{V}[x]}{n}$$

  we assumed the samples are identically and independently distributed.
- The variance shrinks with increasing n and the average becomes more and more concentrated around $\mathbb{E}[x]$.
- Corresponding results exist for the case of statistically dependent samples x_i. Known as “ergodic theorems”.
- Out of scope for PMR but important for the theory of Markov chain Monte Carlo methods.

More general expectations

- So far, we have considered

$$\mathbb{E}[x] = \int x p(x) \mathrm{d}x \approx \frac{1}{n} \sum_{i=1}^{n} x_i$$

  where $x_i \sim p(x)$
- This generalises

$$\mathbb{E}[g(\mathbf{x})] = \int g(\mathbf{x}) p(\mathbf{x}) \mathrm{d}\mathbf{x} \approx \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{x}_i)$$

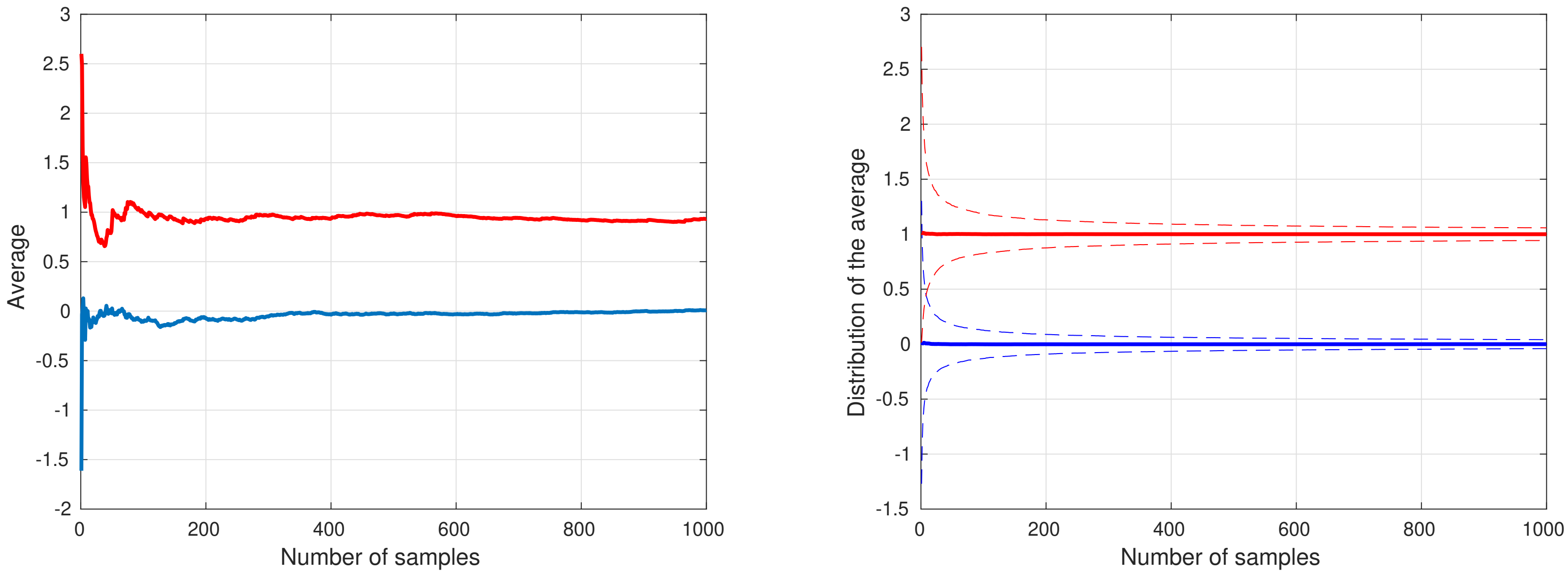
  where $\mathbf{x}_i \sim p(\mathbf{x})$
- Variance of the approximation if the $\mathbf{x}_i$ are iid is $\frac{1}{n}\mathbb{V}[g(\mathbf{x})]$

Example (Based on a slide from Amos Storkey)

$$\mathbb{E}[g(x)] = \int g(x)\mathcal{N}(x;0,1)\mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i) \qquad (x_i \sim \mathcal{N}(x;0,1))$$

for $g(x) = x$ and $g(x) = x^2$

Left: sample average as a function of n
Right: Variability (0.5 quantile: solid, 0.1 and 0.9 quantiles: dashed)



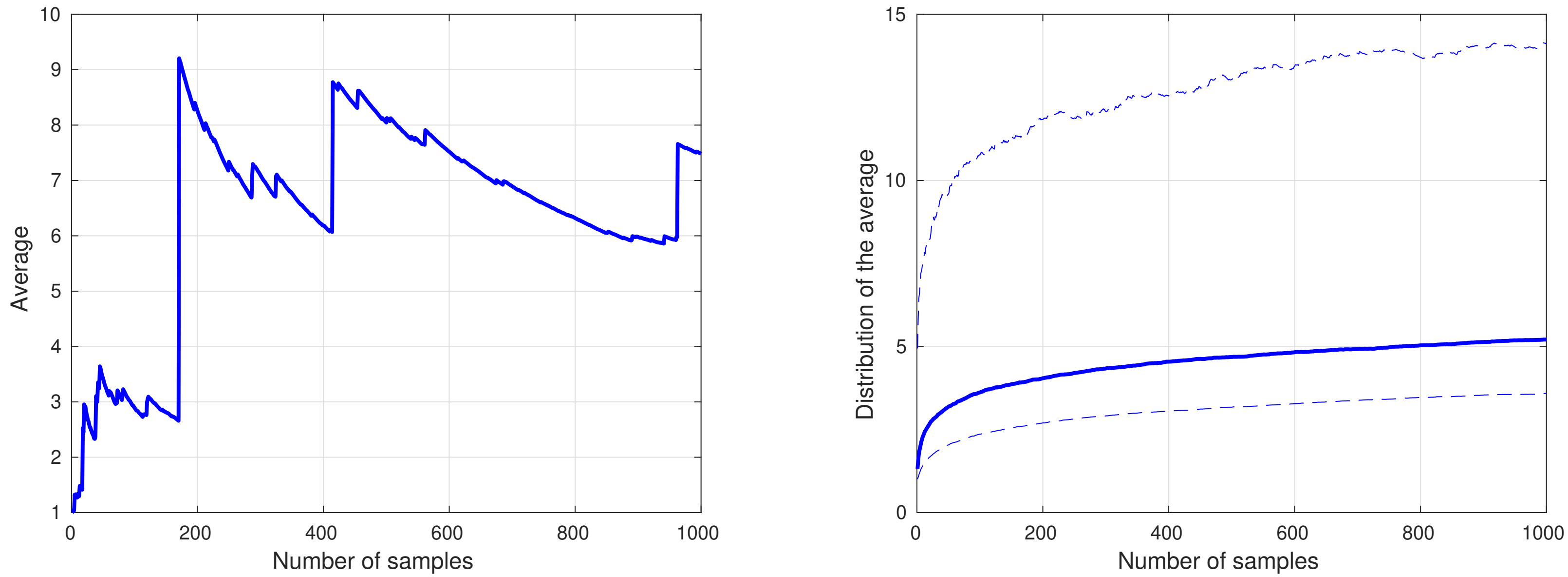
Example (Based on a slide from Amos Storkey)

$$\mathbb{E}[g(x)] = \int g(x)\mathcal{N}(x;0,1)\mathrm{d}x \approx \frac{1}{n}\sum_{i=1}^{n} g(x_i) \qquad (x_i \sim \mathcal{N}(x;0,1))$$

for $g(x) = \exp(0.6x^2)$

Left: sample average as a function of n

Right: Variability (0.5 quantile: solid, 0.1 and 0.9 quantiles: dashed)



Example

- Indicators that something is wrong:
  - Strong fluctuations in the sample average as n increases.
  - Large non-declining variability.
- Note: integral is not finite:

$$\begin{aligned}\int \exp(0.6x^2)\mathcal{N}(x;0,1)\mathrm{d}x &= \frac{1}{\sqrt{2\pi}}\int \exp(0.6x^2)\exp(-0.5x^2)\mathrm{d}x \\ &= \frac{1}{\sqrt{2\pi}}\int \exp(0.1x^2)\mathrm{d}x \\ &= \infty\end{aligned}$$

  but for any n, the sample average is finite and may be mistaken for a good approximation.
- Check variability when approximating the expected value by a sample average!

Importance sampling to approximate integrals

- If the integral does not correspond to an expectation, we can smuggle in a pdf q to rewrite it as an expected value with respect to q

$$
\begin{aligned}
I = \int g(\mathbf{x}) \mathrm{d}\mathbf{x} &= \int g(\mathbf{x}) \frac{q(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} \qquad \text{(assume } q(\mathbf{x}) > 0 \text{ when } g(\mathbf{x}) > 0\text{)} \\
&= \int \frac{g(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&= \mathbb{E}_{q(\mathbf{x})} \left[\frac{g(\mathbf{x})}{q(\mathbf{x})} \right] \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}
\end{aligned}
$$

  with $x_i \sim q(\mathbf{x})$ (iid)
- This is the basic idea of importance sampling.
- q is called the importance (or proposal) distribution

Choice of the importance distribution

- Call the approximation $\widehat{I}$,

$$\widehat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}$$

- $\widehat{I}$ is unbiased by construction

$$\mathbb{E}[\widehat{I}] = \mathbb{E}\left[\frac{g(\mathbf{x})}{q(\mathbf{x})}\right] = \int \frac{g(\mathbf{x})}{q(\mathbf{x})} q(\mathbf{x}) \mathrm{d}\mathbf{x} = \int g(\mathbf{x}) \mathrm{d}\mathbf{x} = I$$

- Variance

$$\mathbb{V}\left[\widehat{I}\right] = \frac{1}{n}\mathbb{V}\left[\frac{g(\mathbf{x})}{q(\mathbf{x})}\right] = \frac{1}{n}\mathbb{E}\left[\left(\frac{g(\mathbf{x})}{q(\mathbf{x})}\right)^2\right] - \frac{1}{n}\underbrace{\left(\mathbb{E}\left[\frac{g(\mathbf{x})}{q(\mathbf{x})}\right]\right)^2}_{I^2}$$

Depends on the second moment.

Choice of the importance distribution

- The second moment is

$$\mathbb{E}\left[\left(\frac{g(\mathbf{x})}{q(\mathbf{x})}\right)^2\right] = \int \left(\frac{g(\mathbf{x})}{q(\mathbf{x})}\right)^2 q(\mathbf{x})\mathrm{d}\mathbf{x} = \int \frac{g(\mathbf{x})^2}{q(\mathbf{x})}\mathrm{d}\mathbf{x}$$
$$= \int |g(\mathbf{x})| \frac{|g(\mathbf{x})|}{q(\mathbf{x})}\mathrm{d}\mathbf{x}$$

- Bad: $q(\mathbf{x})$ is small when $|g(\mathbf{x})|$ is large. Gives large variance.
- Good: $q(\mathbf{x})$ is large when $|g(\mathbf{x})|$ is large.
- Optimal q equals

$$q^*(\mathbf{x}) = \frac{|g(\mathbf{x})|}{\int |g(\mathbf{x})|\mathrm{d}\mathbf{x}}$$

- Optimal q cannot be computed, but justifies the heuristic that $q(\mathbf{x})$ should be large when $|g(\mathbf{x})|$ is large, or that the ratio $|g(\mathbf{x})|/q(\mathbf{x})$ should be approximately constant.

Proof (not examinable)

Since the variance of a random variable $|x|$ is non-negative and can be written as

$$\mathbb{V}[|x|] = \mathbb{E}[x^2] - (\mathbb{E}[|x|])^2,$$

we have

$$\mathbb{E}[x^2] \geq \mathbb{E}[|x|]^2$$

The smallest second moment achieves equality. We now verify that for $q^*(\mathbf{x})$, we have

$$\mathbb{E}\left[\left(\frac{g(\mathbf{x})}{q^*(\mathbf{x})}\right)^2\right] = \mathbb{E}\left[\left|\frac{g(\mathbf{x})}{q^*(\mathbf{x})}\right|\right]^2$$

Proof (not examinable)

Indeed, for the optimal q, we have

$$\mathbb{E}\left[\left(\frac{g(\mathbf{x})}{q^*(\mathbf{x})}\right)^2\right] = \int |g(\mathbf{x})| \frac{|g(\mathbf{x})|}{q^*(\mathbf{x})} \mathrm{d}\mathbf{x}$$
$$= \int |g(\mathbf{x})| \mathrm{d}\mathbf{x} \int |g(\mathbf{x})|^2 \frac{1}{|g(\mathbf{x})|} \mathrm{d}\mathbf{x}$$
$$= \left(\int |g(\mathbf{x})| \mathrm{d}\mathbf{x}\right)^2$$

and

$$\mathbb{E}\left[\left|\frac{g(\mathbf{x})}{q^*(\mathbf{x})}\right|\right]^2 = \left(\int \left|\frac{g(\mathbf{x})}{q^*(\mathbf{x})}\right| q^*(\mathbf{x}) \mathrm{d}\mathbf{x}\right)^2$$
$$= \left(\int |g(\mathbf{x})| \mathrm{d}\mathbf{x}\right)^2,$$

which concludes the proof.

Importance sampling to approximate the partition function

We can use importance sampling to approximate the partition function for unnormalised models $\tilde{p}(\mathbf{x}; \boldsymbol{\theta})$.

$$
\begin{aligned}
Z(\boldsymbol{\theta}) &= \int \tilde{p}(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \\
&= \int \tilde{p}(\mathbf{x}; \boldsymbol{\theta}) \frac{q(\mathbf{x})}{q(\mathbf{x})} \mathrm{d}\mathbf{x} && \text{(assume } q(\mathbf{x}) > 0 \text{ when } \tilde{p}(\mathbf{x}) > 0\text{)} \\
&= \int \frac{\tilde{p}(\mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{x})} q(\mathbf{x}) \mathrm{d}\mathbf{x} \\
&\approx \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(\mathbf{x}_i; \boldsymbol{\theta})}{q(\mathbf{x}_i)} && (\mathbf{x}_i \sim q(\mathbf{x}) \text{ iid})
\end{aligned}
$$

Example

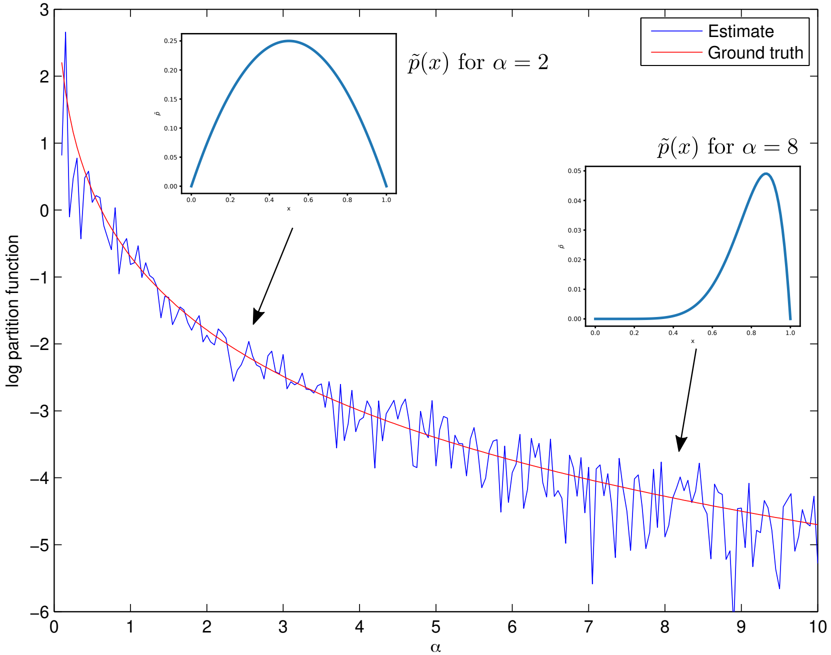
Approximating the log partition function of the unnormalised beta-distribution

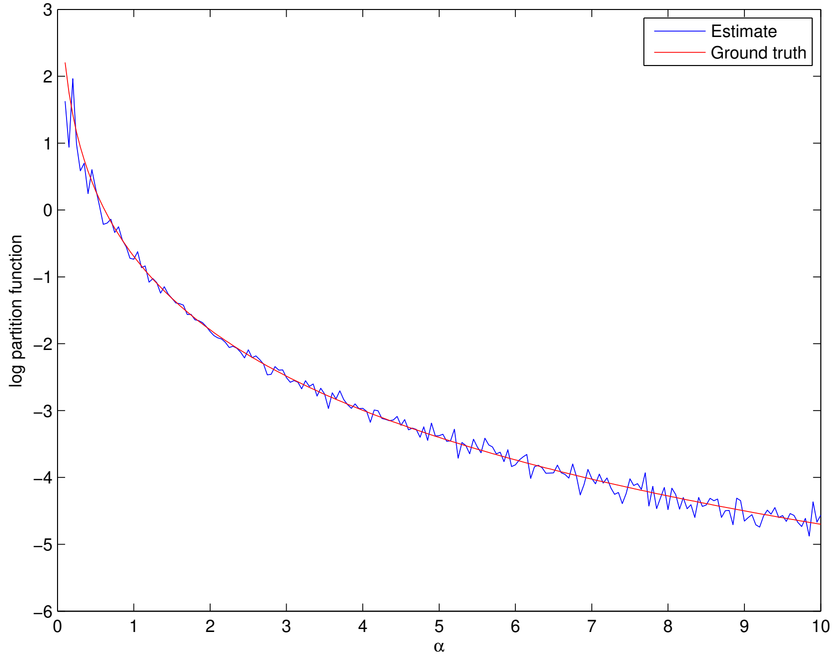
$$\tilde{p}(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}, \qquad x \in [0, 1]$$

for β fixed to $\beta = 2$.

Importance distribution: uniform distribution on $[0, 1]$

Left: $n = 10$, right: $n = 100$.





Importance sampling to approximate expectations

- Assume you would like to approximate $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$ by a sample average but sampling from $p(\mathbf{x})$ is difficult.
- We can write

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] &= \int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \\
&= \int g(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})\mathrm{d}\mathbf{x} \qquad \text{(assume } q(\mathbf{x}) > 0 \text{ when } g(\mathbf{x})p(\mathbf{x}) > 0\text{)} \\
&= \mathbb{E}_{q(\mathbf{x})}\left[g(\mathbf{x})\frac{p(\mathbf{x})}{q(\mathbf{x})}\right] \\
&\approx \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)\frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)}
\end{aligned}$$

  where $\mathbf{x}_i \sim q(\mathbf{x})$ (iid)
- The $w_i = p(\mathbf{x}_i)/q(\mathbf{x}_i)$ are called the importance weights.

Self/auto-normalised importance sampling

- We can combine the above ideas to approximate

$$\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] = \int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x}$$

by importance sampling even if we only know $\tilde{p}(\mathbf{x}) \propto p(\mathbf{x})$ and

$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{\int \tilde{p}(\mathbf{x})\mathrm{d}\mathbf{x}}$$

- Write

$$\begin{aligned}
\int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} &= \frac{\int g(\mathbf{x})\tilde{p}(\mathbf{x})\mathrm{d}\mathbf{x}}{\int \tilde{p}(\mathbf{x})\mathrm{d}\mathbf{x}} \\
&= \frac{\int g(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})\mathrm{d}\mathbf{x}}{\int \frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}q(\mathbf{x})\mathrm{d}\mathbf{x}} \\
&= \frac{\mathbb{E}_{q(\mathbf{x})}\left[g(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}\right]}{\mathbb{E}_{q(\mathbf{x})}\left[\frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}\right]}
\end{aligned}$$

Self/auto-normalised importance sampling

- Since

$$\int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{\mathbb{E}_{q(\mathbf{x})}\left[g(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}\right]}{\mathbb{E}_{q(\mathbf{x})}\left[\frac{\tilde{p}(\mathbf{x})}{q(\mathbf{x})}\right]}$$

$$= \frac{\mathbb{E}_{q(\mathbf{x})}\left[g(\mathbf{x})\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right]}{\mathbb{E}_{q(\mathbf{x})}\left[\frac{\tilde{p}(\mathbf{x})}{\tilde{q}(\mathbf{x})}\right]}$$

we only need to know the importance distribution $q(\mathbf{x})$ up to normalisation constant.

- Approximate both expectations by a sample average

$$\int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \approx \frac{\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)\frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}}{\frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}} = \frac{\sum_{i=1}^{n} g(\mathbf{x}_i)w_i}{\sum_{i=1}^{n} w_i}$$

where $w_i = \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}$ and $\mathbf{x}_i \sim q(\mathbf{x})$ (iid)

Self/auto-normalised importance sampling

$$w_i = \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}, \quad \mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$$

- Called self-normalised or auto-normalised importance sampling

$$\int g(\mathbf{x})p(\mathbf{x})\mathrm{d}\mathbf{x} \approx \sum_{i=1}^{n} W_i g(\mathbf{x}_i) \qquad W_i = \frac{w_i}{\sum_{k=1}^{n} w_k}$$

Note: $\sum_{i=1}^{n} W_i = 1$

- Interpretation in terms of a Dirac-delta approximation of $p(\mathbf{x})$,

$$p(\mathbf{x}) \approx \sum_{i=1}^{n} W_i \delta_{\mathbf{x}_i}(\mathbf{x}) \qquad \mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$$

($\equiv$ mixture of Gaussians with mixture probabilities W_i, expected values $\mathbf{x}_i$, and infinitesimally small variances)

Effective sample size

$$w_i = \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}, \quad \mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$$

- If the weights w_i are constants, the weighted average $\sum_{i=1}^{n} W_i g(\mathbf{x}_i)$ becomes a standard average

$$W_i = \frac{w_i}{\sum_{k=1}^{n} w_k} \overset{w_i=c}{==} \frac{c}{\sum_{k=1}^{n} c} = \frac{1}{N}$$

- But the w_i are typically not all equal, so that some x_i contribute more to the average than others, e.g.

$$w_1 = 10^6, w_k = 1, k > 1 \Longrightarrow W_1 \approx 1, W_k \approx 0, k > 1$$

  We would effectively "average" over 1 data point!
- When working with a weighted average, always compute the"effective sample size" (ESS),

$$\text{ESS} = \frac{\left(\sum_{i=1}^{n} w_i\right)^2}{\sum_{i=1}^{n} w_i^2} = \frac{1}{\sum_{i=1}^{n} W_i^2} \in [1, N]$$

  Small ESS means the average is unreliable (high variance).

Program

1. Monte Carlo integration
   - Approximating expectations by averages
   - Importance sampling
   - Effective sample size
2. Sampling

Program

1. Monte Carlo integration

2. Sampling
   - Simple univariate sampling
   - Rejection sampling
   - Ancestral sampling
   - Gibbs sampling

Assumption

- We assume that we are able to generate iid samples from the uniform distribution on $[0, 1]$.
- How to do that: see e.g.
  `https://statweb.stanford.edu/~owen/mc/Ch-unifrng.pdf`
  (not examinable)

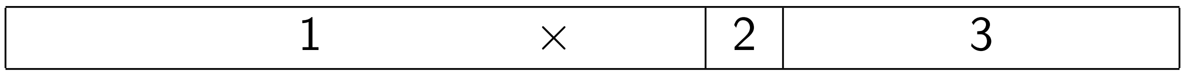
Sampling for univariate discrete random variables

(Based on a slide from David Barber)

- Consider the one dimensional discrete distribution $p(x)$ with $x \in \{1, 2, 3\}$, with

$$p(x) = \begin{cases} 0.6 & x = 1 \\ 0.1 & x = 2 \\ 0.3 & x = 3 \end{cases}$$

- Divide $[0, 1]$ into chunks $[0, 0.6)$, $[0.6, 0.7)$, $[0.7, 1]$



- We then draw a sample u uniformly from $[0, 1]$
- We return the label of the partition in which u fell.
- Example: if $u = 0.53$, we return the sample "1"

Sampling for univariate continuous random variables

- A similar method as the one above exists for continuous random variables.
- Called inverse transform sampling.
- Recall: the cumulative distribution function (cdf) of a random variable x with pdf p_x is

$$F_x(\alpha) = \mathbb{P}(x \le \alpha) = \int_{-\infty}^{\alpha} p_x(u)\mathrm{d}u$$

- To generate n iid samples from x with cdf F_x:
  - calculate the inverse F_x^{-1}
  - sample n iid random variables uniformly distributed on $[0, 1]$: $y_i \sim \mathcal{U}(0, 1)$, $i = 1, \ldots, n$.
  - transform each sample by F_x^{-1}: $x_i = F_x^{-1}(y_i)$, $i = 1, \ldots, n$.

(see exercises for derivation)

Basic principle of rejection sampling

- Assume you can draw iid samples $\mathbf{x}_i \sim q(\mathbf{x})$.
- For each sampled $\mathbf{x}_i$, you draw a Bernoulli random variable $y_i \in \{0, 1\}$ whose success probability depends on $\mathbf{x}_i$

$$\mathbb{P}(y_i = 1|\mathbf{x}_i) = f(\mathbf{x}_i)$$

- You get samples $(y_i, \mathbf{x}_i)$ with joint distribution

$$q(\mathbf{x})f(\mathbf{x})^y(1 - f(\mathbf{x}))^{(1-y)}$$

- Conditional pdf of $\mathbf{x}|y = 1$ is proportional to $q(\mathbf{x})f(\mathbf{x})$
- Keep/"accept" the $\mathbf{x}_i$ with $y_i = 1$, "reject" those with $y_i = 0$.
- Accepted samples follow

$$\mathbf{x}_i \sim \frac{q(\mathbf{x})f(\mathbf{x})}{\int q(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}}$$

- Denominator equals the marginal probability of acceptance

$$\mathbb{P}(y = 1) = \mathbb{E}_{q(\mathbf{x})}\mathbb{P}(y = 1|\mathbf{x}) = \int q(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}$$

Sampling from the posterior by rejection sampling

- Conditional acceptance probability $f(\mathbf{x}) \in [0, 1]$ can be used to shape the distribution of the samples from $q(\mathbf{x})$
- Consider Bayesian inference: prior $p(\boldsymbol{\theta})$, likelihood $L(\boldsymbol{\theta})$
- Using $L(\boldsymbol{\theta})/(\max L(\boldsymbol{\theta}))$ as acceptance probability f transforms the samples $\boldsymbol{\theta}_i$ from the prior into samples from the posterior.
- Accepted parameters follow

$$\boldsymbol{\theta}_i \sim \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})L(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} = p(\boldsymbol{\theta}|\mathcal{D})$$

- More likely parameter configurations are more likely accepted.

Sampling from the posterior by rejection sampling

- For discrete random variables $L(\boldsymbol{\theta}) = \mathbb{P}(\mathbf{x} = \mathcal{D}; \boldsymbol{\theta}) \in [0, 1]$.
- Accepting a $\boldsymbol{\theta}_i$ with probability $L(\boldsymbol{\theta})$ can be implemented by checking whether data simulated from the model with parameter value $\boldsymbol{\theta}_i$ equals the observed data.
- Samples from the posterior = samples from the prior that produce data equal to the observed one.
  (see slides "Basic of Model-Based Learning")

Side-note (not examinable): enables Bayesian inference when the likelihood is intractable (e.g. due to unobserved variables) but sampling from the model is possible. Forms the basis of a set of methods called approximate Bayesian computation, for an introductory review paper see `https://michaelgutmann.github.io/assets/papers/Lintusaari2017.pdf`.

Standard formulation of rejection sampling

- Rejection sampling is typically presented (slightly) differently.
- Goal is to generate samples from $p(\mathbf{x})$ when being able to sample from $q(\mathbf{x})$.
- Since accepted samples follow

$$\mathbf{x}_i \sim \frac{q(\mathbf{x})f(\mathbf{x})}{\int q(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x}}$$

  choose conditional acceptance probability $f(\mathbf{x}) \propto p(\mathbf{x})/q(\mathbf{x})$
- To determine the proportionality factor, note that $f(\mathbf{x})$ must be ≤ 1 since it is a conditional probability. Hence:

$$f(\mathbf{x}) = \frac{1}{M}\frac{p(\mathbf{x})}{q(\mathbf{x})} \qquad\qquad M = \max_{\mathbf{x}} \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

- Acceptance probability: $\mathbb{P}(y=1) = \int q(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{1}{M}$.

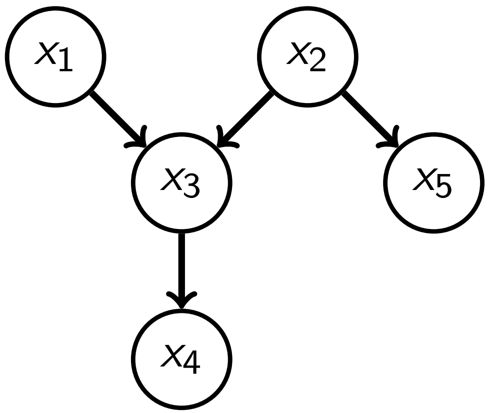
Multivariate by univariate sampling

- Rejection sampling may scale poorly because M increases with dimensionality so that acceptance probability goes down.
- Sampling from high-dimensional multivariate distributions is generally difficult.
- One way to approach the problem of multivariate sampling is to translate it into the task of solving several lower dimensional sampling problems.
- Examples:
  - Ancestral sampling
  - Gibbs sampling

Ancestral sampling

- Factorisation provides a recipe for data generation / sampling from $p(\mathbf{x})$
- Example:
  $p(x_1, \ldots, x_5) = p(x_1)p(x_2)p(x_3|x_1, x_2)p(x_4|x_3)p(x_5|x_2)$
- We can generate samples from the joint distribution $p(x_1, x_2, x_3, x_4, x_5)$ by sampling

  1. $x_1 \sim p(x_1)$
  2. $x_2 \sim p(x_2)$
  3. $x_3 \sim p(x_3|x_1, x_2)$
  4. $x_4 \sim p(x_4|x_3)$
  5. $x_5 \sim p(x_5|x_2)$



- Sets of univariate sampling problems.

Gibbs sampling

(Based on a slide from David Barber)

- Gibbs sampling also reduces the problem of multivariate sampling to the problem of univariate sampling.
- Goal: generate samples $\mathbf{x}^{(k)}$ from $p(\mathbf{x}) = p(x_1, \ldots, x_d)$.
- By product rule

$$\begin{aligned} p(\mathbf{x}) &= p(x_i | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d) \\ &= p(x_i | \mathbf{x}_{\setminus i}) p(\mathbf{x}_{\setminus i}) \end{aligned}$$

- Given a joint initial state $\mathbf{x}^{(1)}$ from which we can read off the 'parental' state $\mathbf{x}^{(1)}_{\setminus i}$

$$\mathbf{x}^{(1)}_{\setminus i} = (x^{(1)}_1, \ldots, x^{(1)}_{i-1}, x^{(1)}_{i+1}, \ldots, x^{(1)}_d),$$

  we can draw a sample $x^{(2)}_i$ from $p(x_i | \mathbf{x}^{(1)}_{\setminus i})$.
- We assume this distribution is easy to sample from since it is univariate.

Gibbs sampling

(Based on a slide from David Barber)

- We call the new joint sample in which only x_i has been updated $\mathbf{x}^{(2)}$,

$$\mathbf{x}^{(2)} = (x_1^{(1)}, \ldots, x_{i-1}^{(1)}, x_i^{(2)}, x_{i+1}^{(1)}, \ldots, x_d^{(1)}).$$

- One then selects another variable x_j to sample and, by continuing this procedure, generates a set $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ of samples in which each $\mathbf{x}^{(k+1)}$ differs from $\mathbf{x}^{(k)}$ in only a single component.
- Since $p(x_i|\mathbf{x}_{\setminus i}) = p(x_i|\mathrm{MB}(x_i))$, we can sample from $p(x_i|\mathrm{MB}(x_i))$ which is easier. ($\mathrm{MB}(x_i)$ is the Markov blanket of x_i)
- Samples are not independent.
- Gibbs sampling is an example of a Markov chain Monte Carlo method (see Barber 27.4 and 27.3.1, and the exercises, not examinable).

Program recap

1. Monte Carlo integration
   - Approximating expectations by averages
   - Importance sampling
   - Effective sample size

2. Sampling
   - Simple univariate sampling
   - Rejection sampling
   - Ancestral sampling
   - Gibbs sampling