Sampling and Monte Carlo Integration

Michael U. Gutmann

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

Autumn Semester 2026

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.

- Likelihood and gradient of the log likelihood in case of unobserved variables

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

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

Recap

- Likelihood and gradient of the log likelihood in case of unnormalised models

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

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

- 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
   - Plain Monte Carlo
   - Importance sampling for integrals
   - Importance sampling for expectations
   - Application to latent variable models
2. Sampling

Main idea

Write the quantity of interest as an expectation under a distribution we can sample from, and replace the expectation by a sample average.

Plain Monte Carlo integration

- Goal: approximate an expectation

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

- Method: draw $\mathbf{x}_1, \ldots, \mathbf{x}_n \overset{\text{iid}}{\sim} p(\mathbf{x})$ and compute the sample average

$$\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})] \approx \bar{g}_n = \frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_i)$$

  (subscript n of $\bar{g}_n$: number of samples used)
- For $g(x) = x$, the sample average $\bar{x}_n$ is the MLE of the mean of a Gaussian, but Gaussianity is not needed for $\bar{x}_n \approx \mathbb{E}[x]$.
- Approximating integrals by sample averages is called Monte Carlo integration.
- We have used this a number of times, e.g. when approximating the ELBO / its gradient.

Properties: unbiasedness and variance

Shown for $g(x) = x$; for general g, replace x_i by $g(\mathbf{x}_i)$.

- 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 $1/n$

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

Properties: law of large numbers

Shown for $g(x) = x$; for general g, replace x_i by $g(\mathbf{x}_i)$.

- Weak law of large numbers:

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

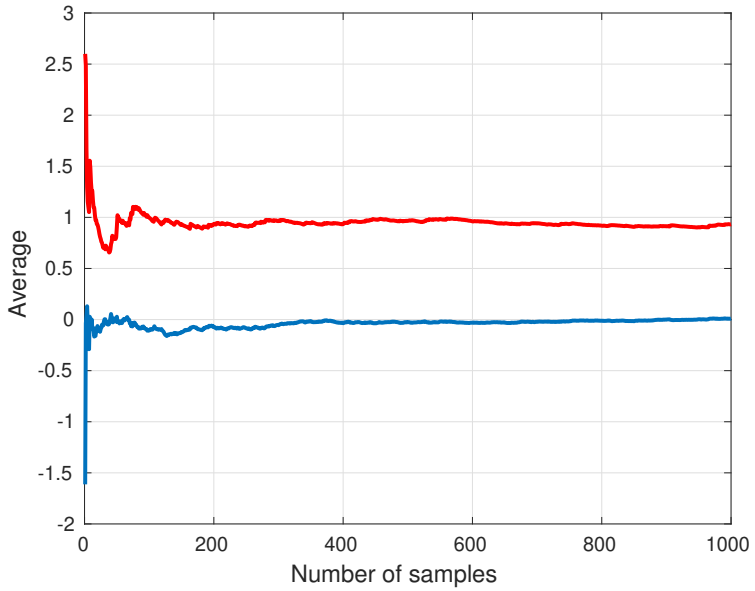
- As $n \to \infty$, the probability for the sample average to deviate from the expected value goes to zero if the variance is finite.
- We say that sample average converges in probability to the expected value.
- Speed of convergence depends on the variance $\mathbb{V}[x]$.
- Follows from Chebyshev's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

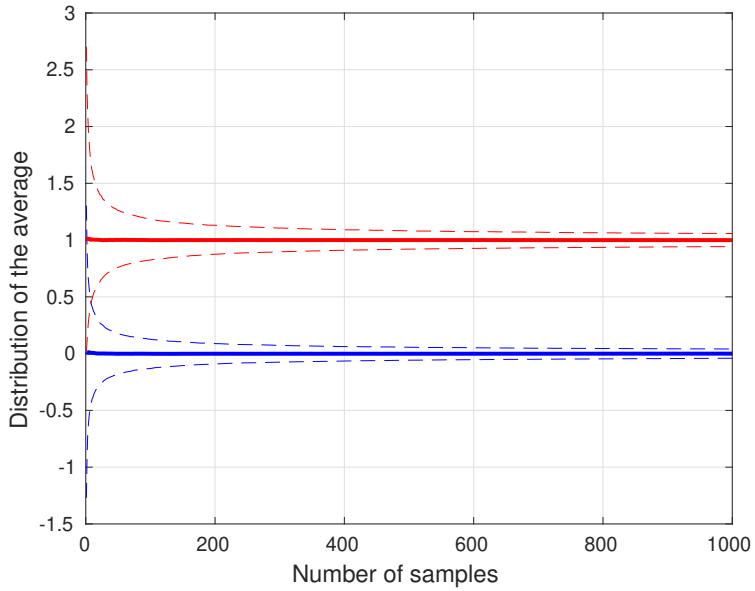
Example (Based on a slide from Amos Storkey)

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

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)



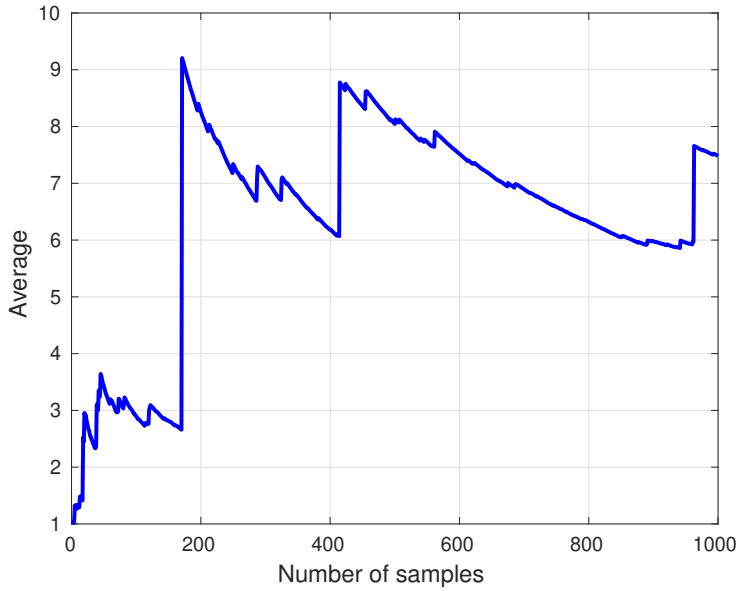


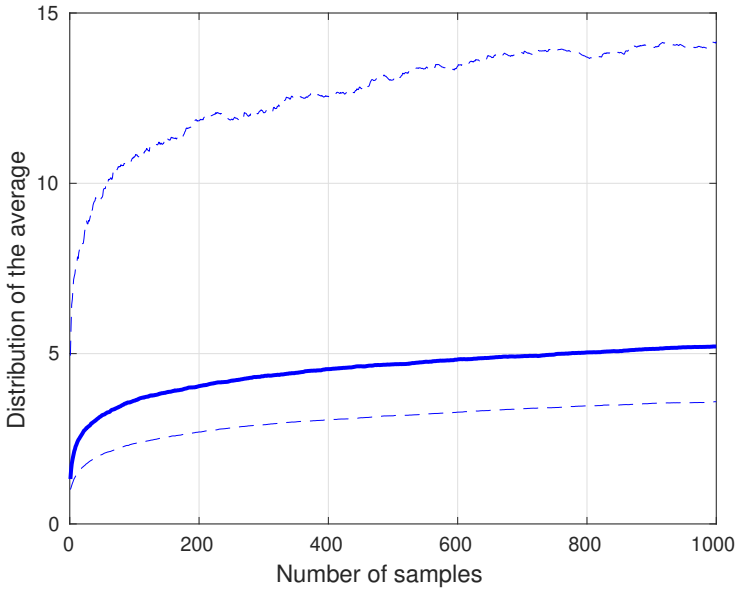
Example (Based on a slide from Amos Storkey)

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

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)





Checking the variability

- 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

- Goal: approximate an integral $I = \int g(\mathbf{x}) \mathrm{d}\mathbf{x}$ that is not an expectation, e.g. $Z(\boldsymbol{\theta})$.
- Method: smuggle in a pdf q to rewrite it as an expectation with respect to q, then use plain Monte Carlo

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

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

Choice of the importance distribution

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

$$\widehat{I}_n = \frac{1}{n}\sum_{i=1}^{n}\frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}, \quad \mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$$

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

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

- Variance

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

  Depends on the second moment.

Choice of the importance distribution

- The second moment is

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

- 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

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

(see appendix for a proof)

- 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.

Normalised weights

$$\widehat{I}_n = \frac{1}{n}\sum_i \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)},\ \mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$$

- Define the (non-negative) weights

$$w_i = \frac{|g(\mathbf{x}_i)|}{q(\mathbf{x}_i)}$$

- The contribution of $\mathbf{x}_i$ to $\widehat{I}_n$ has magnitude w_i/n. Normalised weights: relative size of the contributions,

$$W_i = \frac{w_i}{\sum_k w_k}$$

- Optimal $q = q^*$: all w_i are equal, and $W_i = 1/n$.
- If the weights are uneven, some $\mathbf{x}_i$ dominate, 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 sample!

Effective sample size

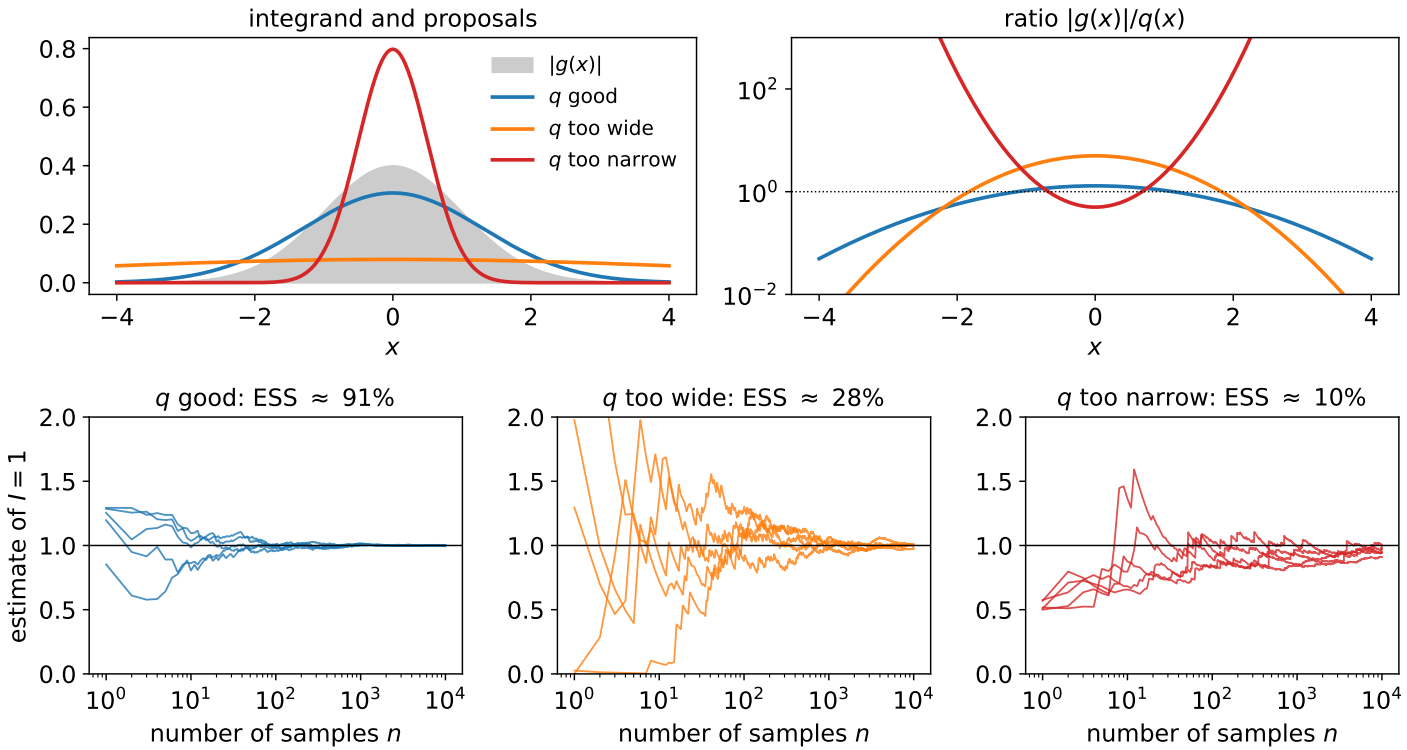
- Effective sample size (ESS) (same definition for all weights used later):

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

  $\mathrm{ESS} = n$ if and only if all weights are equal.
- Interpretation: number of samples that effectively contribute (e.g. one dominating weight $\Rightarrow$ ESS ≈ 1).
- Always compute the ESS. Small ESS means the estimate is unreliable.
- Note: if the variance is infinite, the ESS can look acceptable for moderate n (the rare, huge weights have not been sampled yet). Warning signs: ESS$/n$ keeps decreasing with n, or varies strongly when the estimate is repeated with new samples.

Good and bad importance distributions

$|g(x)| = \mathcal{N}(x; 0, 1)$, so $I = 1$; $q = \mathcal{N}(x; 0, s^2)$ with $s = 1.3$ (good), 5 (too wide), 0.5 (too narrow). (see exercises for more examples.)



- Too wide: $|g(x_i)| \approx 0$ for many $x_i \Rightarrow$ small ESS, slow convergence.
- Too narrow: $|g|/q$ explodes in the tails $\Rightarrow$ infinite variance; underestimates with occasional jumps.

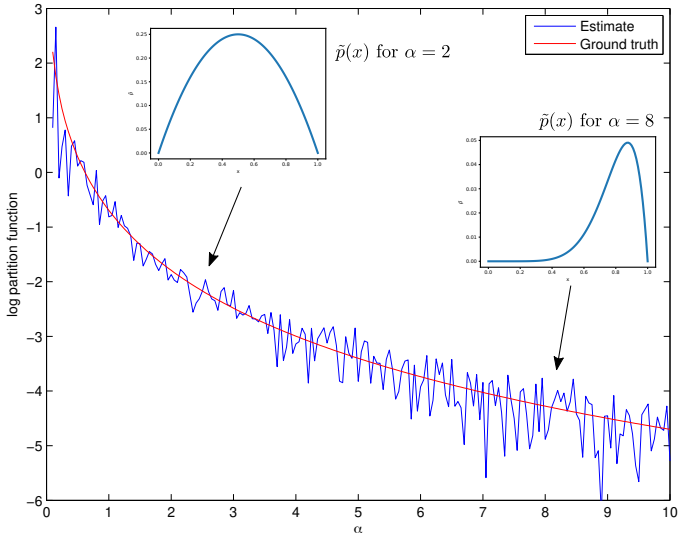
Importance sampling to approximate the partition function

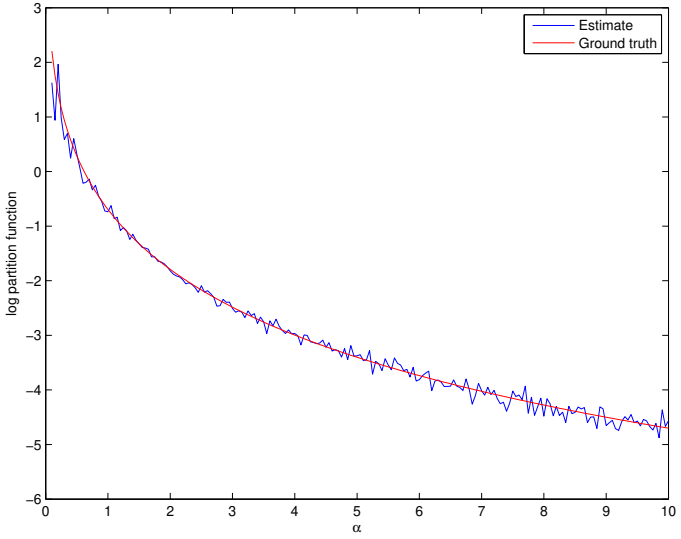
- Special case $g = \tilde{p}$ for unnormalised models $\tilde{p}(\mathbf{x}; \boldsymbol{\theta})$:

$$Z(\boldsymbol{\theta}) = \int \tilde{p}(\mathbf{x}; \boldsymbol{\theta}) \mathrm{d}\mathbf{x} = \mathbb{E}_{q(\mathbf{x})} \left[\frac{\tilde{p}(\mathbf{x}; \boldsymbol{\theta})}{q(\mathbf{x})} \right] \approx \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(\mathbf{x}_i; \boldsymbol{\theta})}{q(\mathbf{x}_i)}$$

  with $\mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$.

- Example: $\log Z(\alpha)$ for the unnormalised beta-distribution $\tilde{p}(x; \alpha, \beta) = x^{\alpha-1}(1-x)^{\beta-1}$, $x \in [0, 1]$, $\beta = 2$, with uniform q. Left: $n = 10$, right: $n = 100$.





Importance sampling for expectations

- Goal: approximate $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$ by a sample average when sampling from $p(\mathbf{x})$ is difficult but from $q(\mathbf{x})$ is possible.
- Method: as for integrals, multiply and divide by $q(\mathbf{x})$

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

  with $\mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$.
- The $w_i = p(\mathbf{x}_i)/q(\mathbf{x}_i)$ are called the importance weights.
- Example: inverse probability weighting (see "Causality"): p interventional, q observational.

Self/auto-normalised importance sampling

- Goal: approximate $\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$ when we only know $\tilde{p}(\mathbf{x}) \propto p(\mathbf{x})$, i.e.

$$p(\mathbf{x}) = \frac{\tilde{p}(\mathbf{x})}{Z}, \qquad Z = \int \tilde{p}(\mathbf{x}) \mathrm{d}\mathbf{x} \text{ intractable}$$

  e.g. posteriors and unnormalised models.
- Method: write numerator and denominator as expectations under q, as before.

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

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 its 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 $\mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$.

Self/auto-normalised importance sampling

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

- 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: $\sum_{i=1}^{n} W_i = 1$
- Reliability: compute the ESS with the importance weights $w_i = \tilde{p}(\mathbf{x}_i)/\tilde{q}(\mathbf{x}_i)$, i.e. ESS $= 1/\sum_i W_i^2$.
- Interpretation: compute expectation with a Dirac-delta approximation of $p(\mathbf{x})$,

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

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

Weights and ESS for all estimators

- Basic idea: write the quantity of interest as an expectation under a proposal $q(\mathbf{x})$ that we can sample from, and replace the expectation by a sample average over $\mathbf{x}_i \overset{\text{iid}}{\sim} q(\mathbf{x})$.

Quantity	**Estimate**	**Weights** w_i
$\int g(\mathbf{x})\mathrm{d}\mathbf{x}$	$\frac{1}{n}\sum_i \frac{g(\mathbf{x}_i)}{q(\mathbf{x}_i)}$	$\lvert g(\mathbf{x}_i)\rvert / q(\mathbf{x}_i)$
$\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$	$\frac{1}{n}\sum_i w_i g(\mathbf{x}_i)$	$p(\mathbf{x}_i)/q(\mathbf{x}_i)$
$\mathbb{E}_{p(\mathbf{x})}[g(\mathbf{x})]$, $p \propto \tilde{p}$	$\sum_i W_i g(\mathbf{x}_i)$	$\tilde{p}(\mathbf{x}_i)/\tilde{q}(\mathbf{x}_i)$

- Effective sample size, with the weights w_i from the table and $W_i = w_i / \sum_k w_k$:

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

$\mathrm{ESS} = n$ iff all weights are equal.

Latent variable models: posterior expectations

- We often need expectations under the posterior $p(\mathbf{h}|\mathbf{v})$ of the unobserved $\mathbf{h}$ given the observed $\mathbf{v}$, e.g.
  - inference: posterior mean $\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[\mathbf{h}]$
  - learning: gradient of the log-likelihood

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

- The posterior is

$$p(\mathbf{h}|\mathbf{v}) = \frac{p(\mathbf{h}, \mathbf{v})}{p(\mathbf{v})}, \qquad p(\mathbf{v}) = \int p(\mathbf{h}, \mathbf{v}) \mathrm{d}\mathbf{h}$$

  Numerator: the model, easy to evaluate. Denominator: intractable integral.

- The posterior is an unnormalised model in $\mathbf{h}$ with $\tilde{p}(\mathbf{h}) = p(\mathbf{h}, \mathbf{v})$ and $Z = p(\mathbf{v})$
  $\Rightarrow$ self-normalised importance sampling applies.

Latent variable models: posterior expectations

- Choose a proposal $q(\mathbf{h}|\mathbf{v})$ that we can sample from, e.g. the prior $p(\mathbf{h})$ or an approximation of the posterior.
- Recipe:
  1. Sample $\mathbf{h}_1, \ldots, \mathbf{h}_n \overset{\text{iid}}{\sim} q(\mathbf{h}|\mathbf{v})$.
  2. Compute the weights $w_i = \dfrac{p(\mathbf{h}_i, \mathbf{v})}{q(\mathbf{h}_i|\mathbf{v})}$ and $W_i = \dfrac{w_i}{\sum_k w_k}$.
  3. Approximate $\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[g(\mathbf{h})] \approx \sum_{i=1}^n W_i g(\mathbf{h}_i)$, and check the ESS.
- Examples:
  - posterior mean: $\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[\mathbf{h}] \approx \sum_{i=1}^n W_i \mathbf{h}_i$
  - gradient of the log-likelihood:

$$\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) \approx \sum_{i=1}^{n} W_i \nabla_{\boldsymbol{\theta}} \log p(\mathbf{h}_i, \mathbf{v}; \boldsymbol{\theta})$$

Latent variable models: marginal likelihood

- The marginal likelihood $p(\mathbf{v}) = \int p(\mathbf{v}, \mathbf{h}) \mathrm{d}\mathbf{h}$ is an integral.
- Importance sampling for integrals with $g(\mathbf{h}) = p(\mathbf{v}, \mathbf{h}) \geq 0$ and proposal $q(\mathbf{h}|\mathbf{v})$ (identity from "Variational Inference"):

$$p(\mathbf{v}) = \mathbb{E}_{q(\mathbf{h}|\mathbf{v})}\left[\frac{p(\mathbf{v}, \mathbf{h})}{q(\mathbf{h}|\mathbf{v})}\right] \approx \frac{1}{n}\sum_{i=1}^{n} w_i$$

  with $w_i = p(\mathbf{v}, \mathbf{h}_i)/q(\mathbf{h}_i|\mathbf{v})$ and $\mathbf{h}_i \overset{\text{iid}}{\sim} q(\mathbf{h}|\mathbf{v})$.
- Same weights as for posterior expectations:
  - average them: marginal likelihood $p(\mathbf{v}) \approx \frac{1}{n}\sum_i w_i$
  - normalise them: gives posterior expectations $\mathbb{E}_{p(\mathbf{h}|\mathbf{v})}[g(\mathbf{h})] \approx \sum_i W_i g(\mathbf{h}_i)$
- Use case: evaluating a trained VAE. The ELBO only gives a lower bound on $\log p(\mathbf{v})$. Importance sampling with the encoder $q_\phi(\mathbf{h}|\mathbf{v})$ as proposal estimates $p(\mathbf{v})$ and $\log p(\mathbf{v})$.

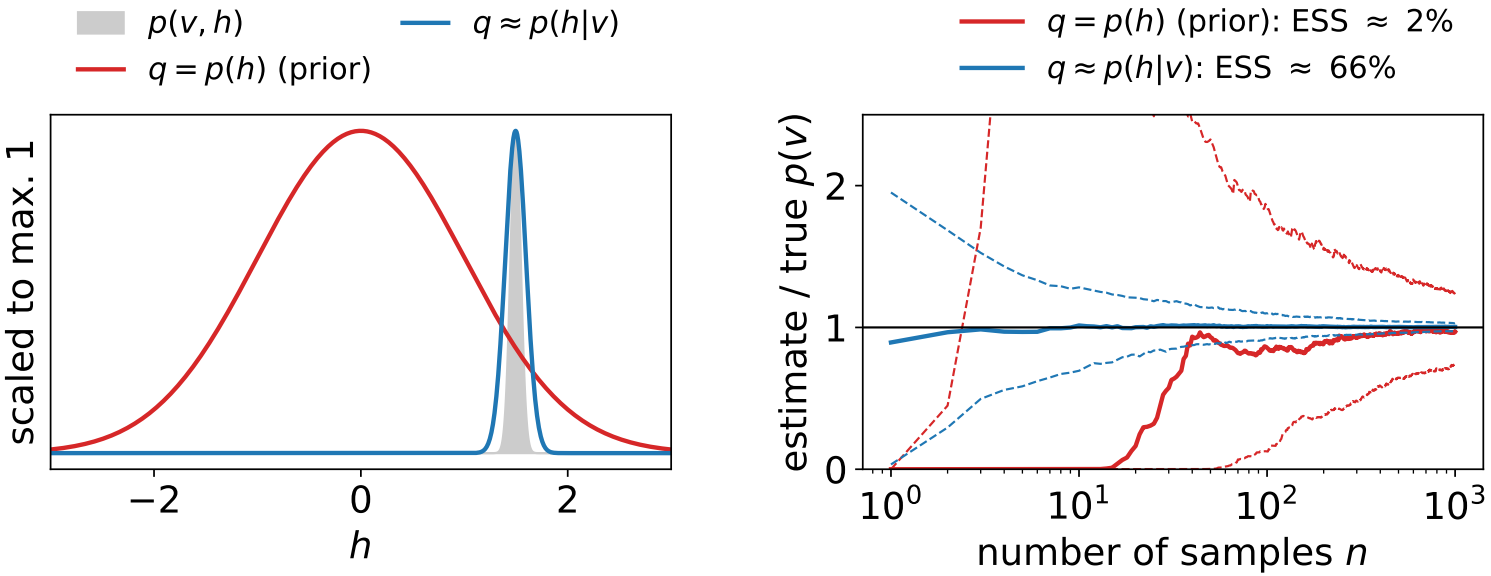
Latent variable models: choice of the proposal

$$w_i = \frac{p(\mathbf{v}, \mathbf{h}_i)}{q(\mathbf{h}_i|\mathbf{v})}, \; \mathbf{h}_i \overset{\text{iid}}{\sim} q(\mathbf{h}|\mathbf{v})$$

- Optimal proposal $q \propto p(\mathbf{v}, \mathbf{h})$, i.e. the posterior $q^*(\mathbf{h}|\mathbf{v}) = p(\mathbf{h}|\mathbf{v})$.
  - All weights are equal: $w_i = p(\mathbf{v}, \mathbf{h}_i)/p(\mathbf{h}_i|\mathbf{v}) = p(\mathbf{v})$.
  - Hence ESS $= n$.
- Prior as proposal, $q = p(\mathbf{h})$:
  - Weights $w_i = p(\mathbf{v}|\mathbf{h}_i)$.
  - Issue: few $\mathbf{h}_i$ land where $p(\mathbf{v}|\mathbf{h})$ is large, which leads to a small ESS.
  - Gets worse in high dimensions (see the overcomplete ICA example in the slides "Intractabilities").
- Use an approximation of the posterior, e.g. the variational distribution from VI or the encoder of a VAE.
- Example on the next slide.

Latent variable models: example

- Model: $h \sim \mathcal{N}(h; 0, 1)$, $v|h \sim \mathcal{N}(v; h, 0.05^2)$; observed $v = 1.5$. Exact $p(v)$ is known.
- Proposals: prior $p(h)$, and $q \approx p(h|v)$ (Gaussian with the posterior mean and twice the posterior standard deviation).
- Left: integrand $p(v, h)$ and proposals (scaled). Right: estimate relative to $p(v)$ (500 runs; 0.5 quantile: solid, 0.1 and 0.9 quantiles: dashed).



Program

1. Monte Carlo integration
   - Plain Monte Carlo
   - Importance sampling for integrals
   - Importance sampling for expectations
   - Application to latent variable models
2. Sampling

Program

1. Monte Carlo integration

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

Sampling for univariate discrete random variables

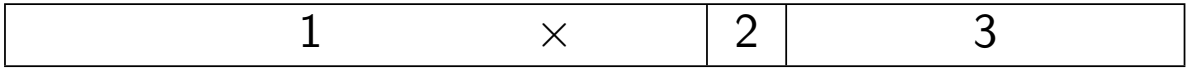
(Based on a slide from David Barber)

Assumption: we can generate iid samples $u \sim \mathcal{U}(0,1)$ (how: e.g. `https://statweb.stanford.edu/~owen/mc/Ch-unifrng.pdf`, not examinable)

- 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.
- Probability of the label equals length of the partition.

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

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

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

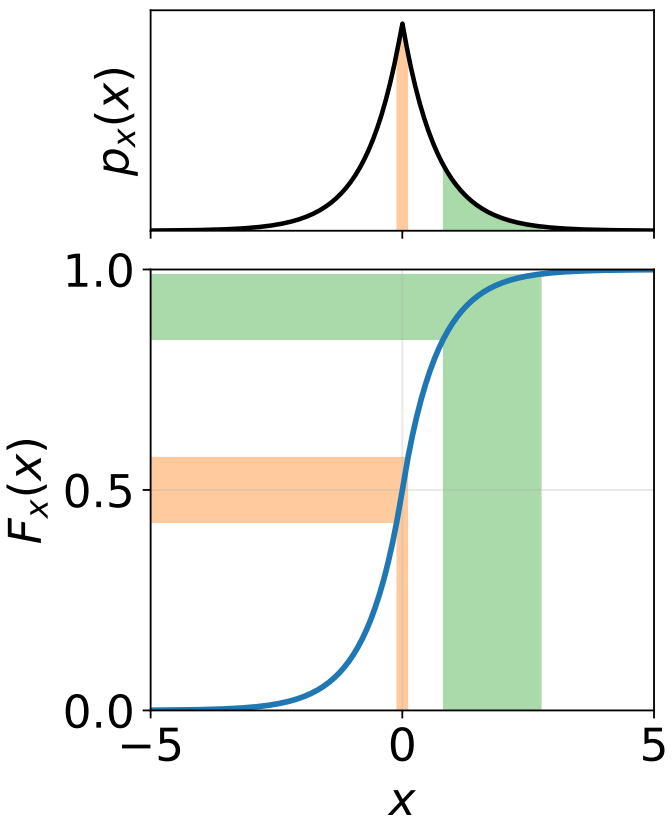
Why does it work?

- For simplicity, assume that F_x is continuous and strictly increasing, and hence invertible.
- Let $u \sim \mathcal{U}(0,1)$. The cdf of the transformed random variable ${F_x}^{-1}(u)$ is

$$\mathbb{P}({F_x}^{-1}(u) \leq \alpha) = \mathbb{P}(u \leq F_x(\alpha)) = F_x(\alpha) \tag{1}$$

  where we have used that $\mathbb{P}(u \leq \beta) = \beta$ if $u \sim \mathcal{U}(0,1)$.
- Hence for $u \sim \mathcal{U}(0,1)$, ${F_x}^{-1}(u)$ has cdf F_x, meaning ${F_x}^{-1}(u) \sim p_x$.

Inverse transform sampling: illustration



Laplace distribution (see exercises). Each band holds 15% of the u: orange → narrow x interval, green → wide x interval.

- Sampling process: draw u on the vertical axis, read off $x = F_x^{-1}(u)$ on the horizontal axis.
- Each band of $u \sim U(0, 1)$ carries the same probability mass to x:
  - steep F_x: narrow x interval ⇒ high density
  - flat F_x: wide x interval ⇒ low density
- density = mass/length = $F'_x(x)$.
- $x = F_x^{-1}(u)$ transforms u, whose distribution does not depend on the parameters of p_x: a reparameterisation (see "VAEs").
- More examples in the exercises.

Reshaping distributions

- Strategy: start from samples of an easy distribution and modify them (move, remove, or reweight) so that their distribution becomes the one we want.
- Inverse transform sampling: move the samples.
  - $x = F_x^{-1}(u)$ compresses or stretches the probability mass of $u \sim \mathcal{U}(0, 1)$.
  - Requires F_x^{-1}: mostly limited to univariate distributions with a tractable cdf.
- Rejection sampling: remove some of the samples.
  - Draw $\mathbf{x}_i$ from an easy $q(\mathbf{x})$ and keep each one with a probability $f(\mathbf{x}_i)$ that depends on where it landed.
  - Thins out the samples where q puts too much mass, so that the kept samples follow a new distribution.
  - Only requires evaluating f.
- Compare importance sampling: keep all samples and reweight them.

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 (those with $y_i = 1$) 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

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

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 "Basics 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: generate samples from $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z$, knowing only $\tilde{p}(\mathbf{x})$, 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 $f(\mathbf{x}) \propto \tilde{p}(\mathbf{x})/q(\mathbf{x})$.
- To determine the proportionality factor, note that $f(\mathbf{x})$ must be ≤ 1 since it is a conditional probability. Hence:

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

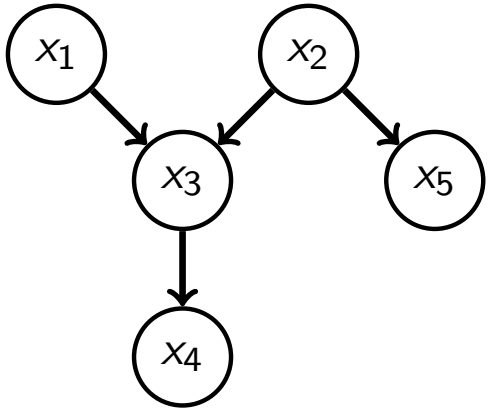
- Acceptance probability: $\mathbb{P}(y = 1) = \int q(\mathbf{x})f(\mathbf{x})\mathrm{d}\mathbf{x} = \frac{Z}{M}$

  ($= 1/M$ if normalised p and not just $\tilde{p}$ is known.)

Ancestral sampling

- High-dimensional sampling is generally difficult, e.g. the acceptance probability of rejection sampling typically decays exponentially with the dimension (see exercises).
- Ancestral sampling decomposes multivariate sampling into a sequence of univariate sampling problems (e.g. solved by inverse transform or rejection sampling).
- Directed graphical models: sample each variable given its parents, in topological order.

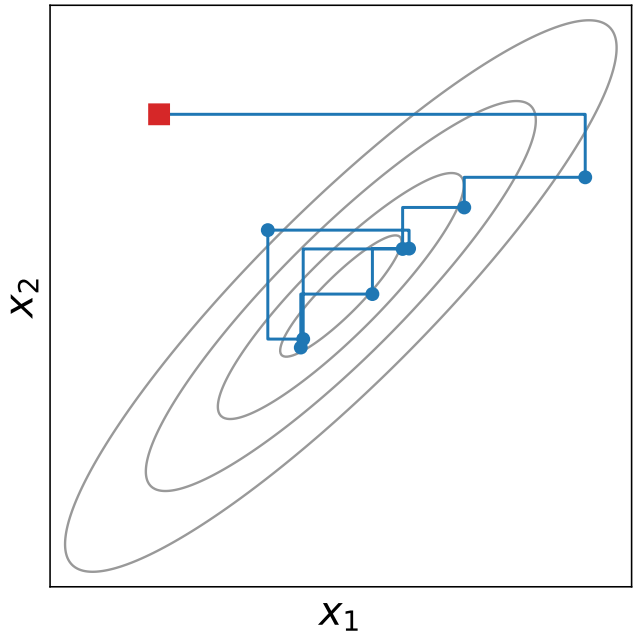
1. $x_1 \sim p(x_1)$
2. $x_2 \sim p(x_2)$
3. $x_3 \sim p(x_3|x_1, x_2)$
4. $x_4 \sim p(x_4|x_3)$
5. $x_5 \sim p(x_5|x_2)$



- Gives exact iid samples of the joint; drop variables to get marginal samples (see "Background").

Gibbs sampling

- Also decomposes multivariate sampling into univariate sampling problems.
- Uses the conditionals $p(x_i|\mathbf{x}_{\setminus i})$ of one variable given all others.
- Algorithm:
  1. Start from some state $\mathbf{x}$.
  2. Pick a variable x_i (in turn or at random).
  3. Sample $x_i \sim p(x_i|\mathbf{x}_{\setminus i})$, keeping the other variables fixed.
  4. Repeat from 2.
- Each step changes one variable only.
- Note: $p(x_i|\mathbf{x}_{\setminus i}) = p(x_i|\mathrm{MB}(x_i))$ (Markov blanket).



Correlated 2D Gaussian; start (red), alternating updates of x_1 and x_2.

Exercises: Gibbs sampling for (restricted) Boltzmann machines, where Z is intractable.

Gibbs sampling: properties

- Compared with ancestral sampling: condition on all other variables (equivalently, the Markov blanket) instead of the parents, and update repeatedly instead of in a single pass $\Rightarrow$ works more widely, including undirected and unnormalised models.
- Unnormalised models $p(\mathbf{x}) = \tilde{p}(\mathbf{x})/Z$: Z cancels in the conditionals,

$$p(x_i|\mathbf{x}_{\setminus i}) = \frac{p(\mathbf{x})}{p(\mathbf{x}_{\setminus i})} = \frac{p(\mathbf{x})}{\int p(\mathbf{x})\mathrm{d}x_i} = \frac{\tilde{p}(\mathbf{x})}{\int \tilde{p}(\mathbf{x})\mathrm{d}x_i}$$

  Requires a one-dimensional integral (sum).
- The samples form a Markov chain that converges to samples from $p(\mathbf{x})$ (proof out of scope, see e.g. Robert and Casella, 2004).
- Gibbs sampling is an example of Markov chain Monte Carlo (MCMC) (see Barber 27.3–27.4 and the exercises, not examinable).

Pros and cons of Gibbs sampling

- Pros:
  - no step-size or tuning required
  - no need for normalised models
  - can handle distributions where some variables are continuous and others discrete
  - can exploit independencies via the Markov blanket
- Cons:
  - conditionals must be sampleable
  - inner one-dimensional integral (sum) to compute
  - can mix (converge) slowly since only one dimension is changed at a time, leads to zig-zaggy sampling paths for correlated variables
- For high-dimensional or strongly correlated posteriors, use other Markov chain Monte Carlo methods, e.g. Hybrid (Hamiltonian) Monte Carlo if gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x})$ can be computed (see e.g., `https://arxiv.org/abs/1206.1901`, not examinable)

Program recap

1. Monte Carlo integration
   - Plain Monte Carlo
   - Importance sampling for integrals
   - Importance sampling for expectations
   - Application to latent variable models

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

Appendix

(not examinable)

Chebyshev's inequality

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

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

- Setting $s = \bar{x}_n$ gives the weak law of large numbers.
- 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.

Proof of Chebyshev's inequality

- 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 $y = |s - \mathbb{E}[s]|$

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

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

Proof of Markov's inequality

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,

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

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

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

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

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

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

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

Proof: optimal importance distribution

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 this is the case for $q^*(\mathbf{x})$, i.e.

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

Proof: optimal importance distribution

Indeed, for the optimal q, we have

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

and

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

which concludes the proof.