Importance Sampling & Sequential Importance Sampling

Arnaud Doucet
Departments of Statistics & Computer Science
University of British Columbia

Generic Problem

- Consider a sequence of probability distributions $\{\pi_n\}_{n\geq 1}$ defined on a sequence of (measurable) spaces $\{(E_n, \mathcal{F}_n)\}_{n\geq 1}$ where $E_1 = E$, $\mathcal{F}_1 = \mathcal{F}$ and $E_n = E_{n-1} \times E$, $\mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F}$.

Generic Problem

- Consider a sequence of probability distributions $\{\pi_n\}_{n\geq 1}$ defined on a sequence of (measurable) spaces $\{(E_n, \mathcal{F}_n)\}_{n\geq 1}$ where $E_1 = E$, $\mathcal{F}_1 = \mathcal{F}$ and $E_n = E_{n-1} \times E$, $\mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F}$.
- Each distribution $\pi_n(dx_{1:n}) = \pi_n(x_{1:n})\, dx_{1:n}$ is known *up to a normalizing constant*, i.e.

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n}$$

Generic Problem

- Consider a sequence of probability distributions $\{\pi_n\}_{n\geq 1}$ defined on a sequence of (measurable) spaces $\{(E_n, \mathcal{F}_n)\}_{n\geq 1}$ where $E_1 = E$, $\mathcal{F}_1 = \mathcal{F}$ and $E_n = E_{n-1} \times E$, $\mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F}$.
- Each distribution $\pi_n(dx_{1:n}) = \pi_n(x_{1:n})\, dx_{1:n}$ is known *up to a normalizing constant*, i.e.

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n}$$

- We want to estimate expectations of test functions $\varphi_n : E_n \rightarrow \mathbb{R}$

$$\mathbb{E}_{\pi_n}(\varphi_n) = \int \varphi_n(x_{1:n})\, \pi_n(dx_{1:n})$$

and/or the normalizing constants Z_n.

Generic Problem

- Consider a sequence of probability distributions $\{\pi_n\}_{n\geq 1}$ defined on a sequence of (measurable) spaces $\{(E_n, \mathcal{F}_n)\}_{n\geq 1}$ where $E_1 = E$, $\mathcal{F}_1 = \mathcal{F}$ and $E_n = E_{n-1} \times E$, $\mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F}$.
- Each distribution $\pi_n(dx_{1:n}) = \pi_n(x_{1:n})\, dx_{1:n}$ is known *up to a normalizing constant*, i.e.

$$\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n}$$

- We want to estimate expectations of test functions $\varphi_n : E_n \to \mathbb{R}$

$$\mathbb{E}_{\pi_n}(\varphi_n) = \int \varphi_n(x_{1:n})\, \pi_n(dx_{1:n})$$

  and/or the normalizing constants Z_n.
- We want to do this **sequentially**; i.e. first π_1 and/or Z_1 at time 1 then π_2 and/or Z_2 at time 2 and so on.

Using Monte Carlo Methods

- **Problem 1**: For most problems of interest, we cannot sample from $\pi_n(x_{1:n})$.

Using Monte Carlo Methods

- **Problem 1**: For most problems of interest, we cannot sample from $\pi_n(x_{1:n})$.
  - A standard approach to sample from high dimensional distribution consists of using iterative Markov chain Monte Carlo algorithms, this is not appropriate in our context.

Using Monte Carlo Methods

- **Problem 1**: For most problems of interest, we cannot sample from $\pi_n(x_{1:n})$.
    - A standard approach to sample from high dimensional distribution consists of using iterative Markov chain Monte Carlo algorithms, this is not appropriate in our context.
- **Problem 2:** Even if we could sample exactly from $\pi_n(x_{1:n})$, then the computational complexity of the algorithm would most likely increase with n but we typically want an algorithm of fixed computational complexity at each time step.

Using Monte Carlo Methods

- **Problem 1**: For most problems of interest, we cannot sample from $\pi_n(x_{1:n})$.
  - A standard approach to sample from high dimensional distribution consists of using iterative Markov chain Monte Carlo algorithms, this is not appropriate in our context.
- **Problem 2:** Even if we could sample exactly from $\pi_n(x_{1:n})$, then the computational complexity of the algorithm would most likely increase with n but we typically want an algorithm of fixed computational complexity at each time step.
- **Summary**: We cannot use standard MC sampling in our case and, even if we could, this would not solve our problem.

Plan of the Lectures

- Review of Importance Sampling.

Plan of the Lectures

- Review of Importance Sampling.
- Sequential Importance Sampling.

Plan of the Lectures

- Review of Importance Sampling.
- Sequential Importance Sampling.
- Applications.

Importance Sampling

- **Importance Sampling** (IS) **identity**. For any distribution q such that $\pi(x) > 0 \Rightarrow q(x) > 0$

$$\pi(x) = \frac{w(x)\,q(x)}{\int w(x)\,q(x)\,dx} \text{ where } w(x) = \frac{\gamma(x)}{q(x)}.$$

where q is called *importance distribution* and w *importance weight.*

Importance Sampling

- **Importance Sampling** (IS) **identity**. For any distribution q such that $\pi(x) > 0 \Rightarrow q(x) > 0$

$$\pi(x) = \frac{w(x)\, q(x)}{\int w(x)\, q(x)\, dx} \text{ where } w(x) = \frac{\gamma(x)}{q(x)}.$$

  where q is called *importance distribution* and w *importance weight.*

- q can be chosen arbitrarily, in particular easy to sample from

$$X^{(i)} \overset{\text{i.i.d.}}{\sim} q(\cdot) \Rightarrow \widehat{q}(dx) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}}(dx)$$

- Plugging this expression in IS identity

$$\begin{aligned}\widehat{\pi}(dx) &= \sum_{i=1}^{N} W^{(i)} \delta_{X^{(i)}}(dx) \text{ where } W^{(i)} \propto w\left(X^{(i)}\right), \\ \widehat{Z} &= \frac{1}{N} \sum_{i=1}^{N} w\left(X^{(i)}\right).\end{aligned}$$

- Plugging this expression in IS identity

$$\widehat{\pi}(dx) = \sum_{i=1}^{N} W^{(i)} \delta_{X^{(i)}}(dx) \text{ where } W^{(i)} \propto w\left(X^{(i)}\right),$$

$$\widehat{Z} = \frac{1}{N} \sum_{i=1}^{N} w\left(X^{(i)}\right).$$

- $\pi(x)$ now approximated by weighted sum of delta-masses $\Rightarrow$ Weights compensate for discrepancy between π and q.

Practical recommendations

- Select q as close to π as possible.

Practical recommendations

- Select q as close to π as possible.
- The varianec of the weights is bounded if and only if

$$\int \frac{\gamma^2(x)}{q(x)} dx < \infty.$$

Practical recommendations

- Select q as close to π as possible.
- The varianec of the weights is bounded if and only if

$$\int \frac{\gamma^2(x)}{q(x)} dx < \infty.$$

- In practice, try to ensure

$$w(x) = \frac{\gamma(x)}{q(x)} < \infty.$$

Note that in this case, rejection sampling could be used to sample from $\pi(x)$.

Example

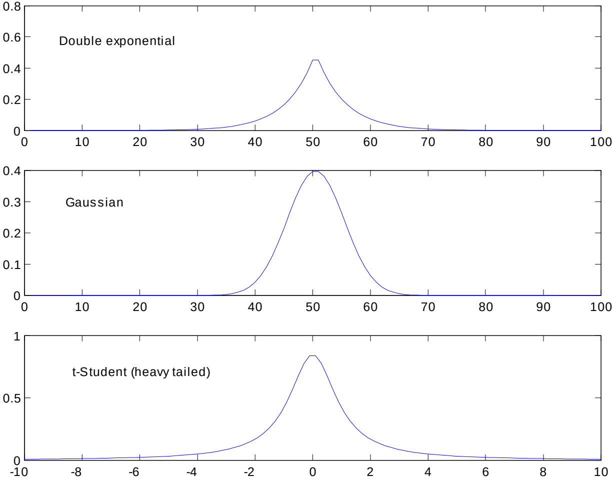

Figure: Target double exponential distributions and two IS distributions

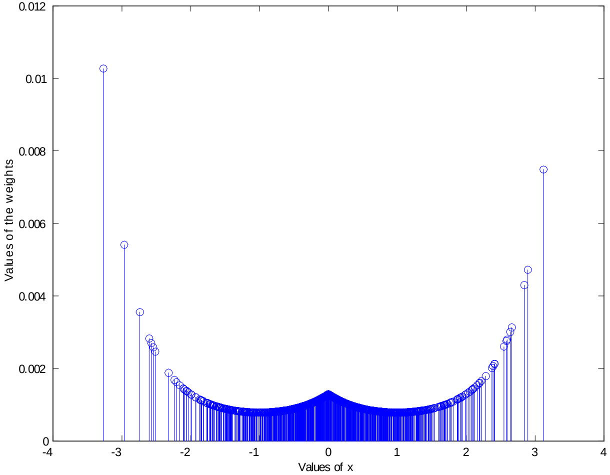

Figure: IS approximation obtained using a Gaussian IS distribution

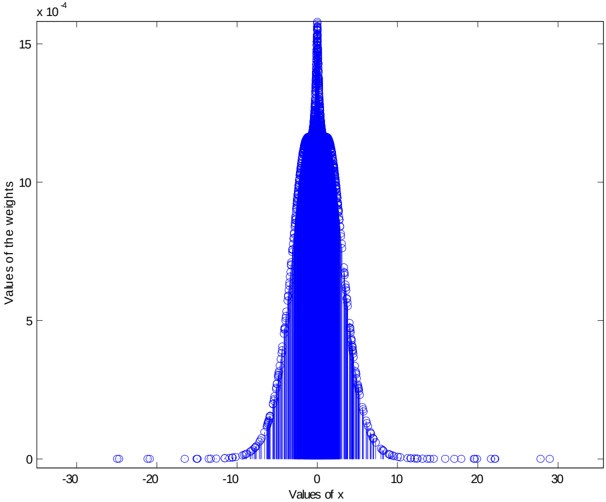

Figure: IS approximation obtained using a Student-t IS distribution

- We try to compute

$$\int \left(\frac{x}{1-x}\right)^2 \pi(x)\, dx$$

where

$$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1+\frac{x}{\nu}\right)^{-(\nu+1)/2}$$

is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0,1) / \mathcal{G}a(\nu/2, \nu/2)$) using Monte Carlo.

- We try to compute
$$\int \left(\frac{x}{1-x}\right)^2 \pi(x)\, dx$$
where
$$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)} \left(1+\frac{x}{\nu}\right)^{-(\nu+1)/2}$$
is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0,1)/\mathcal{G}a(\nu/2,\nu/2)$) using Monte Carlo.
- We use $q_1(x) = \pi(x)$, $q_2(x) = \frac{\Gamma(1)}{\sqrt{\nu\pi}\Gamma(1/2)}\left(1+\frac{x}{\nu\sigma}\right)^{-1}$ (Cauchy distribution) and $q_3(x) = \mathcal{N}\left(x; 0, \frac{\nu}{\nu-2}\right)$ (variance chosen to match the variance of $\pi(x)$)

- We try to compute

$$\int \left(\frac{x}{1-x} \right)^2 \pi(x)\, dx$$

where

$$\pi(x) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x}{\nu}\right)^{-(\nu+1)/2}$$

is a t-student distribution with $\nu > 1$ (you can sample from it by composition $\mathcal{N}(0,1)/\mathcal{G}a(\nu/2, \nu/2)$) using Monte Carlo.

- We use $q_1(x) = \pi(x)$, $q_2(x) = \frac{\Gamma(1)}{\sqrt{\nu\pi}\Gamma(1/2)}\left(1 + \frac{x}{\nu\sigma}\right)^{-1}$ (Cauchy distribution) and $q_3(x) = \mathcal{N}\left(x; 0, \frac{\nu}{\nu-2}\right)$ (variance chosen to match the variance of $\pi(x)$)
- It is easy to see that

$$\frac{\pi(x)}{q_1(x)} < \infty \text{ and } \int \frac{\pi(x)^2}{q_3(x)} dx = \infty, \ \frac{\pi(x)}{q_3(x)} \text{ is unbounded}$$

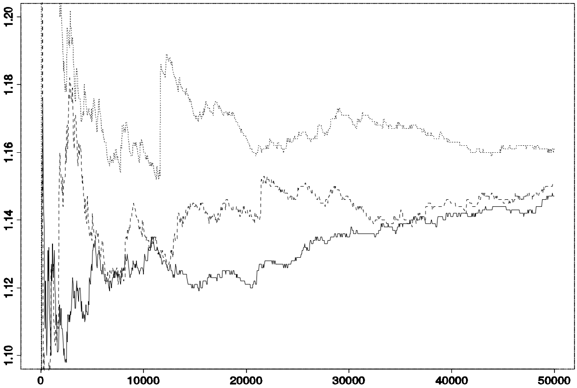

Figure: Performance for $\nu = 12$ with q_1 (solid line), q_2 (dashes) and q_3 (light dots). Final values 1.14, 1.14 and 1.16 vs true value 1.13

- We now try to compute

$$\int_{2.1}^{\infty} x^5 \pi(x)\, dx$$

- We now try to compute

$$\int_{2.1}^{\infty} x^5 \pi(x)\, dx$$

- We try to use the same importance distribution but also use the fact that using a change of variables $u = 1/x$, we have

$$\begin{aligned}\int_{2.1}^{\infty} x^5 \pi(x)\, dx &= \int_0^{1/2.1} u^{-7} \pi(1/u)\, du \\ &= \frac{1}{2.1} \int_0^{1/2.1} 2.1 u^{-7} \pi(1/u)\, du\end{aligned}$$

which is the expectation of $2.1u^{-7}\pi(1/u)$ with respect to $\mathcal{U}[0, 1/2.1]$.

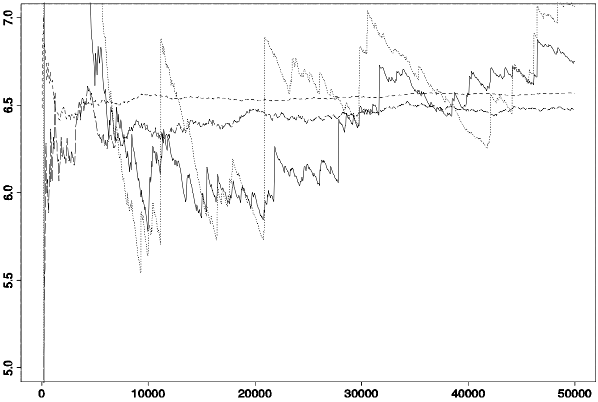

Figure: Performance for $\nu = 12$ with q_1 (solid), q_2 (short dashes), q_3 (dots), uniform (long dashes). Final values 6.75, 6.48, 7.06 and 6.48 vs true value 6.54

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi(\theta)$ and likelihood $f(x|\theta)$.

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi\left(\theta\right)$ and likelihood $f\left(\left.x\right|\theta\right)$.
- The posterior distribution is given by

$$\pi\left(\left.\theta\right|x\right)=\frac{\pi(\theta)f\left(\left.x\right|\theta\right)}{\int_{\Theta}\pi(\theta)f\left(\left.x\right|\theta\right)d\theta}\propto\gamma\left(\left.\theta\right|x\right)$$

where $\gamma\left(\left.\theta\right|x\right)=\pi\left(\theta\right)f\left(\left.x\right|\theta\right)$.

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi\left(\theta\right)$ and likelihood $f\left(x|\,\theta\right)$.
- The posterior distribution is given by

$$\pi\left(\theta|\,x\right) = \frac{\pi(\theta)f\left(x|\theta\right)}{\int_{\Theta}\pi(\theta)f\left(x|\theta\right)d\theta} \propto \gamma\left(\theta|\,x\right)$$
$$\text{where } \gamma\left(\theta|\,x\right) = \pi\left(\theta\right)f\left(x|\,\theta\right).$$

- We can use the prior distribution as a candidate distribution $q\left(\theta\right) = \pi\left(\theta\right)$.

Application to Bayesian Statistics

- Consider a Bayesian model: prior $\pi(\theta)$ and likelihood $f(x|\theta)$.
- The posterior distribution is given by

$$\pi(\theta|x) = \frac{\pi(\theta) f(x|\theta)}{\int_\Theta \pi(\theta) f(x|\theta) d\theta} \propto \gamma(\theta|x)$$

where $\gamma(\theta|x) = \pi(\theta) f(x|\theta)$.

- We can use the prior distribution as a candidate distribution $q(\theta) = \pi(\theta)$.
- We also get an estimate of the marginal likelihood

$$\int_\Theta \pi(\theta) f(x|\theta) d\theta.$$

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr\left(X_{t+1}=1\middle|\, X_t=1\right)=1-\Pr\left(X_{t+1}=2\middle|\, X_t=1\right)=p_1$ and $\Pr\left(X_{t+1}=2\middle|\, X_t=2\right)=1-\Pr\left(X_{t+1}=1\middle|\, X_t=2\right)=p_2$. Physical constraints tell us that $p_1+p_2<1$.

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr(X_{t+1}=1|X_t=1) = 1-\Pr(X_{t+1}=2|X_t=1) = p_1$ and $\Pr(X_{t+1}=2|X_t=2) = 1-\Pr(X_{t+1}=1|X_t=2) = p_2$. Physical constraints tell us that $p_1+p_2<1$.

- Assume we observe $x_1, ..., x_m$ and the prior is

$$\pi(p_1, p_2) = 2\mathbb{I}_{p_1+p_2\leq 1}$$

then the posterior is

$$\pi(p_1, p_2|x_{1:m}) \propto p_1^{m_{1,1}}(1-p_1)^{m_{1,2}}(1-p_2)^{m_{2,1}} p_2^{m_{2,2}}\mathbb{I}_{p_1+p_2\leq 1}$$

where

$$m_{i,j} = \sum_{t=1}^{m-1}\mathbb{I}_{x_t=i}\mathbb{I}_{x_{t+1}=i}$$

- *Example*: Application to Bayesian analysis of Markov chain. Consider a two state Markov chain with transition matrix F

$$\begin{pmatrix} p_1 & 1-p_1 \\ 1-p_2 & p_2 \end{pmatrix}$$

that is $\Pr(X_{t+1}=1|X_t=1) = 1 - \Pr(X_{t+1}=2|X_t=1) = p_1$ and $\Pr(X_{t+1}=2|X_t=2) = 1 - \Pr(X_{t+1}=1|X_t=2) = p_2$. Physical constraints tell us that $p_1 + p_2 < 1$.

- Assume we observe $x_1, ..., x_m$ and the prior is

$$\pi(p_1, p_2) = 2\mathbb{I}_{p_1+p_2\leq 1}$$

then the posterior is

$$\pi(p_1, p_2 | x_{1:m}) \propto p_1^{m_{1,1}} (1-p_1)^{m_{1,2}} (1-p_2)^{m_{2,1}} p_2^{m_{2,2}} \mathbb{I}_{p_1+p_2\leq 1}$$

where

$$m_{i,j} = \sum_{t=1}^{m-1} \mathbb{I}_{x_t=i} \mathbb{I}_{x_{t+1}=i}$$

- The posterior does not admit a standard expression and its normalizing constant is unknown. We can sample from it using rejection sampling.

- We are interested in estimating $\mathbb{E}\left[\left.\varphi_i\left(p_1,p_2\right)\right| x_{1:m}\right]$ for $\varphi_1\left(p_1,p_2\right)=p_1$, $\varphi_2\left(p_1,p_2\right)=p_2$, $\varphi_3\left(p_1,p_2\right)=p_1/\left(1-p_1\right)$, $\varphi_4\left(p_1,p_2\right)=p_2/\left(1-p_2\right)$ and $\varphi_5\left(p_1,p_2\right)=\log\frac{p_1(1-p_2)}{p_2(1-p_1)}$ using Importance Sampling.

- We are interested in estimating $\mathbb{E}\left[\varphi_i(p_1, p_2) \middle| x_{1:m}\right]$ for $\varphi_1(p_1, p_2) = p_1$, $\varphi_2(p_1, p_2) = p_2$, $\varphi_3(p_1, p_2) = p_1 / (1 - p_1)$, $\varphi_4(p_1, p_2) = p_2 / (1 - p_2)$ and $\varphi_5(p_1, p_2) = \log \frac{p_1(1-p_2)}{p_2(1-p_1)}$ using Importance Sampling.
- If there was no on $p_1 + p_2 < 1$ and $\pi(p_1, p_2)$ was uniform on $[0, 1] \times [0, 1]$, then the posterior would be

$$\pi_0(p_1, p_2 | x_{1:m}) = \mathcal{B}e(p_1; m_{1,1} + 1, m_{1,2} + 1) \\ \mathcal{B}e(p_2; m_{2,2} + 1, m_{2,1} + 1)$$

but this is inefficient as for the given data $(m_{1,1}, m_{1,2}, m_{2,2}, m_{2,1})$ we have $\pi_0(p_1 + p_2 < 1 | x_{1:m}) = 0.21$.

- We are interested in estimating $\mathbb{E}\left[\left.\varphi_i\left(p_1, p_2\right)\right| x_{1:m}\right]$ for $\varphi_1\left(p_1, p_2\right)=p_1$, $\varphi_2\left(p_1, p_2\right)=p_2$, $\varphi_3\left(p_1, p_2\right)=p_1/\left(1-p_1\right)$, $\varphi_4\left(p_1, p_2\right)=p_2/\left(1-p_2\right)$ and $\varphi_5\left(p_1, p_2\right)=\log\frac{p_1(1-p_2)}{p_2(1-p_1)}$ using Importance Sampling.
- If there was no on $p_1+p_2<1$ and $\pi\left(p_1, p_2\right)$ was uniform on $[0,1]\times[0,1]$, then the posterior would be

$$\begin{aligned}\pi_0\left(\left.p_1, p_2\right| x_{1:m}\right) &= \mathcal{B}e\left(p_1; m_{1,1}+1, m_{1,2}+1\right)\\ &\quad \mathcal{B}e\left(p_2; m_{2,2}+1, m_{2,1}+1\right)\end{aligned}$$

  but this is inefficient as for the given data $\left(m_{1,1}, m_{1,2}, m_{2,2}, m_{2,1}\right)$ we have $\pi_0\left(\left.p_1+p_2<1\right| x_{1:m}\right)=0.21$.
- The form of the posterior suggests using a Dirichlet distribution with density

$$\pi_1\left(\left.p_1, p_2\right| x_{1:m}\right)\propto p_1^{m_{1,1}} p_2^{m_{2,2}}\left(1-p_1-p_2\right)^{m_{1,2}+m_{2,1}}$$

  but $\pi\left(\left.p_1, p_2\right| x_{1:m}\right)/\pi_1\left(\left.p_1, p_2\right| x_{1:m}\right)$ is unbounded.

- (Geweke, 1989) proposed using the normal approximation to the binomial distribution

$$\begin{aligned}\pi_2\left(p_1, p_2 \middle| x_{1:m}\right) &\propto \exp\left(-\left(m_{1,1}+m_{1,2}\right)\left(p_1-\widehat{p}_1\right)^2 / \left(2\widehat{p}_1\left(1-\widehat{p}_1\right)\right)\right) \\ &\quad\times \exp\left(-\left(m_{2,1}+m_{2,2}\right)\left(p_2-\widehat{p}_2\right)^2 / \left(2\widehat{p}_2\left(1-\widehat{p}_2\right)\right)\right)\end{aligned}$$

where $\widehat{p}_1 = m_{1,1}/\left(m_{1,1}+m_{1,2}\right)$, $\widehat{p}_1 = m_{2,2}/\left(m_{2,2}+m_{2,1}\right)$. Then to simulate from this distribution, we simulate first $\pi_2\left(p_1 \middle| x_{1:m}\right)$ and then $\pi_2\left(p_2 \middle| x_{1:m}, p_1\right)$ which are univariate truncated Gaussian distribution which can be sampled using the inverse cdf method. The ratio $\pi\left(p_1, p_2 \middle| x_{1:m}\right)/\pi_2\left(p_1, p_2 \middle| x_{1:m}\right)$ is upper bounded.

- (Geweke, 1989) proposed using the normal approximation to the binomial distribution

$$\pi_2(p_1, p_2 | x_{1:m}) \propto \exp\left(-(m_{1,1}+m_{1,2})(p_1-\widehat{p}_1)^2 / (2\widehat{p}_1(1-\widehat{p}_1))\right) \times \exp\left(-(m_{2,1}+m_{2,2})(p_2-\widehat{p}_2)^2 / (2\widehat{p}_2(1-\widehat{p}_2))\right)$$

where $\widehat{p}_1 = m_{1,1}/(m_{1,1}+m_{1,2})$, $\widehat{p}_1 = m_{2,2}/(m_{2,2}+m_{2,1})$. Then to simulate from this distribution, we simulate first $\pi_2(p_1 | x_{1:m})$ and then $\pi_2(p_2 | x_{1:m}, p_1)$ which are univariate truncated Gaussian distribution which can be sampled using the inverse cdf method. The ratio $\pi(p_1, p_2 | x_{1:m}) / \pi_2(p_1, p_2 | x_{1:m})$ is upper bounded.
- A final one consists of using

$$\pi_3(p_1, p_2 | x_{1:m}) = \mathcal{B}e(p_1; m_{1,1}+1, m_{1,2}+1)\, \pi_3(p_2 | x_{1:m}, p_1)$$

where $\pi(p_2 | x_{1:m}, p_1) \propto (1-p_2)^{m_{2,1}} p_2^{m_{2,2}} \mathbb{I}_{p_2 \leq 1-p_1}$ is badly approximated through $\pi_3(p_2 | x_{1:m}, p_1) = \frac{2}{(1-p_1)^2} p_2 \mathbb{I}_{p_2 \leq 1-p_1}$. It is straightforward to check that $\pi(p_1, p_2 | x_{1:m}) / \pi_3(p_1, p_2 | x_{1:m}) \propto (1-p_2)^{m_{2,1}} p_2^{m_{2,2}} / \frac{2}{(1-p_1)^2} p_2 < \infty$.

- Performance for $N = 10,000$

Distribution	φ_1	φ_2	φ_3	φ_4	φ_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

- Performance for $N = 10,000$

Distribution	φ_1	φ_2	φ_3	φ_4	φ_5
π_1	0.748	0.139	3.184	0.163	2.957
π_2	0.689	0.210	2.319	0.283	2.211
π_3	0.697	0.189	2.379	0.241	2.358
π	0.697	0.189	2.373	0.240	2.358

- Sampling from π using rejection sampling works well but is computationally expensive. π_3 is computationally much cheaper whereas π_1 does extremely poorly as expected.

Effective Sample Size

- In statistics, we are usually not interested in a specific φ but in several functions and we prefer having $q(x)$ as close as possible to $\pi(x)$.

Effective Sample Size

- In statistics, we are usually not interested in a specific φ but in several functions and we prefer having $q(x)$ as close as possible to $\pi(x)$.
- For flat functions, one can approximate the variance by

$$\mathbb{V}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx (1 + \mathbb{V}_q(w(X)))\frac{\mathbb{V}_\pi(\varphi(X))}{N}.$$

Effective Sample Size

- In statistics, we are usually not interested in a specific φ but in several functions and we prefer having $q(x)$ as close as possible to $\pi(x)$.
- For flat functions, one can approximate the variance by

$$\mathbb{V}\left(\mathbb{E}_{\widehat{\pi}_N}(\varphi(X))\right) \approx (1+\mathbb{V}_q(w(X)))\,\frac{\mathbb{V}_\pi(\varphi(X))}{N}.$$

- **Simple interpretation**: The N weighted samples are approximately equivalent to M unweighted samples from π where

$$M = \frac{N}{1+\mathbb{V}_q(w(X))} \leq N.$$

Limitations of Importance Sampling

- Consider the case where the target is defined on $\mathbb{R}^n$ and

$$\pi(x_{1:n}) = \prod_{k=1}^{n} \mathcal{N}(x_k; 0, 1),$$

$$\gamma(x_{1:n}) = \prod_{k=1}^{n} \exp\left(-\frac{x_k^2}{2}\right),$$

$$Z = (2\pi)^{n/2}.$$

Limitations of Importance Sampling

- Consider the case where the target is defined on $\mathbb{R}^n$ and

$$\pi\left(x_{1:n}\right)=\prod_{k=1}^{n}\mathcal{N}\left(x_k;0,1\right),$$

$$\gamma\left(x_{1:n}\right)=\prod_{k=1}^{n}\exp\left(-\frac{x_k^2}{2}\right),$$

$$Z=\left(2\pi\right)^{n/2}.$$

- We select an importance distribution

$$q\left(x_{1:n}\right)=\prod_{k=1}^{n}\mathcal{N}\left(x_k;0,\sigma^2\right).$$

Limitations of Importance Sampling

- Consider the case where the target is defined on $\mathbb{R}^n$ and

$$\pi\left(x_{1:n}\right) = \prod_{k=1}^{n} \mathcal{N}\left(x_k; 0, 1\right),$$

$$\gamma\left(x_{1:n}\right) = \prod_{k=1}^{n} \exp\left(-\frac{x_k^2}{2}\right),$$

$$Z = \left(2\pi\right)^{n/2}.$$

- We select an importance distribution

$$q\left(x_{1:n}\right) = \prod_{k=1}^{n} \mathcal{N}\left(x_k; 0, \sigma^2\right).$$

- In this case, we have $\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right] < \infty$ only for $\sigma^2 > \frac{1}{2}$ and

$$\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right]}{Z^2} = \frac{1}{N}\left[\left(\frac{\sigma^4}{2\sigma^2 - 1}\right)^{n/2} - 1\right].$$

- The variance increases exponentially with n even in this simple case.

- The variance increases exponentially with n even in this simple case.
- For example, if we select $\sigma^2 = 1.2$ then we have a reasonably good importance distribution as $q(x_k) \approx \pi(x_k)$ but $N \frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right]}{Z^2} \approx (1.103)^{n/2}$ which is approximately equal to 1.9×10^{21} for $n = 1000$!

- The variance increases exponentially with n even in this simple case.
- For example, if we select $\sigma^2 = 1.2$ then we have a reasonably good importance distribution as $q(x_k) \approx \pi(x_k)$ but $N\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right]}{Z^2} \approx (1.103)^{n/2}$ which is approximately equal to 1.9×10^{21} for $n = 1000$!
- We would need to use $N \approx 2 \times 10^{23}$ particles to obtain a relative variance $\frac{\mathbb{V}_{\text{IS}}\left[\widehat{Z}\right]}{Z^2} = 0.01$.

Importance Sampling versus Rejection Sampling

- Given N samples from q, we estimate $\mathbb{E}_{\pi}(\varphi(X))$ through IS

$$\mathbb{E}^{\text{IS}}_{\widehat{\pi}_N}(\varphi(X)) = \frac{\sum_{i=1}^{N} w\left(X^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^{N} w\left(X^{(i)}\right)}$$

or we "filter" the samples through rejection and propose instead

$$\mathbb{E}^{\text{RS}}_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{K}\sum_{k=1}^{K}\varphi\left(X^{(i_k)}\right)$$

where $K \leq N$ is a random variable corresponding to the number of samples accepted.

Importance Sampling versus Rejection Sampling

- Given N samples from q, we estimate $\mathbb{E}_{\pi}\left(\varphi\left(X\right)\right)$ through IS

$$\mathbb{E}^{\text{IS}}_{\widehat{\pi}_N}\left(\varphi\left(X\right)\right)=\frac{\sum_{i=1}^{N} w\left(X^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^{N} w\left(X^{(i)}\right)}$$

  or we "filter" the samples through rejection and propose instead

$$\mathbb{E}^{\text{RS}}_{\widehat{\pi}_N}\left(\varphi\left(X\right)\right)=\frac{1}{K}\sum_{k=1}^{K}\varphi\left(X^{(i_k)}\right)$$

  where $K \leq N$ is a random variable corresponding to the number of samples accepted.

- We want to know which strategy performs the best.

- Define the artificial target $\overline{\pi}(x,y)$ on $E \times [0,1]$ as

$$\overline{\pi}(x,y) = \begin{cases} \frac{Cq(x)}{Z}, & \text{for } \left\{(x,y) : x \in E \text{ and } y \in \left[0, \frac{\gamma(x)}{Cq(x)}\right]\right\} \\ 0 & \text{otherwise} \end{cases}$$

then

$$\int \overline{\pi}(x,y)\, dy = \int_0^{\frac{\gamma(x)}{Cq(x)}} \frac{Cq(x)}{Z} dy = \pi(x).$$

- Define the artificial target $\overline{\pi}(x, y)$ on $E \times [0, 1]$ as

$$\overline{\pi}(x, y) = \begin{cases} \frac{Cq(x)}{Z}, & \text{for } \left\{(x, y) : x \in E \text{ and } y \in \left[0, \frac{\gamma(x)}{Cq(x)}\right]\right\} \\ 0 & \text{otherwise} \end{cases}$$

then

$$\int \overline{\pi}(x, y)\, dy = \int_0^{\frac{\gamma(x)}{Cq(x)}} \frac{Cq(x)}{Z} dy = \pi(x).$$

- Now let us consider the proposal distribution

$$q(x, y) = q(x)\, \mathcal{U}_{[0,1]}(y) \text{ for } (x, y) \in E \times [0, 1].$$

- Then rejection sampling is nothing but IS on $\mathcal{X} \times [0,1]$ where

$$w(x,y) \propto \frac{\overline{\pi}(x,y)}{q(x)\,\mathcal{U}_{[0,1]}(y)} = \begin{cases} \frac{C \int q(x)dx}{Z} & \text{for } y \in \left[0, \frac{\gamma(x)}{Cq(x)}\right] \\ 0, & \text{otherwise.} \end{cases}$$

- Then rejection sampling is nothing but IS on $\mathcal{X} \times [0,1]$ where

$$w(x,y) \propto \frac{\overline{\pi}(x,y)}{q(x)\,\mathcal{U}_{[0,1]}(y)} = \begin{cases} \frac{C \int q(x)dx}{Z} & \text{for } y \in \left[0, \frac{\gamma(x)}{Cq(x)}\right] \\ 0, & \text{otherwise.} \end{cases}$$

- We have

$$\mathbb{E}^{\mathrm{RS}}_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{K}\sum_{k=1}^{K} \varphi\left(X^{(i_k)}\right) = \frac{\sum_{i=1}^{N} w\left(X^{(i)}, Y^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^{N} w\left(X^{(i)}, Y^{(i)}\right)}.$$

- Then rejection sampling is nothing but IS on $\mathcal{X} \times [0,1]$ where

$$w(x,y) \propto \frac{\overline{\pi}(x,y)}{q(x)\,\mathcal{U}_{[0,1]}(y)} = \begin{cases} \frac{C\int q(x)dx}{Z} & \text{for } y \in \left[0, \frac{\gamma(x)}{Cq(x)}\right] \\ 0, & \text{otherwise.} \end{cases}$$

- We have

$$\mathbb{E}^{\mathrm{RS}}_{\widehat{\pi}_N}(\varphi(X)) = \frac{1}{K}\sum_{k=1}^{K}\varphi\left(X^{(i_k)}\right) = \frac{\sum_{i=1}^{N} w\left(X^{(i)}, Y^{(i)}\right)\varphi\left(X^{(i)}\right)}{\sum_{i=1}^{N} w\left(X^{(i)}, Y^{(i)}\right)}.$$

- Compared to standard IS, RS performs IS on an enlarged space.

- The variance of the importance weights from RS is higher than for standard IS:

$$\mathbb{V}\left[w\left(X,Y\right)\right] \geq \mathbb{V}\left[w\left(X\right)\right].$$

More precisely, we have

$$\begin{aligned} \mathbb{V}\left[w\left(X,Y\right)\right] &= \mathbb{V}\left[\mathbb{E}\left[w\left(X,Y\right)\middle| X\right]\right] + \mathbb{E}\left[\mathbb{V}\left[w\left(X,Y\right)\middle| X\right]\right] \\ &= \mathbb{V}\left[w\left(X\right)\right] + \mathbb{E}\left[\mathbb{V}\left[w\left(X,Y\right)\middle| X\right]\right]. \end{aligned}$$

- The variance of the importance weights from RS is higher than for standard IS:

$$\mathbb{V}\left[w\left(X,Y\right)\right]\geq\mathbb{V}\left[w\left(X\right)\right].$$

  More precisely, we have

$$\begin{aligned}\mathbb{V}\left[w\left(X,Y\right)\right] &= \mathbb{V}\left[\mathbb{E}\left[w\left(X,Y\right)\mid X\right]\right]+\mathbb{E}\left[\mathbb{V}\left[w\left(X,Y\right)\mid X\right]\right]\\ &= \mathbb{V}\left[w\left(X\right)\right]+\mathbb{E}\left[\mathbb{V}\left[w\left(X,Y\right)\mid X\right]\right].\end{aligned}$$

- To compute integrals, RS is inefficient and you should simply use IS.

Introduction to Sequential Importance Sampling

- **Aim**: Design an IS method to approximate sequentially $\{\pi_n\}_{n\geq 1}$ and to compute $\{Z_n\}_{n\geq 1}$.

- **Aim**: Design an IS method to approximate sequentially $\{\pi_n\}_{n\geq 1}$ and to compute $\{Z_n\}_{n\geq 1}$.
- At time 1, assume we have approximate $\pi_1(x_1)$ and Z_1 using an IS density $q_1(x_1)$; that is

$$\begin{aligned}
\widehat{\pi}_1(dx_1) &= \sum_{i=1}^{N} W_1^{(i)} \delta_{X_1^{(i)}}(dx) \text{ where } W_1^{(i)} \propto w_1\left(X_1^{(i)}\right), \\
\widehat{Z}_1 &= \frac{1}{N}\sum_{i=1}^{N} w_1\left(X_1^{(i)}\right)
\end{aligned}$$

with

$$w_1(x_1) = \frac{\gamma_1(x_1)}{q_1(x_1)}.$$

- At time 2, we want to approximate $\pi_2(x_{1:2})$ and Z_2 using an IS density $q_2(x_{1:2})$.

- At time 2, we want to approximate $\pi_2(x_{1:2})$ and Z_2 using an IS density $q_2(x_{1:2})$.
- We want to reuse the samples $\left\{X_1^{(i)}\right\}$ from $q_1(x_1)$ use to build the IS approximation of $\pi_1(x_1)$. This only makes sense if $\pi_2(x_1) \approx \pi_1(x_1)$.

- At time 2, we want to approximate $\pi_2(x_{1:2})$ and Z_2 using an IS density $q_2(x_{1:2})$.
- We want to reuse the samples $\left\{X_1^{(i)}\right\}$ from $q_1(x_1)$ use to build the IS approximation of $\pi_1(x_1)$. This only makes sense if $\pi_2(x_1) \approx \pi_1(x_1)$.
- We select

$$q_2(x_{1:2}) = q_1(x_1)\, q_2(x_2 | x_1)$$

so that to obtain $X_{1:2}^{(i)} \sim q_2(x_{1:2})$ we only need to sample $X_2^{(i)} \Big| X_1^{(i)} \sim q_2\left(x_2 | X_1^{(i)}\right)$.

Updating the IS approximation

- We have to compute the weights

$$
\begin{aligned}
w_2(x_{1:2}) &= \frac{\gamma_2(x_{1:2})}{q_2(x_{1:2})} = \frac{\gamma_2(x_{1:2})}{q_1(x_1)\, q_2(x_2|\, x_1)} \\
&= \frac{\gamma_1(x_1)}{q_1(x_1)} \frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2|\, x_1)} \\
&= \underbrace{w_1(x_1)}_{\text{previous weight}} \underbrace{\frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2|\, x_1)}}_{\text{incremental weigh}}
\end{aligned}
$$

Updating the IS approximation

- We have to compute the weights

$$
\begin{aligned}
w_2(x_{1:2}) &= \frac{\gamma_2(x_{1:2})}{q_2(x_{1:2})} = \frac{\gamma_2(x_{1:2})}{q_1(x_1)\, q_2(x_2 | x_1)} \\
&= \frac{\gamma_1(x_1)}{q_1(x_1)} \frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2 | x_1)} \\
&= \underbrace{w_1(x_1)}_{\text{previous weight}} \underbrace{\frac{\gamma_2(x_{1:2})}{\gamma_1(x_1)\, q_2(x_2 | x_1)}}_{\text{incremental weigh}}
\end{aligned}
$$

- For the normalized weights

$$
W_2^{(i)} \propto W_1^{(i)} \frac{\gamma_2\left(X_{1:2}^{(i)}\right)}{\gamma_1\left(X_1^{(i)}\right) q_2\left(X_2^{(i)} \middle| X_1^{(i)}\right)}
$$

- Generally speaking, we use at time n

$$\begin{aligned} q_n(x_{1:n}) &= q_{n-1}(x_{1:n-1})\, q_n(x_n|\, x_{1:n-1}) \\ &= q_1(x_1)\, q_2(x_2|\, x_1) \cdots q_n(x_n|\, x_{1:n-1}) \end{aligned}$$

so if $X_{1:n-1}^{(i)} \sim q_{n-1}(x_{1:n-1})$ then we only need to sample $X_n^{(i)} \Big|\, X_{n-1}^{(i)} \sim q_n\left(x_n|\, X_{1:n-1}^{(i)}\right)$.

- Generally speaking, we use at time n

$$\begin{aligned} q_n(x_{1:n}) &= q_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1}) \\ &= q_1(x_1)\, q_2(x_2 | x_1) \cdots q_n(x_n | x_{1:n-1}) \end{aligned}$$

so if $X_{1:n-1}^{(i)} \sim q_{n-1}(x_{1:n-1})$ then we only need to sample $X_n^{(i)} \Big| X_{n-1}^{(i)} \sim q_n\left(x_n | X_{1:n-1}^{(i)}\right)$.

- The importance weights are updated according to

$$w_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{q_n(x_{1:n})} = w_{n-1}(x_{1:n-1}) \frac{\gamma_n(x_{1:n})}{\gamma_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1})}$$

Sequential Importance Sampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.

Sequential Importance Sampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $n \geq 2$

Sequential Importance Sampling

- At time $n=1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q_n\left(\cdot \mid X_{1:n-1}^{(i)}\right)$

Sequential Importance Sampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q_n\left(\cdot \middle| X_{1:n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = w_{n-1}\left(X_{1:n-1}^{(i)}\right) \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{\gamma_{n-1}\left(X_{1:n-1}^{(i)}\right) q_n\left(X_n^{(i)} \middle| X_{1:n-1}^{(i)}\right)}$.

Sequential Importance Sampling

- At time $n = 1$, sample $X_1^{(i)} \sim q_1(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = \frac{\gamma_1\left(X_1^{(i)}\right)}{q_1\left(X_1^{(i)}\right)}$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim q_n\left(\cdot \middle| X_{1:n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = w_{n-1}\left(X_{1:n-1}^{(i)}\right) \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{\gamma_{n-1}\left(X_{1:n-1}^{(i)}\right) q_n\left(X_n^{(i)} \middle| X_{1:n-1}^{(i)}\right)}$.
- At any time n, we have

$$X_{1:n}^{(i)} \sim q_n(x_{1:n}), \; w_n\left(X_{1:n}^{(i)}\right) = \frac{\gamma_n\left(X_{1:n}^{(i)}\right)}{q_n\left(X_{1:n}^{(i)}\right)}$$

thus we can obtain easily an IS approximation of $\pi_n(x_{1:n})$ and of Z_n.

Sequential Importance Sampling for State-Space Models

- State-space models

$$\text{Hidden Markov process: } X_1 \sim \mu, \;\; X_k | (X_{k-1} = x_{k-1}) \sim f(\cdot | x_{k-1})$$

$$\text{Observation process: } \; Y_k | (X_k = x_k) \sim g(\cdot | x_k)$$

Sequential Importance Sampling for State-Space Models

- State-space models

$$\text{Hidden Markov process: } X_1 \sim \mu, \;\; X_k | (X_{k-1} = x_{k-1}) \sim f(\cdot | x_{k-1})$$

$$\text{Observation process: } \; Y_k | (X_k = x_k) \sim g(\cdot | x_k)$$

- Assume we receive $y_{1:n}$, we are interested in sampling from

$$\pi_n(x_{1:n}) = p(x_{1:n} | y_{1:n}) = \frac{p(x_{1:n}, y_{1:n})}{p(y_{1:n})}$$

and estimating $p(y_{1:n})$ where

$$\gamma_n(x_{1:n}) = p(x_{1:n}, y_{1:n}) = \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \prod_{k=1}^{n} g(y_k | x_k),$$

$$Z_n = p(y_{1:n}) = \int \cdots \int \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \prod_{k=1}^{n} g(y_k | x_k) \, dx_{1:n}.$$

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.
- At time $n = 1$, sample $X_1^{(i)} \sim \mu(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = g\left(y_1 | X_1^{(i)}\right)$.

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.
- At time $n = 1$, sample $X_1^{(i)} \sim \mu(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = g\left(y_1 | X_1^{(i)}\right)$.
- At time $n \geq 2$

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.
- At time $n = 1$, sample $X_1^{(i)} \sim \mu(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = g\left(y_1 | X_1^{(i)}\right)$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim f\left(\cdot | X_{1:n-1}^{(i)}\right)$

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.
- At time $n = 1$, sample $X_1^{(i)} \sim \mu(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = g\left(y_1 | X_1^{(i)}\right)$.
- At time $n \geq 2$
  - sample $X_n^{(i)} \sim f\left(\cdot | X_{1:n-1}^{(i)}\right)$
  - compute $w_n\left(X_{1:n}^{(i)}\right) = w_{n-1}\left(X_{1:n-1}^{(i)}\right) g\left(y_n | X_n^{(i)}\right)$.

- We can select $q_1(x_1) = \mu(x_1)$ and $q_n(x_n | x_{1:n-1}) = q_n(x_n | x_{n-1}) = f(x_n | x_{n-1})$.
- At time $n = 1$, sample $X_1^{(i)} \sim \mu(\cdot)$ and set $w_1\left(X_1^{(i)}\right) = g\left(y_1 | X_1^{(i)}\right)$.
- At time $n \geq 2$
    - sample $X_n^{(i)} \sim f\left(\cdot | X_{1:n-1}^{(i)}\right)$
    - compute $w_n\left(X_{1:n}^{(i)}\right) = w_{n-1}\left(X_{1:n-1}^{(i)}\right) g\left(y_n | X_n^{(i)}\right)$.
- At any time n, we have

$$X_{1:n}^{(i)} \sim \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}), \quad w_n\left(X_{1:n}^{(i)}\right) = \prod_{k=1}^{n} g\left(y_k | X_k^{(i)}\right)$$

thus we can obtain easily an IS approximation of $p(x_{1:n} | y_{1:n})$ and of $p(y_{1:n})$.

Application to Stochastic Volatility Model

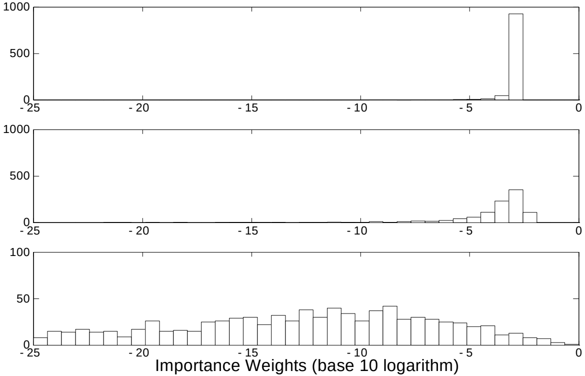

Figure: Histograms of the base 10 logarithm of $W_n^{(i)}$ for $n = 1$ (top), $n = 50$ (middle) and $n = 100$ (bottom).

- The algorithm performance collapse as n increases... After a few time steps, only a very small number of particles have non negligible weights.

Structure of the Optimal Distribution

- The optimal zero-variance density *at time n* is simply given by

$$q_n(x_{1:n}) = \pi_n(x_{1:n}).$$

Structure of the Optimal Distribution

- The optimal zero-variance density *at time n* is simply given by

$$q_n(x_{1:n}) = \pi_n(x_{1:n}).$$

- As we have

$$\pi_n(x_{1:n}) = \pi_n(x_1)\,\pi_n(x_2|\,x_1)\cdots\pi_n(x_n|\,x_{1:n-1}),$$

where $\pi_n(x_k|\,x_{1:k-1}) \propto \gamma_n(x_k|\,x_{1:k-1})$ it means that we have

$$q_k^{\text{opt}}(x_k|\,x_{1:k-1}) = \pi_n(x_k|\,x_{1:k-1}).$$

Structure of the Optimal Distribution

- The optimal zero-variance density *at time n* is simply given by

$$q_n(x_{1:n}) = \pi_n(x_{1:n}).$$

- As we have

$$\pi_n(x_{1:n}) = \pi_n(x_1)\,\pi_n(x_2|\,x_1)\cdots\pi_n(x_n|\,x_{1:n-1}),$$

where $\pi_n(x_k|\,x_{1:k-1}) \propto \gamma_n(x_k|\,x_{1:k-1})$ it means that we have

$$q_k^{\text{opt}}(x_k|\,x_{1:k-1}) = \pi_n(x_k|\,x_{1:k-1}).$$

- Obviously this result does depend on n so it is only useful if we are only interested in a specific target $\pi_n(x_{1:n})$ and in such scenarios we need to typically approximate $\pi_n(x_k|\,x_{1:k-1})$.

Locally Optimal Importance Distribution

- One sensible strategy consists of selecting $q_n\left(x_n \middle| x_{1:n-1}\right)$ at time n so as to minimize the variance of the importance weights.

Locally Optimal Importance Distribution

- One sensible strategy consists of selecting $q_n(x_n | x_{1:n-1})$ at time n so as to minimize the variance of the importance weights.
- We have for the importance weight

$$
\begin{aligned}
w_n(x_{1:n}) &= \frac{\gamma_n(x_{1:n})}{q_{n-1}(x_{1:n-1})\, q_n(x_n | x_{1:n-1})} \\
&= \frac{Z_n \pi_n(x_{1:n-1})}{q_{n-1}(x_{1:n-1})} \frac{\pi_n(x_n | x_{1:n-1})}{q_n(x_n | x_{1:n-1})}
\end{aligned}
$$

Locally Optimal Importance Distribution

- One sensible strategy consists of selecting $q_n\left(x_n \mid x_{1:n-1}\right)$ at time n so as to minimize the variance of the importance weights.
- We have for the importance weight

$$
\begin{aligned}
w_n\left(x_{1:n}\right) &= \frac{\gamma_n\left(x_{1:n}\right)}{q_{n-1}\left(x_{1:n-1}\right) q_n\left(x_n \mid x_{1:n-1}\right)} \\
&= \frac{Z_n \pi_n\left(x_{1:n-1}\right)}{q_{n-1}\left(x_{1:n-1}\right)} \frac{\pi_n\left(x_n \mid x_{1:n-1}\right)}{q_n\left(x_n \mid x_{1:n-1}\right)}
\end{aligned}
$$

- It follows directly that we have

$$
q_n^{\text{opt}}\left(x_n \mid x_{1:n-1}\right) = \pi_n\left(x_n \mid x_{1:n-1}\right)
$$

and

$$
\begin{aligned}
w_n\left(x_{1:n}\right) &= w_{n-1}\left(x_{1:n-1}\right) \frac{\gamma_n\left(x_{1:n}\right)}{\gamma_{n-1}\left(x_{1:n-1}\right) \pi_n\left(x_n \mid x_{1:n-1}\right)} \\
&= w_{n-1}\left(x_{1:n-1}\right) \frac{\gamma_n\left(x_{1:n-1}\right)}{\gamma_{n-1}\left(x_{1:n-1}\right)}
\end{aligned}
$$

- This locally optimal importance density will be used again and again.

- This locally optimal importance density will be used again and again.
- It is often impossible to sample from $\pi_n(x_n | x_{1:n-1})$ and/or computing $\gamma_n(x_{1:n-1}) = \int \gamma_n(x_{1:n})\, dx_n$.

- This locally optimal importance density will be used again and again.
- It is often impossible to sample from $\pi_n\left(x_n | x_{1:n-1}\right)$ and/or computing $\gamma_n\left(x_{1:n-1}\right) = \int \gamma_n\left(x_{1:n}\right) dx_n$.
- In such cases, it is necessary to approximate $\pi_n\left(x_n | x_{1:n-1}\right)$ and $\gamma_n\left(x_{1:n-1}\right)$.

Application to State-Space Models

- In the case of state-space models, we have

$$
\begin{aligned}
q_n^{\text{opt}}\left(x_n \mid x_{1:n-1}\right) &= p\left(x_n \mid y_{1:n}, x_{1:n-1}\right) = p\left(x_n \mid y_n, x_{n-1}\right) \\
&= \frac{g\left(y_n \mid x_n\right) f\left(x_n \mid x_{n-1}\right)}{p\left(y_n \mid x_{n-1}\right)}
\end{aligned}
$$

Application to State-Space Models

- In the case of state-space models, we have

$$\begin{aligned} q_n^{\text{opt}}\left(x_n \mid x_{1:n-1}\right) &= p\left(x_n \mid y_{1:n}, x_{1:n-1}\right) = p\left(x_n \mid y_n, x_{n-1}\right) \\ &= \frac{g\left(y_n \mid x_n\right) f\left(x_n \mid x_{n-1}\right)}{p\left(y_n \mid x_{n-1}\right)} \end{aligned}$$

- In this case,

$$\begin{aligned} w_n\left(x_{1:n}\right) &= w_{n-1}\left(x_{1:n-1}\right) \frac{p\left(x_{1:n}, y_{1:n}\right)}{p\left(x_{1:n-1}, y_{1:n-1}\right) p\left(x_n \mid y_n, x_{n-1}\right)} \\ &= w_{n-1}\left(x_{1:n-1}\right) p\left(y_n \mid x_{n-1}\right). \end{aligned}$$

Application to State-Space Models

- In the case of state-space models, we have

$$\begin{aligned} q_n^{\text{opt}}\left(x_n \mid x_{1:n-1}\right) &= p\left(x_n \mid y_{1:n}, x_{1:n-1}\right) = p\left(x_n \mid y_n, x_{n-1}\right) \\ &= \frac{g\left(y_n \mid x_n\right) f\left(x_n \mid x_{n-1}\right)}{p\left(y_n \mid x_{n-1}\right)} \end{aligned}$$

- In this case,

$$\begin{aligned} w_n\left(x_{1:n}\right) &= w_{n-1}\left(x_{1:n-1}\right) \frac{p\left(x_{1:n}, y_{1:n}\right)}{p\left(x_{1:n-1}, y_{1:n-1}\right) p\left(x_n \mid y_n, x_{n-1}\right)} \\ &= w_{n-1}\left(x_{1:n-1}\right) p\left(y_n \mid x_{n-1}\right). \end{aligned}$$

- **Example**: Consider $f\left(x_n \mid x_{n-1}\right) = \mathcal{N}\left(x_n; \alpha\left(x_{n-1}\right), \beta\left(x_{n-1}\right)\right)$ and $g\left(y_n \mid x_n\right) = \mathcal{N}\left(x_n; \sigma_w^2\right)$ then $p\left(x_n \mid y_n, x_{n-1}\right) = \mathcal{N}\left(x_n; m\left(x_{n-1}\right), \sigma^2\left(x_{n-1}\right)\right)$ with

$$\sigma^2\left(x_{n-1}\right) = \frac{\beta\left(x_{n-1}\right)\sigma_w^2}{\beta\left(x_{n-1}\right) + \sigma_w^2}, \; m\left(x_{n-1}\right) = \sigma^2\left(x_{n-1}\right)\left(\frac{\alpha\left(x_{n-1}\right)}{\beta\left(x_{n-1}\right)} + \frac{y_n}{\sigma_w^2}\right).$$

Application to Linear Gaussian State-Space Models

- Consider the simple model

$$
\begin{aligned}
X_n &= \alpha X_{n-1} + V_n, \\
Y_n &= X_n + \sigma W_n
\end{aligned}
$$

where $X_1 \sim \mathcal{N}(0,1)$, $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Application to Linear Gaussian State-Space Models

- Consider the simple model

$$
\begin{aligned}
X_n &= \alpha X_{n-1} + V_n, \\
Y_n &= X_n + \sigma W_n
\end{aligned}
$$

where $X_1 \sim \mathcal{N}(0,1)$, $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

- We use $q_n(x_n | x_{1:n-1}) = f(x_n | x_{n-1}) = \mathcal{N}(x_n; \alpha x_{n-1}, 1)$,

$$
q_n(x_n | x_{1:n-1}) = f(x_n | x_{n-1}) = \mathcal{N}(x_n; \alpha x_{n-1}, 1),
$$

$$
\begin{aligned}
q_n^{\text{opt}}(x_n | x_{1:n-1}) &= p(x_n | y_n, x_{n-1}) \\
&= \mathcal{N}\left(x_n; \frac{\sigma_w^2}{\sigma_w^2 + 1}\left(\alpha x_{n-1} + \frac{y_n}{\sigma_w^2}\right), \frac{\sigma_w^2}{\sigma_w^2 + 1}\right).
\end{aligned}
$$

Summary

- Sequential Importance Sampling is an attractive idea: sequential and parallelizable, only requires designing low-dimensional proposal distributions.

Summary

- Sequential Importance Sampling is an attractive idea: sequential and parallelizable, only requires designing low-dimensional proposal distributions.
- Sequential Importance Sampling can only work for moderate size problems.

Summary

- Sequential Importance Sampling is an attractive idea: sequential and parallelizable, only requires designing low-dimensional proposal distributions.
- Sequential Importance Sampling can only work for moderate size problems.
- Is there a way to **partially** fix this problem?