Sequential Monte Carlo: An Introduction

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 $\varphi_n : E_n \to \mathbb{R}$

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

and/or the normalizing constants Z_n.

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.

- We could use standard MCMC to sample from $\{\pi_n\}_{n\geq 1}$ but it is slow & it does not provide an estimate of $\{Z_n\}_{n\geq 1}$.

- We could use standard MCMC to sample from $\{\pi_n\}_{n\geq 1}$ but it is slow & it does not provide an estimate of $\{Z_n\}_{n\geq 1}$.
- SMC is a non-iterative alternative class of algorithms to MCMC.

- We could use standard MCMC to sample from $\{\pi_n\}_{n\geq 1}$ but it is slow & it does not provide an estimate of $\{Z_n\}_{n\geq 1}$.
- SMC is a non-iterative alternative class of algorithms to MCMC.
- *Key idea*: if π_{n-1} does not differ too much from π_n then we should be able to reuse our estimate of π_{n-1} to approximate π_n.

Applications

- Optimal estimation in non-linear non-Gaussian dynamic models.

Applications

- Optimal estimation in non-linear non-Gaussian dynamic models.
- Bayesian inference for complex statistical models.

Applications

- Optimal estimation in non-linear non-Gaussian dynamic models.
- Bayesian inference for complex statistical models.
- Global optimization.

Applications

- Optimal estimation in non-linear non-Gaussian dynamic models.
- Bayesian inference for complex statistical models.
- Global optimization.
- Counting problems.

Applications

- Optimal estimation in non-linear non-Gaussian dynamic models.
- Bayesian inference for complex statistical models.
- Global optimization.
- Counting problems.
- Rare events simulation.

State-Space Models

- $\{X_n\}_{n\geq 1}$ latent/hidden Markov process with

$$X_1 \sim \mu(\cdot) \text{ and } X_n|(X_{n-1} = x) \sim f(\cdot|x).$$

State-Space Models

- $\{X_n\}_{n\geq 1}$ latent/hidden Markov process with

$$X_1 \sim \mu(\cdot) \text{ and } X_n|(X_{n-1} = x) \sim f(\cdot|x).$$

- $\{Y_n\}_{n\geq 1}$ observation process such that observations are conditionally independent given $\{X_n\}_{n\geq 1}$ and

$$Y_n|(X_n = x) \sim g(\cdot|x).$$

State-Space Models

- $\{X_n\}_{n\geq 1}$ latent/hidden Markov process with

$$X_1 \sim \mu(\cdot) \text{ and } X_n|(X_{n-1} = x) \sim f(\cdot|x).$$

- $\{Y_n\}_{n\geq 1}$ observation process such that observations are conditionally independent given $\{X_n\}_{n\geq 1}$ and

$$Y_n|(X_n = x) \sim g(\cdot|x).$$

- Very wide class of statistical models also known as hidden Markov models with thousands of applications.

Examples

- *Linear Gaussian state-space model*

$$\begin{aligned} X_1 &\sim \mathcal{N}(m_1, \Sigma_1),\ X_n = AX_{n-1} + BV_n, \\ Y_n &= CX_n + DW_n \end{aligned}$$

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

Examples

- *Linear Gaussian state-space model*

$$\begin{aligned} X_1 &\sim \mathcal{N}(m_1, \Sigma_1),\ X_n = AX_{n-1} + BV_n, \\ Y_n &= CX_n + DW_n \end{aligned}$$

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

- *Stochastic volatility model*

$$\begin{aligned} X_1 &\sim \mathcal{N}\left(0, \frac{\sigma^2}{1-\alpha^2}\right),\ X_n = \alpha X_{n-1} + V_n, \\ Y_n &= \beta \exp(X_n/2)\, W_n \end{aligned}$$

where $|\alpha| < 1$, $V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, $W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

Inference in State-Space Models

- At time n, we have access to the observations are interested in computing

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

and the (marginal) likelihood $p\left(y_{1:n}\right)$ where

$$\begin{aligned} p\left(x_{1:n}, y_{1:n}\right) &= \mu\left(x_1\right) \prod_{k=2}^{n} f\left(x_k \middle| x_{k-1}\right) \prod_{k=1}^{n} g\left(y_k \middle| x_k\right), \\ p\left(y_{1:n}\right) &= \int \cdots \int p\left(x_{1:n}, y_{1:n}\right) dx_{1:n}. \end{aligned}$$

Inference in State-Space Models

- At time n, we have access to the observations are interested in computing

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

and the (marginal) likelihood $p\left(y_{1:n}\right)$ where

$$\begin{aligned} p\left(x_{1:n}, y_{1:n}\right) &= \mu\left(x_1\right) \prod_{k=2}^{n} f\left(x_k \middle| x_{k-1}\right) \prod_{k=1}^{n} g\left(y_k \middle| x_k\right), \\ p\left(y_{1:n}\right) &= \int \cdots \int p\left(x_{1:n}, y_{1:n}\right) dx_{1:n}. \end{aligned}$$

- In our SMC framework,

$$\pi_n\left(x_{1:n}\right) = p\left(x_{1:n} \middle| y_{1:n}\right), \ \gamma_n\left(x_{1:n}\right) = p\left(x_{1:n}, y_{1:n}\right), \ Z_n = p\left(y_{1:n}\right).$$

The Kalman Filter

- For linear Gaussian models, all posteriors are Gaussian and we can compute the likelihood exactly.

The Kalman Filter

- For linear Gaussian models, all posteriors are Gaussian and we can compute the likelihood exactly.
- The marginal distributions $\{p(x_n|y_{1:n})\}_{n\geq 1}$ and $\{p(y_n|y_{1:n-1})\}_{n\geq 1}$ can be computed through the celebrated Kalman filter.

The Kalman Filter

- For linear Gaussian models, all posteriors are Gaussian and we can compute the likelihood exactly.
- The marginal distributions $\{p(x_n|y_{1:n})\}_{n\geq 1}$ and $\{p(y_n|y_{1:n-1})\}_{n\geq 1}$ can be computed through the celebrated Kalman filter.
- To obtain an estimate of the joint distribution, we have

$$\begin{aligned} p(x_{1:n}|y_{1:n}) &= p(x_n|y_{1:n}) \prod_{k=1}^{n-1} p(x_k|y_{1:n}, x_{k+1}) \\ &= p(x_n|y_{1:n}) \prod_{k=1}^{n-1} p(x_k|y_{1:k}, x_{k+1}) \end{aligned}$$

where

$$p(x_k|y_{1:k}, x_{k+1}) = \frac{f(x_{k+1}|x_k)\, p(x_k|y_{1:k})}{p(x_{k+1}|y_{1:k})}.$$

Nonlinear Non-Gaussian Models

- For nonlinear non-Gaussian models, there is *no closed-form expression*.

Nonlinear Non-Gaussian Models

- For nonlinear non-Gaussian models, there is *no closed-form expression*.
- Standard approximations relie on functional approximations: EKF, UKF, Gaussian quadrature, mixture of Gaussians.

Nonlinear Non-Gaussian Models

- For nonlinear non-Gaussian models, there is *no closed-form expression*.
- Standard approximations relie on functional approximations: EKF, UKF, Gaussian quadrature, mixture of Gaussians.
- These functional approximations can be seriously unreliable and are not widely applicable.

Quantum Monte Carlo

- Finding the largest eigenvalue and eigenmeasure of a positive operator

Quantum Monte Carlo

- Finding the largest eigenvalue and eigenmeasure of a positive operator
- Let $K : E \times E \longrightarrow \mathbb{R}^+$ be a positive kernel.

Quantum Monte Carlo

- Finding the largest eigenvalue and eigenmeasure of a positive operator
- Let $K : E \times E \longrightarrow \mathbb{R}^+$ be a positive kernel.
- Find the largest eigenvalue λ ($\lambda > 0$) and associated eigenmeasure μ ($\int \mu(dx) = 1$) of K

$$\int \mu(x) K(y|x) dx = \lambda \mu(y).$$

Quantum Monte Carlo

- Finding the largest eigenvalue and eigenmeasure of a positive operator
- Let $K : E \times E \longrightarrow \mathbb{R}^+$ be a positive kernel.
- Find the largest eigenvalue λ ($\lambda > 0$) and associated eigenmeasure μ ($\int \mu(dx) = 1$) of K

$$\int \mu(x) K(y|x) dx = \lambda \mu(y).$$

- **Basic Idea**: the good old power method.

- *Power method*: A $p \times p$ matrix with p linearly independent eigenvectors $\{V_i\}$ associated to eigenvalues $\{\lambda_i\}$ such that $|\lambda_1| > |\lambda_2| > ... > |\lambda_p|$

$$\begin{aligned} U_1 &= \sum_{i=1}^{p} \alpha_i V_i, \\ &\vdots \\ U_n &= A^{n-1} U_1 = \sum_{i=1}^{p} \alpha_i \lambda_i^{n-1} V_i \end{aligned}$$

- *Power method*: A $p \times p$ matrix with p linearly independent eigenvectors $\{V_i\}$ associated to eigenvalues $\{\lambda_i\}$ such that $|\lambda_1| > |\lambda_2| > ... > |\lambda_p|$

$$\begin{aligned} U_1 &= \sum_{i=1}^{p} \alpha_i V_i, \\ &\vdots \\ U_n &= A^{n-1} U_1 = \sum_{i=1}^{p} \alpha_i \lambda_i^{n-1} V_i \end{aligned}$$

- We have

$$\frac{U_n}{\lambda_1^{n-1}} = \alpha_1 V_1 + \sum_{i=2}^{p} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{n-1} V_i \to \alpha_1 V_1 \text{ and } \frac{U_n^{\mathsf{T}} Y}{U_{n-1}^{\mathsf{T}} Y} \to \lambda_1.$$

- Consider the following artificial sequence of distributions defined through

$$\gamma_n(x_{1:n}) = v(x_1) \prod_{k=2}^{n} K(x_k | x_{k-1})$$

- Consider the following artificial sequence of distributions defined through

$$\gamma_n(x_{1:n}) = v(x_1) \prod_{k=2}^{n} K(x_k | x_{k-1})$$

- As n increases, we have

$$\gamma_n(x_n) = \int \gamma_n(x_{1:n}) \, dx_{1:n-1} \propto \lambda^{n-1} \mu(x_n),$$

and

$$\pi_n(x_n) \to \mu(x_n) \text{ and } \frac{Z_{n+1}}{Z_n} \to \lambda.$$

- Consider the following artificial sequence of distributions defined through

$$\gamma_n(x_{1:n}) = v(x_1) \prod_{k=2}^{n} K(x_k | x_{k-1})$$

- As n increases, we have

$$\gamma_n(x_n) = \int \gamma_n(x_{1:n})\, dx_{1:n-1} \propto \lambda^{n-1} \mu(x_n),$$

and

$$\pi_n(x_n) \to \mu(x_n) \text{ and } \frac{Z_{n+1}}{Z_n} \to \lambda.$$

- SMC methods are widely used to solve this problem.

Self-Avoiding Random Walk (SAW)

- A 2D Self Avoiding Random Walk (SAW). Polymer of size n is characterized by a sequence $x_{1:n}$ on a finite lattice such that $x_i \neq x_j$ for $i \neq j$.

Self-Avoiding Random Walk (SAW)

- A 2D Self Avoiding Random Walk (SAW). Polymer of size n is characterized by a sequence $x_{1:n}$ on a finite lattice such that $x_i \neq x_j$ for $i \neq j$.
- One is interested in the uniform distribution

$$\pi_n(x_{1:n}) = Z_n^{-1} . 1_{D_n}(x_{1:n})$$

where

$$\begin{aligned} D_n &= \{x_{1:n} \in E_n \backslash\ x_k \sim x_{k+1} \text{ and } x_k \neq x_i \text{ for } k \neq i\}, \\ Z_n &= \text{cardinal of } D_n. \end{aligned}$$

Self-Avoiding Random Walk (SAW)

- A 2D Self Avoiding Random Walk (SAW). Polymer of size n is characterized by a sequence $x_{1:n}$ on a finite lattice such that $x_i \neq x_j$ for $i \neq j$.
- One is interested in the uniform distribution

$$\pi_n(x_{1:n}) = Z_n^{-1}.1_{D_n}(x_{1:n})$$

where

$$\begin{aligned} D_n &= \{x_{1:n} \in E_n \backslash\ x_k \sim x_{k+1} \text{ and } x_k \neq x_i \text{ for } k \neq i\}, \\ Z_n &= \text{cardinal of } D_n. \end{aligned}$$

- SMC allow us to simulate from the uniform distribution of SAW of length n and to compute their number.

Particle Motion in Random Medium

- A Markovian particle $\{X_n\}_{n\geq 1}$ evolves in a random medium

$$X_1 \sim \mu(\cdot), \;\; X_{n+1} | \, X_n = x \sim f(\cdot | \, x).$$

Particle Motion in Random Medium

- A Markovian particle $\{X_n\}_{n\geq 1}$ evolves in a random medium

$$X_1 \sim \mu(\cdot), \;\; X_{n+1} | \, X_n = x \sim f(\cdot | \, x).$$

- At time n, its probability to get killed is $1 - g(X_n)$ where $0 \leq g(x) \leq 1$ for any $x \in E$.

Particle Motion in Random Medium

- A Markovian particle $\{X_n\}_{n\geq 1}$ evolves in a random medium

$$X_1 \sim \mu(\cdot), \;\; X_{n+1} | X_n = x \sim f(\cdot | x).$$

- At time n, its probability to get killed is $1 - g(X_n)$ where $0 \leq g(x) \leq 1$ for any $x \in E$.
- One wants to approximate $\Pr(T > n)$ where $T =$ Random time at which the particle is killed.

- One has

$$\begin{aligned} & \Pr(T > n) \\ = \; & \mathbb{E}_{\mu}\left[\text{Proba. of not being killed at } n \text{ given } X_{1:n}\right] \\ = \; & \int \cdots \int \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \underbrace{\prod_{k=1}^{n} g(x_k)}_{\text{Probability to survive at } n} dx_{1:n}. \end{aligned}$$

- One has

$$\begin{aligned} & \Pr(T > n) \\ = \; & \mathbb{E}_{\mu}\left[\text{Proba. of not being killed at } n \text{ given } X_{1:n}\right] \\ = \; & \int \cdots \int \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \underbrace{\prod_{k=1}^{n} g(x_k)}_{\text{Probability to survive at } n} dx_{1:n}. \end{aligned}$$

- Consider

$$\begin{aligned} \gamma_n(x_{1:n}) &= \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \prod_{k=1}^{n} g(x_k), \\ \pi_n(x_{1:n}) &= \frac{\gamma_n(x_{1:n})}{Z_n} \text{ where } Z_n = \Pr(T > n). \end{aligned}$$

- One has

$$\begin{aligned} & \Pr\left(T > n\right) \\ = \; & \mathbb{E}_{\mu}\left[\text{Proba. of not being killed at } n \text{ given } X_{1:n}\right] \\ = \; & \int \cdots \int \mu\left(x_1\right) \prod_{k=2}^{n} f\left(x_k \middle| x_{k-1}\right) \underbrace{\prod_{k=1}^{n} g\left(x_k\right)}_{\text{Probability to survive at } n} dx_{1:n}. \end{aligned}$$

- Consider

$$\begin{aligned} \gamma_n\left(x_{1:n}\right) &= \mu\left(x_1\right) \prod_{k=2}^{n} f\left(x_k \middle| x_{k-1}\right) \prod_{k=1}^{n} g\left(x_k\right), \\ \pi_n\left(x_{1:n}\right) &= \frac{\gamma_n\left(x_{1:n}\right)}{Z_n} \text{ where } Z_n = \Pr\left(T > n\right). \end{aligned}$$

- SMC methods to compute Z_n, the probability of not being killed at time n, and to approximate the distribution of the paths having survived at time n.

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.
- *Examples*

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.
- *Examples*
  - $\pi_n = \pi$ (e.g. Bayesian inference, rare events etc.)

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.
- *Examples*
  - $\pi_n = \pi$ (e.g. Bayesian inference, rare events etc.)
  - $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ where $\gamma_n \to \infty$ (global optimization)

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.
- *Examples*
  - $\pi_n = \pi$ (e.g. Bayesian inference, rare events etc.)
  - $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ where $\gamma_n \longrightarrow \infty$ (global optimization)
  - $\pi_n(x) = p(x|y_{1:n})$ (sequential Bayesian estimation)

Generic Sequence of Target Distributions

- Consider the case where all the target distributions $\{\pi_n\}_{n\geq 1}$ are defined on $E_n = E$.
- *Examples*
  - $\pi_n = \pi$ (e.g. Bayesian inference, rare events etc.)
  - $\pi_n(x) \propto [\pi(x)]^{\gamma_n}$ where $\gamma_n \to \infty$ (global optimization)
  - $\pi_n(x) = p(x|y_{1:n})$ (sequential Bayesian estimation)
- SMC do not apply to this problem as it requires $E_n = E^n$.

- Consider a new sequence of *artificial* distributions $\{\widetilde{\pi}_n\}_{n\geq 1}$ defined on $E_n = E^n$ such that

$$\int \widetilde{\pi}_n\left(x_{1:n-1}, x_n\right) dx_{1:n-1} = \pi_n\left(x_n\right)$$

and apply standard SMC.

- Consider a new sequence of *artificial* distributions $\{\widetilde{\pi}_n\}_{n\geq 1}$ defined on $E_n = E^n$ such that

$$\int \widetilde{\pi}_n\left(x_{1:n-1}, x_n\right) dx_{1:n-1} = \pi_n\left(x_n\right)$$

and apply standard SMC.

- *Example*:

$$\widetilde{\pi}_n\left(x_{1:n-1}, x_n\right) = \pi_n\left(x_n\right) \widetilde{\pi}_n\left(x_{1:n-1} \middle| x_n\right)$$

where $\widetilde{\pi}_n\left(x_{1:n-1} \middle| x_n\right)$ is *any* conditional distribution on E^{n-1}.

- Consider a new sequence of *artificial* distributions $\{\widetilde{\pi}_n\}_{n\geq 1}$ defined on $E_n = E^n$ such that

$$\int \widetilde{\pi}_n\left(x_{1:n-1}, x_n\right) dx_{1:n-1} = \pi_n\left(x_n\right)$$

  and apply standard SMC.

- *Example*:

$$\widetilde{\pi}_n\left(x_{1:n-1}, x_n\right) = \pi_n\left(x_n\right) \widetilde{\pi}_n\left(x_{1:n-1} \middle| x_n\right)$$

  where $\widetilde{\pi}_n\left(x_{1:n-1} \middle| x_n\right)$ is *any* conditional distribution on E^{n-1}.

- How to design $\widetilde{\pi}_n$ optimally will be discussed later.

The Need for Monte Carlo Methods

- Except in trivial cases, one can neither compute $\int \varphi_n(x_{1:n})\,\pi_n(dx_{1:n})$ nor Z_n.

The Need for Monte Carlo Methods

- Except in trivial cases, one can neither compute $\int \varphi_n(x_{1:n})\,\pi_n(dx_{1:n})$ nor Z_n.
- Deterministic numerical integration methods typically inefficient for high-dimensional spaces.

The Need for Monte Carlo Methods

- Except in trivial cases, one can neither compute $\int \varphi_n(x_{1:n})\,\pi_n(dx_{1:n})$ nor Z_n.
- Deterministic numerical integration methods typically inefficient for high-dimensional spaces.
- Monte Carlo methods: simple and flexible.

The Need for Monte Carlo Methods

- Except in trivial cases, one can neither compute $\int \varphi_n(x_{1:n})\,\pi_n(dx_{1:n})$ nor Z_n.
- Deterministic numerical integration methods typically inefficient for high-dimensional spaces.
- Monte Carlo methods: simple and flexible.
- Using Monte Carlo, it is very easy to make “rigourous" your intuition.

Monte Carlo Methods

- For the time being, just concentrate on estimating

$$\mathbb{E}_\pi [\varphi] = \int \varphi(x)\, \pi(dx)$$

where

$$\pi(x) = \frac{\gamma(x)}{Z} \text{ with } \gamma \text{ known pointwise}/Z = \int \gamma(x)\, dx \text{ unknown.}$$

Monte Carlo Methods

- For the time being, just concentrate on estimating

$$\mathbb{E}_\pi[\varphi] = \int \varphi(x)\,\pi(dx)$$

where

$$\pi(x) = \frac{\gamma(x)}{Z} \text{ with } \gamma \text{ known pointwise}/Z = \int \gamma(x)\,dx \text{ unknown.}$$

- Draw a large number samples $X^{(i)} \overset{\text{i.i.d.}}{\sim} \pi$ and build empirical measure

$$\widehat{\pi}(dx) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X^{(i)}}(dx).$$

- *Marginalization is straightforward.* If $x = (x_1, \ldots, x_k)$

$$\widehat{\pi}\left(dx_p\right) = \int \widehat{\pi}\left(dx_{1:p-1}, dx_{p+1:k}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_p^{(i)}}\left(dx\right).$$

- *Marginalization is straightforward*. If $x = (x_1, \ldots, x_k)$

$$\widehat{\pi}\left(dx_p\right) = \int \widehat{\pi}\left(dx_{1:p-1}, dx_{p+1:k}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_p^{(i)}}\left(dx\right).$$

- *Integration is straightforward*. Monte Carlo estimates of $\mathbb{E}_{\pi}\left(\varphi\right)$

$$\mathbb{E}_{\widehat{\pi}}\left(\varphi\right) = \int \varphi\left(x\right) \widehat{\pi}\left(dx\right) = \frac{1}{N}\sum_{i=1}^{N} \varphi\left(X^{(i)}\right).$$

- *Marginalization is straightforward*. If $x = (x_1, \ldots, x_k)$

$$\widehat{\pi}\left(dx_p\right) = \int \widehat{\pi}\left(dx_{1:p-1}, dx_{p+1:k}\right) = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_p^{(i)}}\left(dx\right).$$

- *Integration is straightforward*. Monte Carlo estimates of $\mathbb{E}_{\pi}\left(\varphi\right)$

$$\mathbb{E}_{\widehat{\pi}}\left(\varphi\right) = \int \varphi\left(x\right) \widehat{\pi}\left(dx\right) = \frac{1}{N}\sum_{i=1}^{N} \varphi\left(X^{(i)}\right).$$

- Samples concentrate themselves automatically in regions of high probability mass whatever being the dimension of the space; e.g. $E = \mathbb{R}^{10^6}$.

- Basic results

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}(\varphi)\right] &= \mathbb{E}_{\pi}(\varphi) \text{ unbiased,} \\ \mathbb{V}\left[E_{\widehat{\pi}}(\varphi)\right] &= \frac{1}{N}\mathbb{E}_{\pi}\left((\varphi - \mathbb{E}_{\pi}(\varphi))^2\right)\end{aligned}$$

- Basic results

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}\left(\varphi\right)\right] &= \mathbb{E}_{\pi}\left(\varphi\right) \text{ unbiased,} \\ \mathbb{V}\left[E_{\widehat{\pi}}\left(\varphi\right)\right] &= \frac{1}{N}\mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}\left(\varphi\right)\right)^2\right)\end{aligned}$$

- Rate of convergence to zero **INDEPENDENT** of space E! It breaks the curse of dimensionality... sometimes.

- Basic results

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}\left(\varphi\right)\right] &= \mathbb{E}_{\pi}\left(\varphi\right) \text{ unbiased,} \\ \mathbb{V}\left[E_{\widehat{\pi}}\left(\varphi\right)\right] &= \frac{1}{N}\mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}\left(\varphi\right)\right)^2\right)\end{aligned}$$

- Rate of convergence to zero **INDEPENDENT** of space E! It breaks the curse of dimensionality... sometimes.
- Central limit theorem

$$\sqrt{N}\left(\mathbb{E}_{\widehat{\pi}}\left(\varphi\right) - \mathbb{E}_{\pi}\left(\varphi\right)\right) \Rightarrow \mathcal{N}\left(0, \mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}\left(\varphi\right)\right)^2\right)\right)$$

- Basic results

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}\left(\varphi\right)\right] &= \mathbb{E}_{\pi}\left(\varphi\right) \text{ unbiased,} \\ \mathbb{V}\left[E_{\widehat{\pi}}\left(\varphi\right)\right] &= \frac{1}{N}\mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}\left(\varphi\right)\right)^2\right)\end{aligned}$$

- Rate of convergence to zero **INDEPENDENT** of space E! It breaks the curse of dimensionality... sometimes.
- Central limit theorem

$$\sqrt{N}\left(\mathbb{E}_{\widehat{\pi}}\left(\varphi\right) - \mathbb{E}_{\pi}\left(\varphi\right)\right) \Rightarrow \mathcal{N}\left(0, \mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}\left(\varphi\right)\right)^2\right)\right)$$

- **Problem**: how do you obtain samples from an arbitary high dimensional distribution???

- Basic results

$$\begin{aligned}\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}(\varphi)\right] &= \mathbb{E}_{\pi}(\varphi) \text{ unbiased,} \\ \mathbb{V}\left[E_{\widehat{\pi}}(\varphi)\right] &= \frac{1}{N}\mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}(\varphi)\right)^2\right)\end{aligned}$$

- Rate of convergence to zero **INDEPENDENT** of space E! It breaks the curse of dimensionality... sometimes.
- Central limit theorem

$$\sqrt{N}\left(\mathbb{E}_{\widehat{\pi}}(\varphi) - \mathbb{E}_{\pi}(\varphi)\right) \Rightarrow \mathcal{N}\left(0, \mathbb{E}_{\pi}\left(\left(\varphi - \mathbb{E}_{\pi}(\varphi)\right)^2\right)\right)$$

- **Problem**: how do you obtain samples from an arbitary high dimensional distribution???
- **Answer**: No general answer, typically approximation required.

Standard Monte Carlo Methods

- Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by germans) using inverse method, accept/reject etc.

Standard Monte Carlo Methods

- Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by germans) using inverse method, accept/reject etc.
- Sampling approximately from non standard high dimensional distributions typically done by Markov chain Monte Carlo (e.g. Metropolis-Hastings).

Standard Monte Carlo Methods

- Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by germans) using inverse method, accept/reject etc.
- Sampling approximately from non standard high dimensional distributions typically done by Markov chain Monte Carlo (e.g. Metropolis-Hastings).
- **Basic (bright) idea**: Build an ergodic Markov chain whose stationary distribution is the distribution of interest; i.e.

$$\int \pi(x) K(y|x)\, dx = \pi(y).$$

Standard Monte Carlo Methods

- Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by germans) using inverse method, accept/reject etc.
- Sampling approximately from non standard high dimensional distributions typically done by Markov chain Monte Carlo (e.g. Metropolis-Hastings).
- **Basic (bright) idea**: Build an ergodic Markov chain whose stationary distribution is the distribution of interest; i.e.

$$\int \pi(x) K(y|x) dx = \pi(y).$$

- Iterative algorithm to sample from one distribution, not adapted to our problems.

Standard Monte Carlo Methods

- Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by germans) using inverse method, accept/reject etc.
- Sampling approximately from non standard high dimensional distributions typically done by Markov chain Monte Carlo (e.g. Metropolis-Hastings).
- **Basic (bright) idea**: Build an ergodic Markov chain whose stationary distribution is the distribution of interest; i.e.

$$\int \pi(x) K(y|x) dx = \pi(y).$$

- Iterative algorithm to sample from one distribution, not adapted to our problems.
- **Alternative (not that bright) idea**: Importance sampling $\Rightarrow$ Non iterative, can be understood in one minute.

Importance Sampling

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

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

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

Importance Sampling

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

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

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

$$\begin{aligned}\widehat{\pi}\left(dx\right) &= \frac{w\left(x\right)\widehat{q}\left(dx\right)}{\int w\left(x\right)\widehat{q}\left(dx\right)} = \frac{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)\delta_{X^{(i)}}\left(dx\right)}{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)} \\ &= \sum_{i=1}^{N} W^{(i)}\delta_{X^{(i)}}\left(dx\right)\end{aligned}$$

where

$$W^{(i)} \propto w\left(X^{(i)}\right) \text{ and } \sum_{i=1}^{N} W^{(i)} = 1.$$

- Plugging this expression in IS identity

$$\begin{aligned}\widehat{\pi}\left(dx\right) &= \frac{w\left(x\right)\widehat{q}\left(dx\right)}{\int w\left(x\right)\widehat{q}\left(dx\right)} = \frac{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)\delta_{X^{(i)}}\left(dx\right)}{N^{-1}\sum_{i=1}^{N} w\left(X^{(i)}\right)} \\ &= \sum_{i=1}^{N} W^{(i)}\delta_{X^{(i)}}\left(dx\right)\end{aligned}$$

where

$$W^{(i)} \propto w\left(X^{(i)}\right) \text{ and } \sum_{i=1}^{N} W^{(i)} = 1.$$

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

- Now we can approximate $\mathbb{E}_{\pi}[\varphi]$ by

$$\mathbb{E}_{\widehat{\pi}}[\varphi] = \int \varphi(x)\,\widehat{\pi}(dx) = \sum_{i=1}^{N} W^{(i)} \varphi\left(X^{(i)}\right).$$

- Now we can approximate $\mathbb{E}_\pi[\varphi]$ by

$$\mathbb{E}_{\widehat{\pi}}[\varphi] = \int \varphi(x)\,\widehat{\pi}(dx) = \sum_{i=1}^{N} W^{(i)} \varphi\left(X^{(i)}\right).$$

- Statistics for $N \gg 1$

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &= \mathbb{E}_\pi[\varphi] - \underbrace{N_\pi^{-1}\mathbb{E}\left[W(X)\left(\varphi(X) - \mathbb{E}_\pi[\varphi]\right)\right]}_{\text{negligible bias}}, \\
\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &= N_\pi^{-1}\mathbb{E}\left[W(X)\left(\varphi(X) - \mathbb{E}_\pi[\varphi]\right)^2\right].
\end{aligned}$$

- Now we can approximate $\mathbb{E}_\pi[\varphi]$ by

$$\mathbb{E}_{\widehat{\pi}}[\varphi] = \int \varphi(x)\,\widehat{\pi}(dx) = \sum_{i=1}^{N} W^{(i)} \varphi\left(X^{(i)}\right).$$

- Statistics for $N \gg 1$

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &= \mathbb{E}_\pi[\varphi] - \underbrace{N_\pi^{-1}\mathbb{E}\left[W(X)\left(\varphi(X) - \mathbb{E}_\pi[\varphi]\right)\right]}_{\text{negligible bias}}, \\
\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}[\varphi]\right] &= N_\pi^{-1}\mathbb{E}\left[W(X)\left(\varphi(X) - \mathbb{E}_\pi[\varphi]\right)^2\right].
\end{aligned}$$

- Estimate of normalizing constant

$$\widehat{Z} = \int \frac{\gamma(x)}{q(x)}\widehat{q}(dx) = \frac{1}{N}\sum_{i=1}^{N} \frac{\gamma\left(X^{(i)}\right)}{q\left(X^{(i)}\right)}$$

and $\mathbb{E}\left[\widehat{Z}\right] = Z$, $\mathbb{V}\left[\widehat{Z}\right] = N^{-1}\left(\mathbb{E}_q\left[\left(\frac{\gamma(X)}{q(X)} - Z\right)^2\right]\right)$.

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.
- In applications we are interested in, there is typically no specific φ of interest.

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.
- In applications we are interested in, there is typically no specific φ of interest.
- Practical recommendations

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.
- In applications we are interested in, there is typically no specific φ of interest.
- Practical recommendations
  - Select q as close to π as possible.

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.
- In applications we are interested in, there is typically no specific φ of interest.
- Practical recommendations
  - Select q as close to π as possible.
  - Ensure

$$w\left(x\right) = \frac{\pi\left(x\right)}{q\left(x\right)} < \infty.$$

- For a given φ, importance distribution minimizing $\mathbb{V}\left[\mathbb{E}_{\widehat{\pi}}\left[\varphi\right]\right]$ is

$$q^{\text{opt}}\left(x\right) = \frac{\left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right)}{\int \left|\varphi\left(x\right) - \mathbb{E}_{\pi}\left[\varphi\right]\right| \pi\left(x\right) dx}.$$

- Useless as sampling from q^{opt} as complex as solving the original problem.
- In applications we are interested in, there is typically no specific φ of interest.
- Practical recommendations
  - Select q as close to π as possible.
  - Ensure

$$w\left(x\right) = \frac{\pi\left(x\right)}{q\left(x\right)} < \infty.$$

- IS methods typically used for problems of limited dimension; say $E = \mathbb{R}^{25} \Rightarrow$ For more complex problems, MCMC are favoured.