Sequential Monte Carlo: An Introduction

Arnaud Doucet
Departments of Statistics & Computer Science
University of British Columbia
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 \( E_n = E_{n-1} \times E \), \( \mathcal{F}_n = \mathcal{F}_{n-1} \times \mathcal{F} \).
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 \( \pi_n \left( dx_{1:n} \right) = \pi_n \left( x_{1:n} \right) dx_{1:n} \) is known up to a normalizing constant, i.e.

\[
\pi_n \left( x_{1:n} \right) = \frac{\gamma_n \left( x_{1:n} \right)}{Z_n}
\]
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 $Z_n$. 
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 \( \pi_1 \) and/or \( Z_1 \) at time 1 then \( \pi_2 \) 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 \( \pi_{n-1} \) does not differ too much from \( \pi_n \) then we should be able to reuse our estimate of \( \pi_{n-1} \) to approximate \( \pi_n \).
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

\[
X_1 \sim \mu(\cdot) \quad \text{and} \quad X_n \mid (X_{n-1} = x) \sim f(\cdot \mid x).
\]

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

- \( \{X_n\}_{n \geq 1} \) latent/hidden Markov process with
  \[
  X_1 \sim \mu(\cdot) \quad \text{and} \quad X_n| (X_{n-1} = x) \sim f(\cdot|x).
  \]

- \( \{Y_n\}_{n \geq 1} \) 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
  \[ X_1 \sim \mu(\cdot) \quad \text{and} \quad X_n | (X_{n-1} = x) \sim f(\cdot | x). \]

- \( \{Y_n\}_{n \geq 1} \) 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**

  \[ X_1 \sim \mathcal{N}(m_1, \Sigma_1), \quad X_n = AX_{n-1} + BV_n, \]
  \[ Y_n = CX_n + DW_n \]

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

- **Linear Gaussian state-space model**

\[ X_1 \sim \mathcal{N}(m_1, \Sigma_1), \quad X_n = AX_{n-1} + BV_n, \]
\[ Y_n = CX_n + DW_n \]

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**

\[ X_1 \sim \mathcal{N}\left(0, \frac{\sigma^2}{1 - \alpha^2}\right), \quad X_n = \alpha X_{n-1} + V_n, \]
\[ Y_n = \beta \exp\left(\frac{X_n}{2}\right) W_n \]

where \(|\alpha| < 1\), \( V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2) \), \( W_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \).
At time $n$, we have access to the observations are interested in computing

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

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

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

$$p(y_{1:n}) = \int \cdots \int p(x_{1:n}, y_{1:n}) \, dx_{1:n}.$$
At time $n$, we have access to the observations and are interested in computing

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

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

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

$$p(y_{1:n}) = \int \cdots \int p(x_{1:n}, y_{1:n}) \, dx_{1:n}.$$ 

In our SMC framework,

$$\pi_n(x_{1:n}) = p(x_{1:n} \mid y_{1:n}), \quad \gamma_n(x_{1:n}) = p(x_{1:n}, y_{1:n}), \quad Z_n = p(y_{1:n}).$$
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.
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

\[
p(x_{1:n} | y_{1:n}) = p(x_n | y_{1:n}) \prod_{k=1}^{n-1} p(x_k | y_{1:n}, x_{k+1})
\]

\[
= p(x_n | y_{1:n}) \prod_{k=1}^{n-1} p(x_k | y_{1:k}, x_{k+1})
\]

where

\[
p(x_k | y_{1:k}, x_{k+1}) = \frac{f(x_{k+1} | x_k) p(x_k | y_{1:k})}{p(x_{k+1} | y_{1:k})}.
\]
For nonlinear non-Gaussian models, there is \textit{no closed-form expression}. 
For nonlinear non-Gaussian models, there is *no closed-form expression*. Standard approximations rely 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 rely on functional approximations: EKF, UKF, Gaussian quadrature, mixture of Gaussians.
- These functional approximations can be seriously unreliable and are not widely applicable.
Finding the largest eigenvalue and eigenmeasure of a positive operator
Quantum Monte Carlo

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

Let $K : E \times E \to \mathbb{R}^+$ be a positive kernel.

Find the largest eigenvalue $\lambda$ ($\lambda > 0$) and associated eigenmeasure $\mu$ ($\int \mu\,dx = 1$) of $K$

$$\int \mu(x) K(y|x)\,dx = \lambda \mu(y).$$
Finding the largest eigenvalue and eigenmeasure of a positive operator

Let $K : E \times E \rightarrow \mathbb{R}^+$ be a positive kernel.

Find the largest eigenvalue $\lambda$ ($\lambda > 0$) and associated eigenmeasure $\mu$ ($\int \mu (dx) = 1$) 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 $|\lambda_1| > |\lambda_2| > ... > |\lambda_p|$

\[
U_1 = \sum_{i=1}^{p} \alpha_i V_i,
\]

\[
U_n = A^{n-1} U_1 = \sum_{i=1}^{p} \alpha_i \lambda_i^{n-1} V_i
\]
**Power method:** A $p \times p$ matrix with $p$ linearly independent eigenvectors $\{V_i\}$ associated to eigenvalues $\{\lambda_i\}$ such that $|\lambda_1| > |\lambda_2| > ... > |\lambda_p|$ 

\[
U_1 = \sum_{i=1}^{p} \alpha_i V_i,
\]

\[
\vdots
\]

\[
U_n = A^{n-1} U_1 = \sum_{i=1}^{p} \alpha_i \lambda_i^{n-1} V_i
\]

- We have

\[
\frac{U_n}{\lambda_1^{n-1}} = \alpha_1 V_1 + \sum_{i=2}^{p} \alpha_i \left(\frac{\lambda_i}{\lambda_1}\right)^{n-1} V_i \rightarrow \alpha_1 V_1 \text{ and } \frac{U_n^T Y}{U_{n-1}^T Y} \rightarrow \lambda_1.
\]
Consider the following artificial sequence of distributions defined through

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

As \( n \) increases, we have

\[ \gamma_n(x_{1:n}) = Z_{\gamma_n(x_{1:n})} dx_1: \]

and

\[ \pi_n(x_{1:n}) = \mu(x_{1:n}) Z_{n+1} Z_{n} \lambda. \]

SMC methods are widely used to solve this problem.
Consider the following artificial sequence of distributions defined through

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

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

\[ \pi_n (x_n) \rightarrow \mu (x_n) \quad \text{and} \quad \frac{Z_{n+1}}{Z_n} \rightarrow \lambda. \]
Consider the following artificial sequence of distributions defined through

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

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

\[ \pi_n(x_n) \rightarrow \mu(x_n) \text{ and } \frac{Z_{n+1}}{Z_n} \rightarrow \lambda. \]

SMC methods are widely used to solve this problem.
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

$$D_n = \{ x_{1:n} \in E_n \setminus x_k \sim x_{k+1} \text{ and } x_k \neq x_i \text{ for } k \neq i \},$$

$$Z_n = \text{cardinal of } D_n.$$
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

$$
D_n = \{x_{1:n} \in E_n \setminus x_k \sim x_{k+1} \text{ and } x_k \neq x_i \text{ for } k \neq i \},
$$

$$
Z_n = \text{cardinal of } D_n.
$$

- SMC allow us to simulate from the uniform distribution of SAW of length $n$ and to compute their number.
A Markovian particle $\{X_n\}_{n \geq 1}$ evolves in a random medium

$$X_1 \sim \mu(\cdot), \ X_{n+1}|X_n = x \sim f(\cdot|x).$$
A Markovian particle $\{X_n\}_{n \geq 1}$ evolves in a random medium

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

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

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

---

A.D. ()
A Markovian particle \( \{X_n\}_{n \geq 1} \) evolves in a random medium

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

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

\[
\Pr(T > n) = \mathbb{E}_\mu [\text{Proba. of not being killed at } n \text{ given } X_{1:n}]
\]

\[
= \int \cdots \int \mu(x_1) \prod_{k=2}^{n} f(x_k | x_{k-1}) \prod_{k=1}^{n} g(x_k) \, dx_{1:n}.
\]

Probability to survive at \(n\)
One has

\[
\Pr(T > n) = \mathbb{E}_\mu \left[ \text{Proba. of not being killed at } n \text{ given } X_{1:n} \right] = \int \cdots \int \mu(x_1) \prod_{k=2}^{n} f(x_k \mid x_{k-1}) \prod_{k=1}^{n} g(x_k) \, dx_{1:n}.
\]

Probability to survive at \( n \)

Consider

\[
\gamma_n(x_{1:n}) = \mu(x_1) \prod_{k=2}^{n} f(x_k \mid x_{k-1}) \prod_{k=1}^{n} g(x_k),
\]

\[
\pi_n(x_{1:n}) = \frac{\gamma_n(x_{1:n})}{Z_n} \text{ where } Z_n = \Pr(T > n).
\]
One has

\[
\Pr ( T > n ) = \mathbb{E}_\mu [ \text{Proba. of not being killed at } n \text{ given } X_{1:n}] = \int \cdots \int \mu ( x_1 ) \prod_{k=2}^{n} f ( x_k | x_{k-1} ) \prod_{k=1}^{n} g ( x_k ) \ dx_{1:n}.
\]

Probability to survive at \( n \)

Consider

\[
\gamma_n ( x_{1:n} ) = \mu ( x_1 ) \prod_{k=2}^{n} f ( x_k | x_{k-1} ) \prod_{k=1}^{n} g ( x_k ),
\]

\[
\pi_n ( x_{1:n} ) = \frac{\gamma_n ( x_{1:n} )}{Z_n} \text{ where } Z_n = \Pr ( T > n ).
\]

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

\[ \pi_n(x) \propto \left[ \pi(x) \right]^\gamma_n \text{ where } \gamma_n! \to \infty \text{ (global optimization)} \]

SMC do not apply to this problem as it requires \( E_n = E \).
Consider the case where all the target distributions \( \{\pi_n\}_{n \geq 1} \) are defined on \( E_n = E \).

**Examples**
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.)
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)
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)
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 \{\tilde{\pi}_n\}_{n \geq 1} defined on \(E_n = E^n\) such that

\[
\int \tilde{\pi}_n (x_{1:n-1}, x_n) \, dx_{1:n-1} = \pi_n (x_n)
\]

and apply standard SMC.
Consider a new sequence of *artificial* distributions \( \{ \tilde{\pi}_n \}_{n \geq 1} \) defined on \( E_n = E^n \) such that

\[
\int \tilde{\pi}_n (x_{1:n-1}, x_n) \, dx_{1:n-1} = \pi_n (x_n)
\]

and apply standard SMC.

**Example:**

\[
\tilde{\pi}_n (x_{1:n-1}, x_n) = \pi_n (x_n) \tilde{\pi}_n (x_{1:n-1} \mid x_n)
\]

where \( \tilde{\pi}_n (x_{1:n-1} \mid x_n) \) is *any* conditional distribution on \( E^{n-1} \).

How to design \( \tilde{\pi}_n \) optimally will be discussed later.
Consider a new sequence of artificial distributions \( \{ \tilde{\pi}_n \}_{n \geq 1} \) defined on \( E_n = E^n \) such that

\[
\int \tilde{\pi}_n (x_{1:n-1}, x_n) \, dx_{1:n-1} = \pi_n (x_n)
\]

and apply standard SMC.

**Example:**

\[
\tilde{\pi}_n (x_{1:n-1}, x_n) = \pi_n (x_n) \tilde{\pi}_n (x_{1:n-1} | x_n)
\]

where \( \tilde{\pi}_n (x_{1:n-1} | x_n) \) is any conditional distribution on \( E^{n-1} \).

How to design \( \tilde{\pi}_n \) optimally will be discussed later.
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.
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

\[ E_\pi [\varphi] = \int \varphi (x) \pi (dx) \]

where

\[ \pi (x) = \frac{\gamma (x)}{Z} \]

with \( \gamma \) known pointwise/\( Z = \int \gamma (x) dx \) unknown.

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

\[ \hat{\pi} (dx) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}} (dx). \]
Marginalization is straightforward. If \( x = (x_1, \ldots, x_k) \)

\[
\hat{\pi}(dx_p) = \int \hat{\pi}(dx_{1:p-1}, dx_{p+1:k}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X_p^{(i)}}(dx).
\]
• **Marginalization is straightforward.** If $x = (x_1, \ldots, x_k)$

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

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

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

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

A.D. () 20 / 26
Marginalization is straightforward. If \( x = (x_1, \ldots, x_k) \)

\[
\hat{\pi} (dx_p) = \int \hat{\pi} (dx_1:p-1, dx_{p+1:k}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}} (dx).
\]

Integration is straightforward. Monte Carlo estimates of \( \mathbb{E}_{\pi} (\varphi) \)

\[
\mathbb{E}_{\hat{\pi}} (\varphi) = \int \varphi (x) \hat{\pi} (dx) = \frac{1}{N} \sum_{i=1}^{N} \varphi (X^{(i)}).
\]

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

\[ \mathbb{E} [ \mathbb{E}_{\tilde{\pi}} (\varphi)] = \mathbb{E}_{\pi} (\varphi) \text{ unbiased}, \]
\[ \mathbb{V} [ \mathbb{E}_{\tilde{\pi}} (\varphi)] = \frac{1}{N} \mathbb{E}_{\pi} \left( (\varphi - \mathbb{E}_{\pi} (\varphi))^2 \right) \]
Basic results

\[ E[\hat{E}_\pi(\phi)] = E\pi(\phi) \text{ unbiased}, \]
\[ \text{Var}[\hat{E}_\pi(\phi)] = \frac{1}{N} E\pi((\phi - E\pi(\phi))^2) \]

Rate of convergence to zero \textbf{INDEPENDENT} of space \( E \)! It breaks the curse of dimensionality... sometimes.
Basic results

\[ \mathbb{E} [ \mathbb{E}_{\hat{\pi}} (\varphi)] = \mathbb{E}_{\pi} (\varphi) \text{ unbiased}, \]

\[ \mathbb{V} [ \mathbb{E}_{\hat{\pi}} (\varphi)] = \frac{1}{N} \mathbb{E}_{\pi} \left( (\varphi - \mathbb{E}_{\pi} (\varphi))^2 \right) \]

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

Central limit theorem

\[ \sqrt{N} (\mathbb{E}_{\hat{\pi}} (\varphi) - \mathbb{E}_{\pi} (\varphi)) \Rightarrow \mathcal{N} \left( 0, \mathbb{E}_{\pi} \left( (\varphi - \mathbb{E}_{\pi} (\varphi))^2 \right) \right) \]
Basic results

\[ E \left[ E_{\tilde{\pi}} (\phi) \right] = E_{\pi} (\phi) \text{ unbiased}, \]

\[ \nabla \left[ E_{\tilde{\pi}} (\phi) \right] = \frac{1}{N} E_{\pi} \left( (\phi - E_{\pi} (\phi))^2 \right) \]

Rate of convergence to zero \textbf{INDEPENDENT} of space \( E \)! It breaks the curse of dimensionality... sometimes.

Central limit theorem

\[ \sqrt{N} \left( E_{\tilde{\pi}} (\phi) - E_{\pi} (\phi) \right) \Rightarrow \mathcal{N} \left( 0, E_{\pi} \left( (\phi - E_{\pi} (\phi))^2 \right) \right) \]

Problem: how do you obtain samples from an arbitrary high dimensional distribution???
Basic results

\[
\mathbb{E} [\mathbb{E}_{\hat{\pi}} (\varphi)] = \mathbb{E}_{\pi} (\varphi) \text{ unbiased},
\]

\[
\mathbb{V} [\mathbb{E}_{\hat{\pi}} (\varphi)] = \frac{1}{N} \mathbb{E}_{\pi} \left( (\varphi - \mathbb{E}_{\pi} (\varphi))^2 \right)
\]

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

Central limit theorem

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

**Problem**: how do you obtain samples from an arbitrary 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.

- 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.

  \[ Z \pi(x) K(y|x) \, dx = \pi(y) \]

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

  Alternative (not that bright) idea: Importance sampling. Non-iterative, can be understood in one minute.
Sampling from standard distributions (Gaussian, Gamma, Poisson...) can be done exactly (see articles by Germans) using inverse methods, 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.

$$\pi(x) = \int \pi(y) \text{d}x = \pi(y).$$

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

Alternative (not that bright) idea: Importance sampling.

Non-iterative, can be understood in one minute.
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 ⇒ Non iterative, can be understood in one minute.
**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 (IS) identity. For any distribution $q$ such that $\pi(x) > 0 \Rightarrow q(x) > 0$

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

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

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

$$X^{(i)} \overset{\text{i.i.d.}}{\sim} q(\cdot) \Rightarrow \hat{q}(dx) = \frac{1}{N} \sum_{i=1}^{N} \delta_{X^{(i)}}(dx)$$
Plugging this expression in IS identity

\[
\hat{\pi}(dx) = \frac{w(x) \hat{q}(dx)}{\int w(x) \hat{q}(dx)} = \frac{N^{-1} \sum_{i=1}^{N} w(X^{(i)}) \delta_{X^{(i)}}(dx)}{N^{-1} \sum_{i=1}^{N} w(X^{(i)})} = \sum_{i=1}^{N} W^{(i)} \delta_{X^{(i)}}(dx)
\]

where

\[W^{(i)} \propto w(X^{(i)}) \text{ and } \sum_{i=1}^{N} W^{(i)} = 1.\]
Plugging this expression in IS identity

\[
\hat{\pi}(dx) = \frac{w(x) \hat{q}(dx)}{\int w(x) \hat{q}(dx)} = \frac{N^{-1} \sum_{i=1}^{N} w(X^{(i)}) \delta_{X^{(i)}}(dx)}{N^{-1} \sum_{i=1}^{N} w(X^{(i)})} \\
= \sum_{i=1}^{N} W^{(i)} \delta_{X^{(i)}}(dx)
\]

where

\[
W^{(i)} \propto w(X^{(i)}) \quad \text{and} \quad \sum_{i=1}^{N} W^{(i)} = 1.
\]

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

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

Statistics for $N \gg 1$

$$
\mathbb{E} [\mathbb{E}_{\hat{\pi}} [\varphi]] = \mathbb{E}_\pi [\varphi] - N^{-1}_\pi \mathbb{E} \left[ W (X) (\varphi (X) - \mathbb{E}_\pi [\varphi]) \right],
$$

with negligible bias.

$$
\text{Var} [\mathbb{E}_{\hat{\pi}} [\varphi]] = N^{-1}_\pi \mathbb{E} \left[ W (X) (\varphi (X) - \mathbb{E}_\pi [\varphi])^2 \right].
$$
• Now we can approximate $\mathbb{E}_\pi [\varphi]$ by

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

• Statistics for $N \gg 1$

$$
\mathbb{E} [\mathbb{E}_\hat{\pi} [\varphi]] = \mathbb{E}_\pi [\varphi] - \frac{1}{N_{\pi}} \mathbb{E} [W(X)(\varphi(X) - \mathbb{E}_\pi [\varphi])] \text{, negligible bias},
$$

$$
\mathbb{V} [\mathbb{E}_\hat{\pi} [\varphi]] = N_{\pi}^{-1} \mathbb{E} [W(X)(\varphi(X) - \mathbb{E}_\pi [\varphi])^2].
$$

• Estimate of normalizing constant

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

and $\mathbb{E} [\hat{Z}] = Z$, $\mathbb{V} [\hat{Z}] = N^{-1} \left( \mathbb{E}_q \left[ \left( \frac{\gamma(X)}{q(X)} - Z \right)^2 \right] \right)$.
For a given $\varphi$, importance distribution minimizing $\mathbb{V} [\mathbb{E}_\pi [\varphi]]$ is

$$q^{\text{opt}} (x) = \frac{\left| \varphi (x) - \mathbb{E}_\pi [\varphi] \right| \pi (x)}{\int \left| \varphi (x) - \mathbb{E}_\pi [\varphi] \right| \pi (x) \, dx}.$$
For a given $\varphi$, importance distribution minimizing $\nabla \left[ \mathbb{E}_{\pi} \left[ \varphi \right] \right]$ is

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

Useless as sampling from $q^{\text{opt}}$ as complex as solving the original problem.

Useless as sampling from $q^{\text{opt}}$ as complex as solving the original problem.
For a given $\phi$, importance distribution minimizing $V[\mathbb{E}_{\pi}[\phi]]$ is

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

- Useless as sampling from $q^{\text{opt}}$ as complex as solving the original problem.
- In applications we are interested in, there is typically no specific $\phi$ of interest.
For a given $\varphi$, importance distribution minimizing $V[\mathbb{E}_\pi[\varphi]]$ is

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

- Useless as sampling from $q^{\text{opt}}$ as complex as solving the original problem.
- In applications we are interested in, there is typically no specific $\varphi$ of interest.
- Practical recommendations
For a given $\varphi$, importance distribution minimizing $V[\mathbb{E}_{\tilde{\pi}}[\varphi]]$ is

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

Useless as sampling from $q^{\text{opt}}$ as complex as solving the original problem.

In applications we are interested in, there is typically no specific $\varphi$ of interest.

Practical recommendations

- Select $q$ as close to $\pi$ as possible.
For a given $\varphi$, importance distribution minimizing $V [\mathbb{E}_\pi [\varphi]]$ is

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

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

In applications we are interested in, there is typically no specific $\varphi$ of interest.

**Practical recommendations**

- Select $q$ as close to $\pi$ as possible.
- Ensure

$$w (x) = \frac{\pi (x)}{q (x)} < \infty.$$
For a given \( \varphi \), importance distribution minimizing \( \nabla \mathbb{E}_\pi [\varphi] \) is

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

Useless as sampling from \( q^{\text{opt}} \) as complex as solving the original problem.

In applications we are interested in, there is typically no specific \( \varphi \) of interest.

Practical recommendations

- Select \( q \) as close to \( \pi \) as possible.
- Ensure

\[
\omega (x) = \frac{\pi (x)}{q (x)} < \infty.
\]

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