Adaptive Multiple Importance Sampling

Pierre Pudlo
⟨pierre.pudlo@univ-amu.fr⟩

Aix-Marseille Université
Faculté des Sciences
Institut de Mathématiques de Marseille (I2M)

02 / 29 / 2016

Table of Contents

Aim

Approximate a **target** distribution $\Pi(dx) = \pi(x)dx$ with a weighted Monte Carlo sample:

$$\Pi \approx \frac{1}{N} \sum_{i=1}^{N} w_i \delta_{x_i}$$

by sampling from an **instrumental** distribution $Q(dx) = q(x)\,dx$:

$$x_i \sim^{\text{iid}} Q \quad \text{and} \quad w_i = w(x_i) = \pi(x_i) \Big/ q(x_i)$$

Aim

Approximate a **target** distribution $\Pi(dx) = \pi(x)dx$ with a weighted Monte Carlo sample:

$$\Pi \approx \frac{1}{N} \sum_{i=1}^{N} w_i \delta_{x_i}$$

by sampling from an **instrumental** distribution $Q(dx) = q(x)\,dx$:

$$x_i \sim^{\text{iid}} Q \quad \text{and} \quad w_i = w(x_i) = \pi(x_i) \Big/ q(x_i)$$

- Approximating the target means that, for a large class of function ψ,

$$\int \psi(x)\Pi(dx) \approx \frac{1}{N} \sum_{i=1}^{N} w_i \ \psi(x_i)$$

Aim

Approximate a **target** distribution $\Pi(dx) = \pi(x)dx$ with a weighted Monte Carlo sample:

$$\Pi \approx \frac{1}{N}\sum_{i=1}^{N} w_i \delta_{x_i}$$

by sampling from an **instrumental** distribution $Q(dx) = q(x)\,dx$:

$$x_i \overset{\text{iid}}{\sim} Q \quad \text{and} \quad w_i = w(x_i) = \pi(x_i) \Big/ q(x_i)$$

- Approximating the target means that, for a large class of function ψ,

$$\int \psi(x)\Pi(dx) \approx \frac{1}{N}\sum_{i=1}^{N} w_i\ \psi(x_i)$$

- If $\Pi(dx) \ll Q(dx)$, the approximation is **unbiased**:

$$\int \psi(x)\pi(x)\,dx = \int \psi(x)\frac{\pi(x)}{q(x)}q(x)\,dx$$

Accuracy depends heavily on the spread of the w_i's:

1. if $w_1 = \mathcal{O}(N)$ and $w_2 \ll 1, \ldots, w_N \ll 1$, then
$$\frac{1}{N}\sum_{i=1}^{N} w_i \ \psi(x_i) \approx \frac{w_1}{N}\psi(x_1)$$
$\implies$ same accuracy as a Monte Carlo sample of size 1

Accuracy depends heavily on the spread of the w_i's:

1. if $w_1 = \mathcal{O}(N)$ and $w_2 \ll 1, \ldots, w_N \ll 1$, then
$$\frac{1}{N}\sum_{i=1}^{N} w_i\ \psi(x_i) \approx \frac{w_1}{N}\psi(x_1)$$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size 1

2. if $Q = \Pi$, then $w_1 = \ldots = w_N = 1$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size N

Accuracy depends heavily on the spread of the w_i's:

1. if $w_1 = \mathcal{O}(N)$ and $w_2 \ll 1, \ldots, w_N \ll 1$, then
$$\frac{1}{N}\sum_{i=1}^{N} w_i \; \psi(x_i) \approx \frac{w_1}{N}\psi(x_1)$$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size 1

2. if $Q = \Pi$, then $w_1 = \ldots = w_N = 1$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size N

Effective Sample Size

$$\text{ESS} = \left(\sum_{i=1}^{N} w_i\right)^2 \Big/ \sum_{i=1}^{N} w_i^2$$

Accuracy depends heavily on the spread of the w_i's:

1. if $w_1 = \mathcal{O}(N)$ and $w_2 \ll 1, \dots, w_N \ll 1$, then
$$\frac{1}{N}\sum_{i=1}^{N} w_i\ \psi(x_i) \approx \frac{w_1}{N}\psi(x_1)$$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size 1

2. if $Q = \Pi$, then $w_1 = \dots = w_N = 1$
$\Longrightarrow$ same accuracy as a Monte Carlo sample of size N

Effective Sample Size

$$\text{ESS} = \left(\sum_{i=1}^{N} w_i\right)^2 \Bigg/ \sum_{i=1}^{N} w_i^2$$

1. if $w_1 = \mathcal{O}(N)$ and $w_2 \ll 1, \dots, w_N \ll 1$, then ESS ≈ 1
2. if $w_1 = \dots = w_N = 1$, then ESS $= N$

Table of Contents

Multiple Importance Sampling

At our disposal: T instrumental distributions $Q^t(dx) = q^t(x)dx$, $t = 1, \ldots, T$

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$\begin{array}{ccc} x_1^1, \ldots, x_{N_1}^1 & \stackrel{\text{iid}}{\sim} q^1(x)dx & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\ \vdots & \vdots & \vdots \\ x_1^T, \ldots, x_{N_T}^T & \stackrel{\text{iid}}{\sim} q^T(x)dx & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T) \end{array}$$

Multiple Importance Sampling

At our disposal: T instrumental distributions $Q^t(dx) = q^t(x)dx$, $t = 1, \ldots, T$

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$\begin{array}{ccc} x_1^1, \ldots, x_{N_1}^1 & \overset{\text{iid}}{\sim} q^1(x)dx & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\ \vdots & \vdots & \vdots \\ x_1^T, \ldots, x_{N_T}^T & \overset{\text{iid}}{\sim} q^T(x)dx & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T) \end{array}$$

- Merge weighted samples: $\Pi \approx \frac{1}{\Omega_T} \sum_{t=1}^{T} \sum_{i=1}^{N_T} w_i^t \delta_{x_i^t}$

Multiple Importance Sampling

At our disposal: T instrumental distributions $Q^t(dx) = q^t(x)dx$, $t = 1, \ldots, T$

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$
\begin{array}{ccc}
x_1^1, \ldots, x_{N_1}^1 & \sim^{\text{iid}} q^1(x)dx & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\
\vdots & \vdots & \vdots \\
x_1^T, \ldots, x_{N_T}^T & \sim^{\text{iid}} q^T(x)dx & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T)
\end{array}
$$

- Merge weighted samples: $\Pi \approx \dfrac{1}{\Omega_T} \displaystyle\sum_{t=1}^{T} \sum_{i=1}^{N_T} w_i^t \delta_{x_i^t}$
- Is still unbiased

At our disposal: T instrumental distributions $Q^t(dx) = q^t(x)dx$, $t = 1, \ldots, T$

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$\begin{array}{ccc} x_1^1, \ldots, x_{N_1}^1 \sim^{\text{iid}} q^1(x)dx & & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\ \vdots \qquad\quad \vdots & & \vdots \\ x_1^T, \ldots, x_{N_T}^T \sim^{\text{iid}} q^T(x)dx & & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T) \end{array}$$

- Merge weighted samples: $\Pi \approx \dfrac{1}{\Omega_T} \displaystyle\sum_{t=1}^{T} \sum_{i=1}^{N_T} w_i^t \delta_{x_i^t}$
- Is still unbiased
- But, if one weight is much larger than all the others, merging does not solve the issue

Multiple Importance Sampling

At our disposal: T instrumental distributions $Q^t(dx) = q^t(x)dx$, $t = 1, \ldots, T$

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$
\begin{array}{lll}
x_1^1, \ldots, x_{N_1}^1 & \overset{\text{iid}}{\sim} q^1(x)dx & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\
\vdots & \vdots & \vdots \\
x_1^T, \ldots, x_{N_T}^T & \overset{\text{iid}}{\sim} q^T(x)dx & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T)
\end{array}
$$

- Merge weighted samples: $\Pi \approx \dfrac{1}{\Omega_T} \displaystyle\sum_{t=1}^{T} \sum_{i=1}^{N_T} w_i^t \delta_{x_i^t}$
- Is still unbiased
- But, if one weight is much larger than all the others, merging does not solve the issue

Basic merging inherits property of the worst instrumental distribution among $Q^1, \ldots, Q^T$.

Multiple Importance Sampling (2)

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$
\begin{array}{ccccc}
x_1^1, \ldots, x_{N_1}^1 & \stackrel{\text{iid}}{\sim} & q^1(x)dx & \quad \text{and } & w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\
\vdots & & \vdots & & \vdots \\
x_1^T, \ldots, x_{N_T}^T & \stackrel{\text{iid}}{\sim} & q^T(x)dx & \quad \text{and } & w_i^T = \pi(x_i^T)/q^T(x_i^T)
\end{array}
$$

- Interpret all x_i^t as drawn from the mixture $q_{\text{mixt}}(x) = \sum_{t=1}^{T} \frac{N_t}{\Omega_T} q^t(x)$

  & replace all weights with $\widetilde{w}_i^t = \pi(x_i^t) \big/ q_{\text{mixt}}(x_i^t)$

Multiple Importance Sampling (2)

Several instrumental distributions

$\Omega_T = N_1 + \ldots + N_T$ simulations from T instrumental distributions:

$$\begin{array}{ccc} x_1^1, \ldots, x_{N_1}^1 \stackrel{\text{iid}}{\sim} q^1(x)dx & & \text{and } w_i^1 = \pi(x_i^1)/q^1(x_i^1) \\ \vdots \quad\quad \vdots & & \vdots \\ x_1^T, \ldots, x_{N_T}^T \stackrel{\text{iid}}{\sim} q^T(x)dx & & \text{and } w_i^T = \pi(x_i^T)/q^T(x_i^T) \end{array}$$

- Interpret all x_i^t as drawn from the mixture $q_{\text{mixt}}(x) = \sum_{t=1}^{T} \frac{N_t}{\Omega_T} q^t(x)$

  & replace all weights with $\widetilde{w}_i^t = \pi(x_i^t) \big/ q_{\text{mixt}}(x_i^t)$

- Stabilises the approximation by reducing the variance of the weights
  & remains unbiased

[Veach and Guibas (1995); Owen and Zhou (2000)]

Multiple Importance Sampling (3)

Why does the above trick stabilize the approximation?

- $w_i^t = \pi(x_i^t)/q^t(x_i^t)$ is large when $q^t(x_i^t) \ll \pi(x_i^t)$

Multiple Importance Sampling (3)

Why does the above trick stabilize the approximation?

- $w_i^t = \pi(x_i^t)/q^t(x_i^t)$ is large when $q^t(x_i^t) \ll \pi(x_i^t)$
- which means that x_i^t is in the tail of q^t and
  1. either x_i^t is not in the tail of the target Π
  2. or Π has larger tails than the instrumental Q^t

Multiple Importance Sampling (3)

Why does the above trick stabilize the approximation?

- $w_i^t = \pi(x_i^t)/q^t(x_i^t)$ is large when $q^t(x_i^t) \ll \pi(x_i^t)$
- which means that x_i^t is in the tail of q^t and
  1. either x_i^t is not in the tail of the target Π
  2. or Π has larger tails than the instrumental Q^t

- The mixture distribution Q_{mixt} of density $q_{\text{mixt}}(x) = \sum_{t=1}^{T} \frac{N_t}{\Omega_T} q^t(x)$:
  1. has relatively high density as soon as one of the instrumentals has relatively high density
  2. has tails which decrease as the instrumental of largest tails.

Multiple Importance Sampling (3)

Why does the above trick stabilize the approximation?

- $w_i^t = \pi(x_i^t)/q^t(x_i^t)$ is large when $q^t(x_i^t) \ll \pi(x_i^t)$
- which means that x_i^t is in the tail of q^t and
  1. either x_i^t is not in the tail of the target Π
  2. or Π has larger tails than the instrumental Q^t

- The mixture distribution Q_{mixt} of density $q_{\text{mixt}}(x) = \sum_{t=1}^{T} \frac{N_t}{\Omega_T} q^t(x)$:
  1. has relatively high density as soon as one of the instrumentals has relatively high density
  2. has tails which decrease as the instrumental of largest tails.

The clever merging with "mixture" weights inherits properties of the best instrumental distributions among $Q^1, \dots, Q^T$.

Table of Contents

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\, \pi(x) dx$, where h is known.

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1}\displaystyle\sum_{i=1}^{N_1} w_i^1 h(x_i^1)$

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \dots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \dots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1}\sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \ldots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_2}\sum_{i=1}^{N_2} w_i^2 h(x_i^2)$

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \ldots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_2} \sum_{i=1}^{N_2} w_i^2 h(x_i^2)$
- Draw a third sample ...

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \displaystyle\sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \ldots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{N_2} w_i^2 h(x_i^2)$
- Draw a third sample ...

→ Return the last sample $\dfrac{1}{N_T} \displaystyle\sum_{i=1}^{N_T} w_i^T \delta_{x_i^T}$

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\, \pi(x) dx$, where h is known.

- Draw a first sample from $x_1^1, \ldots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \displaystyle\sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \ldots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{N_2} w_i^2 h(x_i^2)$
- Draw a third sample ...

→ Return the last sample $\dfrac{1}{N_T} \displaystyle\sum_{i=1}^{N_T} w_i^T \delta_{x_i^T}$

Can we do better with merging?

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\,\pi(x)dx$, where h is known.

- Draw a first sample from $x_1^1, \dots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \dots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_1 + N_2} \sum_{t=1}^{2} \sum_{i=1}^{N_t} \widetilde{w}_i^t h(x_i^t)$, with "mixture" weights
- Draw a third sample ...

→ Return the whole sample $\dfrac{1}{\Omega_T} \sum_{t=1}^{T} \sum_{i=1}^{N_t} \widetilde{w}_i^t \delta_{x_i^t}$, with "mixture" weights

[Cornuet, Marin, Mira, Robert (2012)]

Adaptive Multiple Importance Sampling (2)

- AMIS uses a clever recycling strategy ("mixture" weights)
  1. at the end of the t-iteration to adapting θ:

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  2. at the end of the algorithm to compute the final weights of the output

$$\frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

Adaptive Multiple Importance Sampling (2)

- AMIS uses a clever recycling strategy ("mixture" weights)
  1. at the end of the t-iteration to adapting θ:

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  2. at the end of the algorithm to compute the final weights of the output

$$\frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

- AMIS has good numerical properties, see, e.g.,
  - Cornuet, Marin, Mira and Robert (2012)
  - Sirén, Marttinen and Corander (2011)
  - Šmídl and Hofman (2013)
  - Bugallo, Martino and Corander (2015)
  - Martino, Elvira, Luengo and Corander (2015)
  - ...

Adaptive Multiple Importance Sampling (2)

- AMIS uses a clever recycling strategy ("mixture" weights)
  1. at the end of the t-iteration to adapting θ:

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  2. at the end of the algorithm to compute the final weights of the output

$$\frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

- AMIS has good numerical properties, see, e.g.,
  - Cornuet, Marin, Mira and Robert (2012)
  - Sirén, Marttinen and Corander (2011)
  - Šmídl and Hofman (2013)
  - Bugallo, Martino and Corander (2015)
  - Martino, Elvira, Luengo and Corander (2015)
  - ...

- **But no proof of AMIS' consistency**

Why is it difficult?

- At time t, θ is adapted with

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

Why is it difficult?

- At time t, θ is adapted with

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

➔ $\widehat{\theta}_{t+1}$ depends on the whole set of simulations

Why is it difficult?

- At time t, θ is adapted with

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  - ➔ $\widehat{\theta}_{t+1}$ depends on the whole set of simulations
  - ➔ the weight $\widetilde{w}_1^1$ of the first x_1^1 in $\widehat{\theta}_{t+1}$ depends on the whole set of simulations via $\widehat{\theta}_s$, $s = 1, \ldots, t$

Why is it difficult?

- At time t, θ is adapted with

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  - ➔ $\widehat{\theta}_{t+1}$ depends on the whole set of simulations
  - ➔ the weight $\widetilde{w}_1^1$ of the first x_1^1 in $\widehat{\theta}_{t+1}$ depends on the whole set of simulations via $\widehat{\theta}_s$, $s = 1, \ldots, t$
  - ➔ cannot even compute $\mathbb{E}\left(\widehat{\theta}_{t+1}\right)$ and study the bias $\mathbb{E}\left(\widehat{\theta}_{t+1}\right) - \theta^*$

Why is it difficult?

- At time t, θ is adapted with

$$\widehat{\theta}_{t+1} = \frac{1}{\Omega_t} \sum_{s=1}^{t} \sum_{i=1}^{N_s} \widetilde{w}_i^s h(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{t} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r)$$

  - → $\widehat{\theta}_{t+1}$ depends on the whole set of simulations
  - → the weight $\widetilde{w}_1^1$ of the first x_1^1 in $\widehat{\theta}_{t+1}$ depends on the whole set of simulations via $\widehat{\theta}_s$, $s = 1, \ldots, t$
  - → cannot even compute $\mathbb{E}\left(\widehat{\theta}_{t+1}\right)$ and study the bias $\mathbb{E}\left(\widehat{\theta}_{t+1}\right) - \theta^*$

- Same issues with the output
  - → cannot even study the bias between

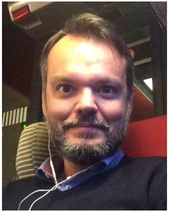
$$\frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_T} \widetilde{w}_i^T \psi(x_i^s) \quad \text{and} \quad \int \psi(x)\pi(x)\,dx$$

on test functions ψ

Table of Contents

Joint work with

Jean-Michel **Marin** (U. Montpellier)



& Mohammed **Sedki** (U. Paris Sud)

Consistency of Adaptive Importance Sampling and Recycling Schemes,
http://arxiv.org/abs/1211.2548

Modified Adaptive Multiple Importance Sampling

Adaptive

A parametrized family of distributions: $\{Q(\theta),\ \theta \in \Theta\}$

& adapt the instrumental distribution sequentially by fitting moments.

Targeted instrumental distribution $\theta^* = \int h(x)\pi(x)\ dx$, where h is known.

- Draw a first sample from $x_1^1, \dots, x_{N_1}^1$ from $Q(\widehat{\theta}_1)$ where $\widehat{\theta}_1$ is a first guess
- Adapt θ with $\widehat{\theta}_2 = \dfrac{1}{N_1} \displaystyle\sum_{i=1}^{N_1} w_i^1 h(x_i^1)$
- Draw a second sample $x_1^2, \dots, x_{N_2}^2$ from $Q(\widehat{\theta}_2)$
- Adapt θ with $\widehat{\theta}_3 = \dfrac{1}{N_2} \displaystyle\sum_{i=1}^{N_2} w_i^2 h(x_i^2)$ (no recycling here)
- Draw a third sample ...

→ Return the whole sample $\dfrac{1}{\Omega_T} \displaystyle\sum_{t=1}^{T} \sum_{i=1}^{N_t} \widetilde{w}_i^t \delta_{x_i^t}$, with "mixture" weights

Modified Adaptive Multiple Importance Sampling (2)

- MAMIS uses a clever recycling strategy ("mixture" weights) only at the end of the algorithm to compute the weights of the output:

$$\frac{1}{\Omega_T}\sum_{s=1}^{T}\sum_{i=1}^{N_s}\widetilde{w}_i^s\delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s)\Big/\sum_{r=1}^{T}\frac{N_r}{\Omega_t}q(x_i^s,\widehat{\theta}_r)$$

- But adapt θ naively:

$$\widehat{\theta}_{t+1} = \frac{1}{N_t}\sum_{i=1}^{N_t} w_i^t h(x_i^t) \quad \text{where } w_i^t = \pi(x_i^t)/q(x_i^t,\widehat{\theta}_t).$$

Modified Adaptive Multiple Importance Sampling (2)

- MAMIS uses a clever recycling strategy ("mixture" weights) only at the end of the algorithm to compute the weights of the output:

$$\frac{1}{\Omega_T}\sum_{s=1}^{T}\sum_{i=1}^{N_s}\widetilde{w}_i^s\delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s)\Big/\sum_{r=1}^{T}\frac{N_r}{\Omega_t}q(x_i^s,\widehat{\theta}_r)$$

- But adapt θ naively:

$$\widehat{\theta}_{t+1} = \frac{1}{N_t}\sum_{i=1}^{N_t} w_i^t h(x_i^t) \quad \text{where } w_i^t = \pi(x_i^t)/q(x_i^t,\widehat{\theta}_t).$$

- MAMIS has almost the same good numerical properties, see references below

Modified Adaptive Multiple Importance Sampling (2)

- MAMIS uses a clever recycling strategy ("mixture" weights) only at the end of the algorithm to compute the weights of the output:

$$\frac{1}{\Omega_T}\sum_{s=1}^{T}\sum_{i=1}^{N_s}\widetilde{w}_i^s\delta_{x_i^s} \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s)\Big/\sum_{r=1}^{T}\frac{N_r}{\Omega_t}q(x_i^s,\widehat{\theta}_r)$$

- But adapt θ naively:

$$\widehat{\theta}_{t+1} = \frac{1}{N_t}\sum_{i=1}^{N_t} w_i^t h(x_i^t) \quad \text{where } w_i^t = \pi(x_i^t)/q(x_i^t,\widehat{\theta}_t).$$

- MAMIS has almost the same good numerical properties, see references below

- **MAMIS is much simple to study**

Table of Contents

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

  - Has been used by Douc, Guillin, Marin, Robert (2007) to prove consistency
  - The proof is sequential: if $\widehat{\theta}_t \to \theta^*$ at time t, does $\widehat{\theta}_{t+1} \to \theta^*$?
  - Does not indicate how Monte Carlo errors accumulate (or not) over time

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

  - Has been used by Douc, Guillin, Marin, Robert (2007) to prove consistency
  - The proof is sequential: if $\widehat{\theta}_t \to \theta^*$ at time t, does $\widehat{\theta}_{t+1} \to \theta^*$?
  - Does not indicate how Monte Carlo errors accumulate (or not) over time

- Instead we assume that

$$N_1, N_2, \ldots \text{ fixed}, \qquad T \to \infty$$

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

  - Has been used by Douc, Guillin, Marin, Robert (2007) to prove consistency
  - The proof is sequential: if $\widehat{\theta}_t \to \theta^*$ at time t, does $\widehat{\theta}_{t+1} \to \theta^*$?
  - Does not indicate how Monte Carlo errors accumulate (or not) over time

- Instead we assume that

$$N_1, N_2, \ldots \text{ fixed}, \qquad T \to \infty$$

  - Models the situation where we add iterations over time until being happy with the output

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

  - Has been used by Douc, Guillin, Marin, Robert (2007) to prove consistency
  - The proof is sequential: if $\widehat{\theta}_t \to \theta^*$ at time t, does $\widehat{\theta}_{t+1} \to \theta^*$?
  - Does not indicate how Monte Carlo errors accumulate (or not) over time

- Instead we assume that

$$N_1, N_2, \ldots \text{ fixed}, \qquad T \to \infty$$

  - Models the situation where we add iterations over time until being happy with the output
  - Is more difficult to study because, at time t, we have a value $\widehat{\theta}_t$ that comes from a finite sample (of fixed size)

The asymptotic framework

- A first asymptotic framework we do not use is

$$N_1 = N_2 = \cdots = N_T = N, \qquad T \text{ fixed}, \qquad N \to \infty$$

  - Has been used by Douc, Guillin, Marin, Robert (2007) to prove consistency
  - The proof is sequential: if $\widehat{\theta}_t \to \theta^*$ at time t, does $\widehat{\theta}_{t+1} \to \theta^*$?
  - Does not indicate how Monte Carlo errors accumulate (or not) over time

- Instead we assume that

$$N_1, N_2, \ldots \text{ fixed}, \qquad T \to \infty$$

  - Models the situation where we add iterations over time until being happy with the output
  - Is more difficult to study because, at time t, we have a value $\widehat{\theta}_t$ that comes from a finite sample (of fixed size)

- We also assume that $N_t \to \infty$ when $t \to \infty$.

(H1) $\sum_{t=1}^{\infty} 1/N_t$ is finite

(H2) $\int \|h(x)\|^2 \frac{\pi(x)}{q(x,\theta)} \; \pi(x)dx$ is finite for all θ and depends continuously on θ

(H1) $\sum_{t=1}^{\infty} 1/N_t$ is finite

(H2) $\int \|h(x)\|^2 \frac{\pi(x)}{q(x,\theta)} \pi(x)dx$ is finite for all θ and depends continuously on θ

Theorem 1

Under (H1) and (H2), when $T \to \infty$, $\lim \widehat{\theta}_T = \theta^*$ **almost surely**

Consistency of the learning scheme

(H1) $\sum_{t=1}^{\infty} 1/N_t$ is finite

(H2) $\int \|h(x)\|^2 \frac{\pi(x)}{q(x,\theta)} \pi(x)dx$ is finite for all θ and depends continuously on θ

Theorem 1

Under (H1) and (H2), when $T \to \infty$, $\lim \widehat{\theta}_T = \theta^*$ **almost surely**

Remark 1. Almost sure convergence is needed to deal with

$$q^T_{\text{mixt}}(x) = \sum_{t=1}^{T} \frac{N_t}{\Omega_T} q\big(x, \widehat{\theta}_t\big)$$

because it depends on the path $\widehat{\theta}_1, \ldots, \widehat{\theta}_T$

Consistency of the learning scheme

(H1) $\sum_{t=1}^{\infty} 1/N_t$ is finite

(H2) $\int \|h(x)\|^2 \frac{\pi(x)}{q(x,\theta)}\, \pi(x)dx$ is finite for all θ and depends continuously on θ

Theorem 1

Under (H1) and (H2), when $T \to \infty$, $\lim \widehat{\theta}_T = \theta^*$ **almost surely**

Remark 2. $\widehat{\theta}_{t+1}$ is an average over a new sample when compared to $\widehat{\theta}_t$

$\Longrightarrow$ A price to pay to get almost sure convergence.
Here L^2 instead of L^1, see (H2)

Theorem 2

Assume that $\sum 1/N_t$ is finite, and that $\widehat{\theta}_T \to \theta^*$ almost surely.
Let

$$\widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \psi(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r).$$

Then, when $T \to \infty$, over a large class of functions ψ,

$$\lim \widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \int \psi(x)\pi(x)dx \quad \text{almost surely.}$$

Theorem 2

Assume that $\sum 1/N_t$ is finite, and that $\widehat{\theta}_T \to \theta^*$ almost surely.
Let

$$\widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \psi(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r).$$

Then, when $T \to \infty$, over a large class of functions ψ,

$$\lim \widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \int \psi(x)\pi(x)dx \quad \text{almost surely.}$$

Theorem 2

Assume that $\sum 1/N_t$ is finite, and that $\widehat{\theta}_T \to \theta^*$ almost surely.
Let

$$\widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \frac{1}{\Omega_T} \sum_{s=1}^{T} \sum_{i=1}^{N_s} \widetilde{w}_i^s \psi(x_i^s) \quad \text{where } \widetilde{w}_i^s = \pi(x_i^s) \Big/ \sum_{r=1}^{T} \frac{N_r}{\Omega_t} q(x_i^s, \widehat{\theta}_r).$$

Then, when $T \to \infty$, over a large class* of functions ψ,

$$\lim \widehat{\Pi}_T^{\text{MAMIS}}(\psi) = \int \psi(x)\pi(x)dx \quad \text{almost surely.}$$

*The class depends on the tails of the instrumentals and the target

E.g., if $\Pi(dx)$ has Gaussian tails or exponentially decreasing tails, and $Q(dx, \theta)$ has polynomials tails in a neighborhood of θ^*, then every polynomials $\psi(x)$ are in this class.