An auto-validating trans-dimensional von Neumann rejection sampler

Raazesh Sainudiin

Biomathematics Research Centre, Department of Mathematics & Statistics,
University of Canterbury, Christchurch, NZ

[with Tom York, Cornell University, Ithaca, NY, USA]
Outline

• A Statistical Sampling Problem

• Validated Numerics
  • What is it ?
  • How does it work ?

• A Solution to the Sampling Problem
  • Moore Rejection Sampling
  • Examples

• Summary and Extensions
Problem: Sampling From a Density
Sampling from a Density - MCMC (M-H)

Support : $\Theta \ni \theta$  
Target : $p := p^*/N_p$ (Unknown $N_p$)  
Proposal : $q := q^*/N_q$

1. Choose an arbitrary starting point $\theta_0$ and set $i = 0$.
2. Generate a candidate point $\theta' \sim q(\theta_i, \cdot)$ and $u \sim U(0, 1)$.
3. Set:

   $$\theta_{i+1} = \begin{cases} 
   \theta' & \text{if } u \leq \frac{p^*(\theta')q(\theta', \theta_i)}{p^*(\theta_i)q(\theta_i, \theta')} \\
   \theta_i & \text{otherwise}
   \end{cases}$$

4. Set $i = i + 1$ and GO TO 2
Sampling from a Density - Rejection Sampling

Support : $\Theta \ni \theta$  Target : $p := p^*/N_p$ (Unknown $N_p$)  Proposal : $q := q^*/N_q$

Find “envelope function” $f_q(\theta) = cq^*(\theta)$ such that $f_q(\theta) \geq p^*(\theta)$, $\forall \theta \in \Theta$

1. Generate a candidate point $\theta \sim q(\cdot)$.
2. Draw $u \sim U(0, 1)$.
3. If $(u < p^*(\theta)/f_q(\theta))$, then $\theta$ is an exact and independent sample from $p$. DONE.
4. Else GO TO 1
MCMC vs Rejection Sampling

MCMC (extensions of Metropolis-Hastings Chains)

- “Easy” to implement; almost any proposal works – BUT ONLY asymptotically...
- Poor proposal $\Rightarrow$ slow convergence – heuristic proposal ‘tuning’,
- Convergence diagnostics are generally not rigorous – can be misleading.

Rejection Sampling (due to Von Neumann)

- “Hard” to implement; envelope property is NECESSARY – or will NOT sample $p$,
- Poor proposal $\Rightarrow$ low acceptance probability – MUCH $\ll 10^{-10}$,
- Perfectly independent samples – and NO convergence issues.
What makes a density hard to sample?

• Generally:
  1 Complexity – many peaks and valleys – size of $\Theta$.
  2 Curse of dimensionality.

• If ignorant of global behavior of density:
  3 Widely separated peaks (hard to get from one to next),
  4 Narrow peaks on smooth background (hard to find),
  5 Peaks of strange shapes (e.g. Rosenbrock’s banana density)
  6 Others... ?

Don’t treat target density as black box.

• Look at expression (or code) for $p^*(x)$.
• Find locations, widths, shapes of peaks, etc.
• Construct a better proposal.
• Interval arithmetic is way to do this in an auto-validated manner
  • AND we get envelope function.
Trivariate Needle in a Haystack – Heuristic Diagnostics

\[ p^*(x) = \frac{1}{\sigma_1^3} \exp\left\{-\frac{1}{2} \left( \frac{x - \mu_1}{\sigma_1} \right)^2 \right\} + \frac{1}{\sigma_2^3} \exp\left\{-\frac{1}{2} \left( \frac{x - \mu_2}{\sigma_2} \right)^2 \right\} \]

\[ \mu_1 = (0, 0, 0), \mu_2 = (1, 1, 1), \sigma_1 = 1, \sigma_2 = 0.006 \]
**Trivariate Needle in a Haystack – Heuristic Diagnostics**

\[
p^*(x) = \frac{1}{\sigma_1^3} \exp\left\{-\frac{1}{2} \left(\frac{x - \mu_1}{\sigma_1}\right)^2\right\} + \frac{1}{\sigma_2^3} \exp\left\{-\frac{1}{2} \left(\frac{x - \mu_2}{\sigma_2}\right)^2\right\}
\]

\[
\mu_1 = (0, 0, 0), \mu_2 = (1, 1, 1), \sigma_1 = 1, \sigma_2 = 0.006
\]
Some possibilities through

Validated Numerics ...
**Validated Numerics**

What is it?
- set-valued mathematics
- intervals replace real numbers

Why use it?
- provides rigorous error bounds
- naturally models uncertainty in data
- may produce faster numerical methods

Where has it been used recently?

Early work:
**Intervals**

*Notations, Definitions and Features*

\[
\begin{align*}
& \overline{y} - \underline{y} \\
\hline
& \langle X \rangle \\
& m(X) \\
& d(X)
\end{align*}
\]

\[\begin{align*}
\begin{array}{c}
\overline{y} - \underline{y} \\
\hline
\langle X \rangle \\
\end{array} & \mp \begin{array}{c}
\overline{y} - \underline{y} \\
\hline
\langle X \rangle \\
\end{array}
\]

- real: \( x \in \mathbb{R} \)
- (compact) real interval: \( X = [\underline{x}, \overline{x}] = [\inf(X), \sup(X)] \)
- set of all real intervals: \( \mathbb{IR} := \{ [a, b] : a \leq b, a, b \in \mathbb{R} \} \)
  - **Example:** \([1, \pi], 17, \sqrt{2} \in \mathbb{IR}, \) but not \([2, 1] \) or \([1, \infty] \).
- real interval vector or **box:** \( X = (X_1, \cdots, X_n)^T \in \mathbb{IR}^n \), where \( X_i = [\underline{x}_i, \overline{x}_i] \in \mathbb{IR} \), \( 1 \leq i \leq n \)
- a thin interval \( X = [x, x] \) has 0 diameter with \( \underline{x} = \overline{x} = x \Rightarrow \mathbb{R} \subset \mathbb{IR} \)
**Arithmetic Over** \( \mathbb{IR} \)

**Definition.** If \( \circ \) is one of the operators \(+, -, /, \cdot\) and if \( X, Y \in \mathbb{IR} \)

\[
X \circ Y := \{ x \circ y : x \in X, y \in Y \}
\]

*except that* \( X/Y \) *is undefined if* \( 0 \in Y \).

Uncountable many cases to consider!

Continuity, Monotonicity, and Compactness \( \Rightarrow \)

\[
\begin{align*}
X + Y &= [x + y, \overline{x} + \overline{y}], & X \cdot Y &= [\min\{xy, x\overline{y}, \overline{x}y, \overline{x}\overline{y}\}, \max\{xy, x\overline{y}, \overline{x}y, \overline{x}\overline{y}\}], \\
X - Y &= [x - \overline{y}, \overline{x} - y], \text{ and} & X/Y &= X \cdot [1/\overline{y}, 1/y], \ 0 \notin Y.
\end{align*}
\]

On a computer we use directed rounding:

\[
X + Y = [\bigvee (x \oplus y), \bigtriangleup (\overline{x} \oplus \overline{y})]
\]

We then have \( X \circ Y \supseteq \{ x \circ y : x \in X, y \in Y \} \)
Interval Extensions

One of the main goals is to enclose the range of a real-valued function $f : \mathbb{R} \rightarrow \mathbb{R}$:

$$R(f; D) := \{ f(x) : x \in D \}$$

This is achieved by constructing an interval extension $F : \mathbb{IR} \rightarrow \mathbb{IR}$ of $f : \mathbb{R} \rightarrow \mathbb{R}$.

Monotone functions are easy!

$$\exp(X) = [\exp(x), \exp(\bar{x})]$$

$$\sqrt(X) = [\sqrt(x), \sqrt(\bar{x})], \quad \text{if } 0 \leq x$$

$$\log(X) = [\log(x), \log(\bar{x})], \quad \text{if } 0 < x$$

$$\arctan(X) = [\arctan(x), \arctan(\bar{x})].$$

Piecewise monotone functions are also OK!

$$X^n = \begin{cases} 
[x^n, \bar{x}^n] & : n \in \mathbb{Z}^+ \text{ is odd}, \\
[(X)^n, |X|^n] & : n \in \mathbb{Z}^+ \text{ is even}, \\
[1, 1] & : n = 0, \\
[1/\bar{x}, 1/x]^{-n} & : n \in \mathbb{Z}^-; 0 \notin X
\end{cases}$$
We define the class of standard functions to be the set
\[ \mathcal{S} = \{ \exp x, \log x, x^a, |x|, \sin x, \cos x, \tan x, \cdots, \arccos x, \arctan x, \sinh x, \cosh x, \tanh x \} \]

For any \( f \in \mathcal{S} \), we can construct a \textit{sharp} interval extension \( F \), i.e.,
\[ f \in \mathcal{S} \Rightarrow R(f; X) = F(X). \]

Building new functions is easy...

We use finite combinations of constants, elements of \( \mathcal{S} \), \{\( +, -, \cdot, / \}\}, and their compositions to build the elementary functions \( \mathcal{E} \). Interval versions of \( \mathcal{S} \) and \{\( +, -, \cdot, / \}\} provide the corresponding interval extensions.

But, we may now over-estimate the range ...

If \( f(x) = \frac{x}{1+x^2} \), then \( F(X) = \frac{X}{1+X^2} \). For the interval \( X = [1, 2] \), we have
\[ R(f; [1, 2]) = \left[ \frac{2}{5}, \frac{1}{2} \right] \subseteq \left[ \frac{1}{5}, 1 \right] = F([1, 2]). \]
Interval Enclosures

Theorem (1). If \( f(x) \in \mathcal{E} \), and \( F(X) \) is well-defined, then

\[
R(f;X) \subseteq F(X)
\]

How tight is the enclosure?

Theorem (2). If \( f \in \mathcal{E}, X = X_1 \cup X_2 \cdots \cup X_k, \) and \( F(X) \) is well-defined, then

\[
R(f;X) \subseteq \bigcup_{i=1}^{k} F(X_i) \subseteq F(X)
\]

If \( f \) is Lipschitz on \( X \) there is a \( K \geq 0 \) s.t.

\[
d \left( \bigcup_{i=1}^{k} F(X_i) \right) - d (R(f;X)) = K \max_{i} d(X_i)
\]

I.A. (almost) gives us access to \( R(f;X) \).
A consequence of Theorem 1: \( y \notin F(X) \Rightarrow y \notin R(f; X) \)

**Exercise 1:** Let \( f(x) = \cos(x)^3 + \sin(x) \). Prove that \( f(x) \neq 0 \), \( \forall x \in [0, \frac{5}{4}] \).

**Solution:** Define \( F(X) = \cos(X)^3 + \sin(X) \). Then, by Theorem 1, we have

\[
R(f; [0, \frac{5}{4}]) \subseteq F([0, \frac{5}{4}]) = [\cos(\frac{5}{4})^3, 1] + [0, \sin(\frac{5}{4})] \subseteq [0.0313, 1.9490].
\]
‘Impossible’ cases too

Exercise 2: Draw the graph of the function

\[ f_a(x) = x^2 - \frac{3}{10} e^{-(a(x-\frac{1}{2}))^2}, \text{ for } a = 200, \text{ over the interval } [-1, 1]. \]

Even for huge \(a\), the I.A.-methods cannot miss the sharp bend! Conventional methods do.
Rejection Envelopes via Computer-Aided Proofs

Obtain an envelope of the Lipschitz function \(-\sum_{k=1}^{5} k \cdot x \cdot \sin\left(\frac{k(x-3)}{3}\right)\).

**Solution:** Define \(F(X) = -\sum_{k=1}^{5} k \cdot X \cdot \sin\left(\frac{k(X-3)}{3}\right)\). Then, by

**Theorem 1** \(R(f; X) \subseteq F(X)\)

**Theorem 2** we can bisect the domain into smaller pieces until \(\max_i d(F(X_i)) \leq TOL\)

Recursive evaluation of the sub-expressions \(s_i\) by \(f_i\) on the DAG for \(x \cdot \sin((x - 3)/3)\)

\[
\begin{align*}
  s_1 &= x \\
  f_1 &= s_1 \\
  s_3 &= x - 3 \\
  f_3 &= - \\
  s_4 &= \frac{x-3}{3} \\
  f_4 &= / \\
  s_5 &= \sin\left(\frac{x-3}{3}\right) \\
  f_5 &= \sin \\
  s_6 &= x \sin\left(\frac{x-3}{3}\right) \\
  f_6 &= .
\end{align*}
\]
Auto-validating von Neumann RS ⇔ Moore RS (MRS)

Suppose,

- Compact domain \( \Theta = [\theta, \bar{\theta}] \)
- Target shape \( p^*(\theta) : \Theta \to \mathbb{R} \)
- Target integral \( N_p := \int_\Theta p^*(\theta) \, d\theta \)
- Target density \( p(\theta) := \frac{p^*(\theta)}{N_p} : \Theta \to \mathbb{R} \)
- Interval extension of \( p^* \)
- Partition of \( \Theta \)

\( P^*(\Theta) : \mathbb{I}\Theta \to \mathbb{I}\mathbb{R} \)
\( \mathcal{T} := \{ \Theta(1), \Theta(2), ..., \Theta(|\mathcal{T}|) \} \)

then, by Theorem 1

\[
p^*(\Theta(i)) \subseteq P^*(\Theta(i)) := [P^*(\Theta(i)), P^*(\Theta(i))] \quad \forall i \in \{1, 2, ..., |\mathcal{T}|\}.
\]

Construct the \( \mathcal{T} \)-specific proposal \( q^{\mathcal{T}}(\theta) \) as a normalized simple function over \( \Theta \)

\[
q^{\mathcal{T}}(\theta) = \left( N_{q^{\mathcal{T}}} \right)^{-1} \sum_{i=1}^{\mid \mathcal{T} \mid} P^*(\Theta(i)) \, \mathbf{1}_{\{ \theta \in \Theta(i) \}}, \quad N_{q^{\mathcal{T}}} := \sum_{i=1}^{\mid \mathcal{T} \mid} \left( d(\Theta(i)) \cdot P^*(\Theta(i)) \right).
\]

Then an envelope function \( f(q^{\mathcal{T}}(\theta)) \) guaranteeing the necessary inequality is

\[
f_{q^{\mathcal{T}}}(\theta) = \sum_{i=1}^{\mid \mathcal{T} \mid} P^*(\Theta(i)) \, \mathbf{1}_{\{ \theta \in \Theta(i) \}} \geq p^*(\theta), \quad \forall \theta \in \Theta
\]
Auto-validating von Neumann RS ⇔ Moore RS (MRS)

Suppose,

Compact domain  \( \Theta = [\theta, \overline{\theta}] \)
Target shape  \( p^*(\theta) : \Theta \to \mathbb{R} \)
Target integral  \( N_p := \int_{\Theta} p^*(\theta) \, d\theta \)
Target density  \( p(\theta) := \frac{p^*(\theta)}{N_p} : \Theta \to \mathbb{R} \)
Interval extension of \( p^* \)
Partition of \( \Theta \)

\[ \Xi := \{ \Theta(1), \Theta(2), \ldots, \Theta(|\Xi|) \} \]

Efficiency  \( \Leftrightarrow \) Large average acceptance probability

\[ A^p_{\Xi} = \frac{\int_{\Theta} p^*(\theta) \, d\theta}{\int_{\Theta} f_q(\theta) \, d\theta} = \frac{N_p}{\sum_{i=1}^{|\Xi|} \left( d(\Theta(i)) \cdot \overline{P}^*(\Theta(i)) \right)} \geq \frac{\sum_{i=1}^{|\Xi|} \left( d(\Theta(i)) \cdot P^*(\Theta(i)) \right)}{\sum_{i=1}^{|\Xi|} \left( d(\Theta(i)) \cdot \overline{P}^*(\Theta(i)) \right)} \]

Furthermore, if \( p^* \in \mathcal{E}_\Xi \), the Lipschitz class of elementary functions

\[ A^p_{\Xi} \rightarrow 1 - \mathcal{O} \left( \max_{i \in \{1, \ldots, |\Xi|\}} d(\Theta(i)) \right) \]

Efficiency can be further improved by:

- tighter range enclosures through automatic differentiation
- clever partitioning strategies
MRS – Bivariate Levy Densities

\[ E(X_1, X_2) = \sum_{i=1}^{5} i \cos ((i - 1)X_1 + i) \sum_{j=1}^{5} j \cos ((j + 1)X_2 + j) + (X_1 + 1.42513)^2 + (X_2 + 0.80032)^2 \]

\[ l_T(X_1, X_2) = \exp\{-E(X_1, X_2)/T\} \]

There are 700 modes!
MRS – Bivariate Levy Densities

Adaptive partitioning of the domain $[-10, 10] \times [-10, 10]$ into 150 rectangles for Moore rejection sampling from the Levy target density $l_{40}$ (acceptance probab. $= 0.01$).
Acceptance probability \((A_{\mathcal{V}_{\alpha}}^{l_T})\) versus partition size \((|\mathcal{V}_{\alpha}|)\) for Levy targets \(l_T\), where \(T\) is the temperature parameter. There is an optimal CPU time (2.0GHz) to generate \(10^4\) samples.
MRS – Trivariate Needle in a Haystack

\[ p^*(x) = \frac{1}{\sigma_1^3} \exp\left\{-\frac{1}{2}((x - \mu_1)/\sigma_1)^2\right\} + \frac{1}{\sigma_2^3} \exp\left\{-\frac{1}{2}((x - \mu_2)/\sigma_2)^2\right\} \]

Can we find needles in haystacks? – Yes! at least those with a natural interval extension.
MRS – Trivariate Needle in a Haystack

1. MCMC with Metropolis proposal (uniform in cube of side $6\sigma_1$ centered at $x$)
   - With burn-in defined as ending at $B/W = 0.05$
   - Run length is 10 times burn-in (typical run length 20000-50000 for $\sigma_2 = 0.01$).

2. MRS with 1000 boxes and 10000 samples.

Raazesh Sainudiin, Dept. of Maths & Stats, University of Canterbury, Christchurch, NZ
\[ e^c_d(X) = \sum_{i=1}^{c} \exp \left( -|X - \alpha^{(i)}| \right), \quad \alpha_j^{(i)} = a^{(i)} \in \Theta = [-100, 100]^D, \quad j = 1, \ldots, D, \quad i = 1, \ldots, c \]
MRS – Multivariate Rosenbrock’s Density

\[ r_D(X) = \exp\left\{ -\sum_{i=2}^{D} (100(X_i - X_{i-1}^2)^2 + (1 - x_{i-1})^2) \right\} \]
MRS – Multivariate Rosenbrock’s Density

Acceptance probability ($A^{r_D}_{{\mathcal{V}}_\alpha}$) versus partition size ($|{\mathcal{V}}_\alpha|$) for Rosenbrock targets $r_D$, where $D$ is the dimension ($\Theta = [-10, 10]^D$). CPU time (2.0GHz) to generate $10^4$ samples.
MRS – Multivariate Witch’s Hats

Acceptance probability ($A_{h^D_r}$) versus partition size ($|\mathcal{V}_\alpha|$) for Witch’s Hat targets $h^D_r$, where $D$ is the support dimension and $R = 10^{-r}$ is the hat’s radius.
**Phylogenetic Likelihood**

1: **input:** (i) a tree $T_k$, (ii) branch lengths $t = (t_1, t_2, \ldots, t_{b_k})$, (iii) transition probability $P_{a_i, a_j}(t)$ for any $a_i, a_j \in \mathcal{A}$, (iv) stationary distribution $\pi(a_i)$ over each characters $a_i \in \mathcal{A}$, (v) site pattern (data) at site $q$

2: **output:** $\ell_q(k, t)$, the likelihood at site $q$

3: **initialize:** For a leaf node $h$ with observed character $a_i$ at site $q$, set $l^{(a_i)}_h = 1$ and $l^{(a_j)}_h = 0$ for all $j \neq i$. For any internal node $h$, set $l_h := (1, 1, \ldots, 1)$.

4: **recurse:** compute $l_h$ for each sub-terminal node $h$, then those of their ancestors recursively to finally compute $l_r$ for the root node $r$ to obtain the likelihood for site $q$,

$$
\ell_q(k, t) = l_r = \sum_{a_i \in \mathcal{A}} (\pi(a_i) \cdot l^{(a_i)}_r).
$$

For an internal node $h$ with descendants $s_1, s_2, \ldots, s_h$,

$$
l^{(a_i)}_h = \sum_{j_1, \ldots, j_h \in \mathcal{A}} \{ l^{(j_1)}_{s_1} \cdot P_{a_i, j_1}(t_{s_1}) \cdot l^{(j_2)}_{s_2} \cdot P_{a_i, j_2}(t_{s_2}) \cdots l^{(j_h)}_{s_h} \cdot P_{a_i, j_h}(t_{s_h}) \}.
$$
Auto-validating Posterior Estimates of Human-Neandertal Divergence Time

Envelope via Interval-extended post-order traversals

site : 1 1 1 1 1 1
pattern : 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
neandertal : t t c a g g t g t c a a c a a
human : t t c a g g t a c c a g t a g
chimpanzee : t c c a g a a a t t g a c t g

THE Pos. Distn. of Human-Neandertal Divergence Time

4MY H-C Div : (272680, 571124, 1073375)
8MY H-C Div : (545360, 1142248, 2146749)

Human-Neandertal Divergence Estimate (461000, 821000) of Green et. al. (Nature, 2006) is too narrow

10,000 i.i.d. samples from the posterior over the Chimpanzee, Human and Neandertal phylogenies
Model Selection in Primate Inter-relatedness across $3 \times 10^9$ Human-Chimp-Gorilla Genomes

$0.8668 \pm 0.0067, 0.1127 \pm 0.0062, 0.0086 \pm 0.0018, 0.0075 \pm 0.0017, 0.0044 \pm 0.0013$
Summary

• Advantages:
  • Exploits the DAG encoding of the function in the machine
  • Automatically constructs proposal density shape adapted to the target density
  • Produce guaranteed independent samples from inclusion isotonic densities
  • Can be used to account for physical limits on numerical and empirical resolutions in inferential procedures

• Limitations:
  • Can be overwhelmed by complexity, domain size, many dimensions.
  • Much work refining partition before first sample.
  • The MRS is ultimately RAM limited
  • Discrete spaces without an apparent metric structure?

• POA:
  • Pre-enclosures – DAG dissection – Hash Access
  • Algebraic Statistics for dissolving symmetries in DAG (‘minimal sufficiency’)
  • Tighter enclosures via AD – higher-order Taylor expansions
  • Extending arithmetic and $\varepsilon$ to regular sub-pavings
  • The support need not necessarily be Euclidean (CAT(0) space of trees is OK!)
  • The puniest(< 1 ulp)-headed 11-dimensional witch "takes off her hat" to MRS!

• Moore Rejection Sampler is an Auto-validating von Neumann Rejection Sampler

http://www.math.canterbury.ac.nz/~r.sainudiin/codes/mrs/index.shtml
Acknowledgments

People:
Rick Durrett (MATH@Cornell) – Listening and guiding
Michael Nussbaum (MATH@Cornell) – Math. stats. seminars & discussions
Laurent Saloff-Coste (MATH@Cornell) – MCMC asymptotics
Rob Strawderman (STATISTICS@Cornell) – Finite mixtures
Warwick Tucker (MATH@Uppsala, SW) – Introduction to Interval analysis
Karen Vogtmann (MATH@Cornell) – Geometry of tree space
Marty Wells (STATISTICS@Cornell) – Various statistical insights

Funding:
Research Fellow of the Royal Commission for the Exhibition of 1851 (Oxford)