Pliable rejection sampling

Michal Valko

Inria Lille - Nord Europe, France

with

Akram Erraqabi — *Montréal Institute of Learning Algorithms, Canada*
Alexandra Carpentier — *Otto-von-Guericke-Universität Magdeburg, Germany*
Odalric-Ambrym Maillard — *Inria Lille - Nord Europe, France*

SequeL – Inria Lille

GdR ISIS

Paris, February 2018

Adapting to unknown smoothness

Learning the envelope for rejection sampling

Adapting to unknown smoothness

Learning the envelope for rejection sampling

Smooth functions are easier to learn

Adapting to unknown smoothness

Learning the envelope for rejection sampling

Smooth functions are easier to learn

How to adapt to the unknown smoothness?

Adapting to unknown smoothness

Learning the envelope for rejection sampling

Smooth functions are easier to learn

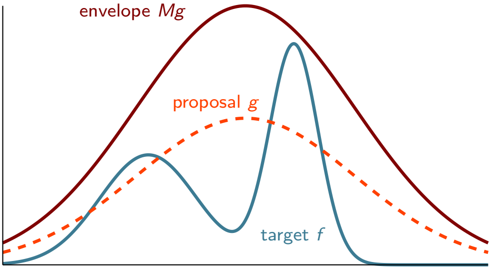
How to adapt to the unknown smoothness?

How to trade off between learning and sampling?

Vanilla rejection sampling

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



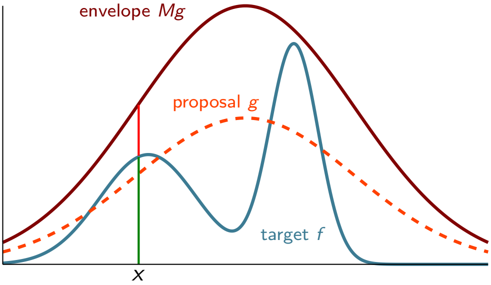
M verifies $f \leq Mg$
Rejection sampling:

1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

Vanilla rejection sampling

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



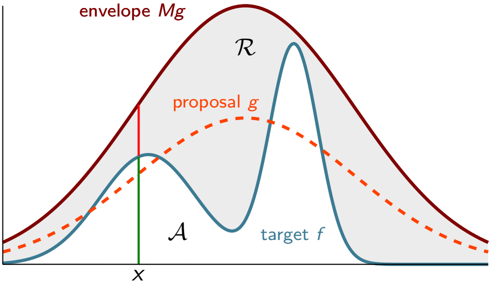
M verifies $f \leq Mg$
Rejection sampling:

1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

Vanilla rejection sampling

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



M verifies $f \leq Mg$
Rejection sampling:

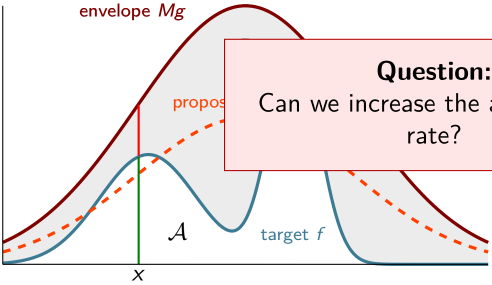
1. Sample x from g
2. Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{M}$$

Vanilla rejection sampling

Rejection Sampling

Goal: Sample from a target density f (not easy to sample from)
Tool: Use a proposal density g (from which sampling is quite easy)



M verifies $f \leq Mg$
... sampling:
... ple x from g
... pt x as a sample
... f with probability $\frac{f(x)}{Mg(x)}$

Question:
Can we increase the acceptance rate?

acceptance rate $= \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{M}$

The setting

Let $d \geq 1$ and let f be a density on $\mathbb{R}^d$.

Goal:

Given a number n of requests to f, what is the number T of samples $Y_1, \ldots, Y_T$ that we can generate such that they are i.i.d. and sampled according to f?

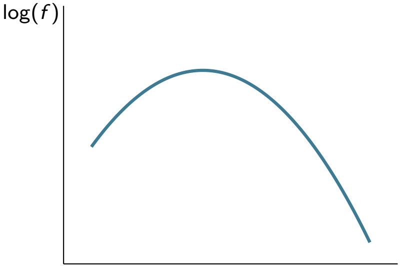
$$\text{acceptance rate} = \frac{T}{n}$$

Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be *log-concave* (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$

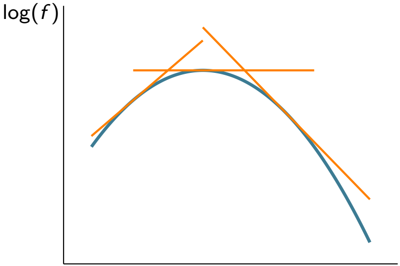


Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be *log-concave* (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$

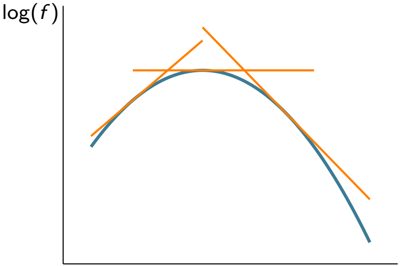


Can we increase the acceptance rate?

Adaptive Rejection Sampling

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]

- The target f is assumed to be ***log-concave*** (unimodal)
- The envelope is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$



Very strong assumption!

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Correlated samples!

Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Can we increase the acceptance rate?

Improved ARS versions

Adaptive Rejection Metropolis Sampling (ARMS) [Gilks, Best and Tan 1995]

- Can deal with non-log-concave densities.
- Performs a Metropolis-Hastings control for each accepted sample
- At each rejection, the sample is added to $\mathcal{S}$

Convex-Concave Adaptive Rejection Sampling [Gorur and Tuh 2011]

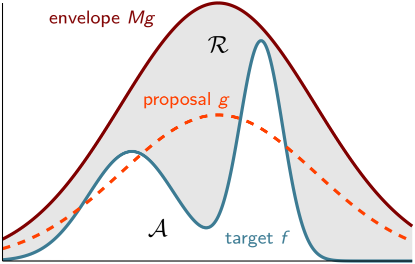
- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)
- At each rejection, the sample is added to $\mathcal{S}$

Correlated samples!

Convexity assumption!

Pliable solution

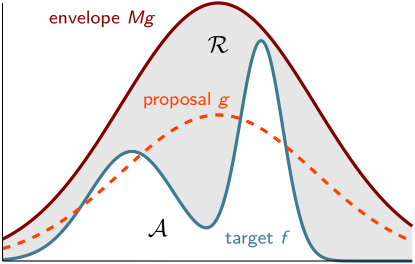
Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Pliable solution

Folding the envelope



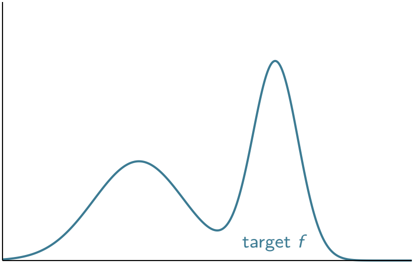
$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A} + \mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

Pliable solution

Folding the envelope



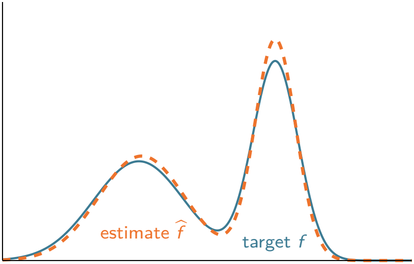
$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A} + \mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

Pliable solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

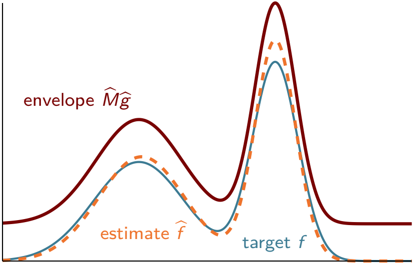
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$

Pliable solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

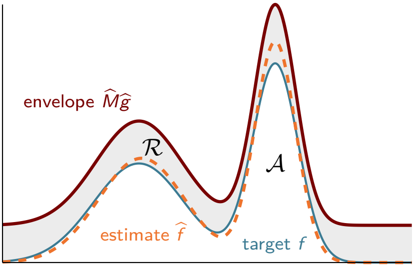
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

Pliable solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

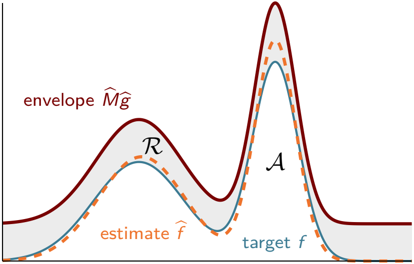
Smaller $\mathcal{R}$ means g should have a similar "shape" to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

Pliable solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A} + \mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar "shape" to f

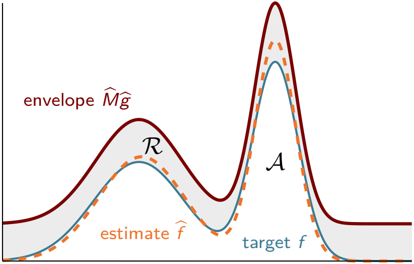
For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

⚠ It should be easy to sample from $\widehat{g}$ ...

Pliable solution

Folding the envelope



$$\text{acceptance rate} = \frac{\mathcal{A}}{\mathcal{A}+\mathcal{R}} = \frac{1}{\widehat{M}}$$

Better proposal means smaller rejection area $\mathcal{R}$

Smaller $\mathcal{R}$ means g should have a similar “shape” to f

For this purpose:

- Build an estimate $\widehat{f}$
- Translate it *uniformly*

⚠ It should be easy to sample from $\widehat{g}$ ... <u>and $\widehat{f}$!</u>

Assumption on the target density f

- The positive function f, defined on $[0, A]^d$ is bounded i.e., there exists $c > 0$ such that the density f satisfies $f(x) \leq c$.

Assumption on the target density f

- The positive function f, defined on $[0, A]^d$ is bounded i.e., there exists $c > 0$ such that the density f satisfies $f(x) \leq c$.

- f can be uniformly expanded by a Taylor expansion in any point up to some degree $0 < s \leq 2$,

$$|f(x+u) - f(x) - \langle \nabla f(x), u \rangle \mathbf{1}\{s > 1\}| \leq c'' \|u\|_2^s.$$

Assumption on the target density f

- The positive function f, defined on $[0, A]^d$ is bounded i.e., there exists $c > 0$ such that the density f satisfies $f(x) \leq c$.

- f can be uniformly expanded by a Taylor expansion in any point up to some degree $0 < s \leq 2$,

$$|f(x+u) - f(x) - \langle \triangledown f(x), u \rangle \mathbf{1}\{s > 1\}| \leq c'' \|u\|_2^s.$$

  - f is in a Hölder ball of smoothness s

Assumption on the target density f

- The positive function f, defined on $[0, A]^d$ is bounded i.e., there exists $c > 0$ such that the density f satisfies $f(x) \leq c$.

- f can be uniformly expanded by a Taylor expansion in any point up to some degree $0 < s \leq 2$,

$$|f(x+u) - f(x) - \langle \triangledown f(x), u \rangle \mathbf{1}\{s > 1\}| \leq c'' \|u\|_2^s.$$

  - f is in a Hölder ball of smoothness s
  - not very restrictive, for a small s

Assumption on the target density f

- The positive function f, defined on $[0, A]^d$ is bounded i.e., there exists $c > 0$ such that the density f satisfies $f(x) \leq c$.

- f can be uniformly expanded by a Taylor expansion in any point up to some degree $0 < s \leq 2$,

$$|f(x+u) - f(x) - \langle \nabla f(x), u \rangle \mathbf{1}\{s > 1\}| \leq c'' \|u\|_2^s.$$

  - f is in a Hölder ball of smoothness s
  - not very restrictive, for a small s
  - f can be an unnormalized density (useful for some Bayesian methods)

Visualizing a 2D example

Multimodal case

$$f(x, y) \propto \left(1 + \sin\left(4\pi x - \frac{\pi}{2}\right)\right)\left(1 + \sin\left(4\pi y - \frac{\pi}{2}\right)\right)$$

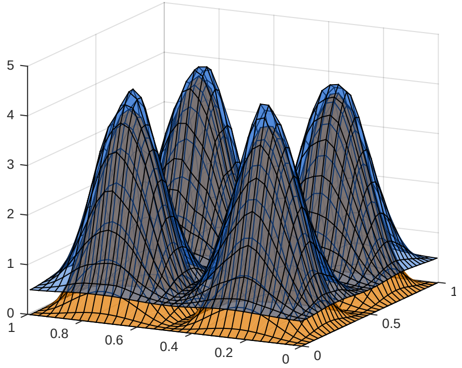


Figure: 2D target density (orange) and the pliable proposal (blue)

Pliable Rejection Sampling

Step 1: Estimating f

- f is defined on $[0, A]^d$, bounded and smooth.
- K is a positive kernel on $\mathbb{R}^d$ (product kernel).
- Let $X_1, \ldots, X_N \sim \mathcal{U}_{[0,A]^d}$. The (modified) kernel regression estimate is

$$\widehat{f}(x) = \frac{A^d}{Nh^d} \sum_{k=1}^{N} f(X_i) K\left(\frac{X_i - x}{h}\right)$$

For an unbounded support density, some extra information is needed to construct a kernel-based estimate.

Pliable Rejection Sampling

Step 1: Estimating f

- f is defined on $[0, A]^d$, bounded and smooth.
- K is a positive kernel on $\mathbb{R}^d$ (product kernel).
- Let $X_1, \ldots, X_N \sim \mathcal{U}_{[0,A]^d}$. The (modified) kernel regression estimate is

$$\widehat{f}(x) = \frac{A^d}{Nh^d} \sum_{k=1}^{N} f(X_i) K\left(\frac{X_i - x}{h}\right)$$

Cost: N requests to f out of n.

For an unbounded support density, some extra information is needed to construct a kernel-based estimate.

Assumption on the kernel K

K_0 be a positive univariate density kernel defined on $\mathbb{R}$

$$K = \prod_{i=1}^{d} K_0$$

Furthermore, it is also of degree 2, i.e., it satisfies

$$\int_{\mathbb{R}} x K_0(x) dx = 0,$$

and, for some $C' > 0$

$$\int_{\mathbb{R}} x^2 K_0(x) dx \leq C'.$$

K_0 is ε-Hölder for some $\varepsilon > 0$, i.e., $\exists C'' > 0$ s.t., for any $(x, y) \in \mathbb{R}^2$,

$$|K_0(y) - K_0(x)| \leq C'' |x - y|^{\varepsilon}.$$

Assumption on the kernel K

K_0 be a positive univariate density kernel defined on $\mathbb{R}$

$$K = \prod_{i=1}^{d} K_0$$

Furthermore, it is also of degree 2, i.e., it satisfies

$$\int_{\mathbb{R}} x K_0(x) dx = 0,$$

and, for some $C' > 0$

$$\int_{\mathbb{R}} x^2 K_0(x) dx \leq C'.$$

K_0 is ε-Hölder for some $\varepsilon > 0$, i.e., $\exists C'' > 0$ s.t., for any $(x, y) \in \mathbb{R}^2$,

$$|K_0(y) - K_0(x)| \leq C'' |x - y|^{\varepsilon}.$$

Gaussian kernel satisfies this with $C = 1$, $C' = 1$, $C'' = 4$, and $\varepsilon = 1$

Pliable Rejection Sampling

Bounding the gap

Theorem 1

The estimate $\widehat{f}$ is such that with probability larger than $1-\delta$, for any point $x \in [0, A]^d$,

$$\left|\widehat{f}(x) - f(x)\right| \leq H_0 \left(\left(\frac{\log(NAd/\delta)}{N} \right)^{\frac{s}{2s+d}} \right)$$

where H_0 is a constant that depends on the problem parameters.

s is the degree to which f can be expanded as a Taylor expression.

Pliable Rejection Sampling

Bounding the gap

Theorem 1

The estimate $\widehat{f}$ is such that with probability larger than $1-\delta$, for any point $x \in [0, A]^d$,

$$\left|\widehat{f}(x) - f(x)\right| \leq H_0 \left(\left(\frac{\log(NAd/\delta)}{N} \right)^{\frac{s}{2s+d}} \right)$$

where H_0 is a constant that depends on the problem parameters.

s is the degree to which f can be expanded as a Taylor expression.

Remaing Budget: $n - N$.

Pliable Rejection Sampling

Step 2: Generating Samples

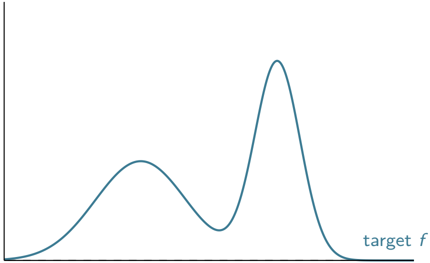
- Remaining requests to f: $n - N$
- Let $r_N = A^d H_C \left(\frac{\log(NAd/\delta)}{N}\right)^{\frac{s}{2s+d}}$
- Construct the *pliable* proposal $\widehat{g}$ out of $\widehat{f}$:

$$\widehat{g} = \frac{\widehat{f} + r_N\, \mathcal{U}_{[0,A]^d}}{\frac{1}{N}\sum_{i=1}^{N} f(X_i) + r_N}$$

- Perform rejection sampling using $\widehat{g}$ and the empirical rejection sampling constant

$$\widehat{M} = \frac{\frac{1}{N}\sum_i f(X_i) + r_N}{\frac{1}{N}\sum_i f(X_i) - 5r_N}$$

The algorithm

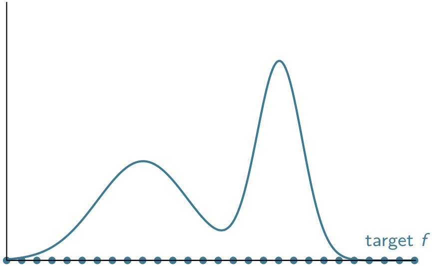


Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

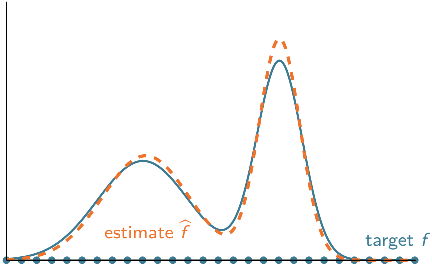
Input: s, n, δ, H_C

Initial Sampling

Draw uniformly at random N samples on $[0, A]^d$

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Initial Sampling

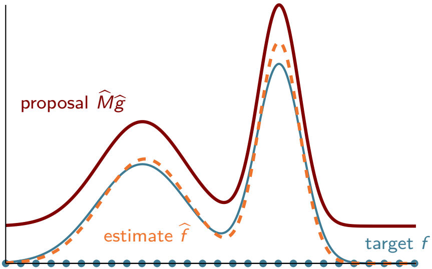
Draw uniformly at random N samples on $[0, A]^d$

Estimation of f

Estimate f using these N samples by kernel regression

Output: $\widehat{n}$ accepted samples

The algorithm



Algorithm: Pliable Rejection Sampling (PRS)

Input: s, n, δ, H_C

Initial Sampling

Draw uniformly at random N samples on $[0, A]^d$

Estimation of f

Estimate f using these N samples by kernel regression

Generating the samples

Sample $n - N$ samples from the *pliable proposal* $\widehat{g}$ and perform Rejection Sampling using $\widehat{M}$ as the envelope constant

Output: $\widehat{n}$ accepted samples

Is the sampling correct?

Theorem 1: w.p. $1-\delta$, for any $x \in [0, A]^d$

$$\xi' \stackrel{\text{def}}{=} \left|\widehat{f}(x) - f(x)\right| \leq r_N \frac{1}{A^d} = r_N \mathcal{U}_{[0,A]^d}.$$

Is the sampling correct?

Theorem 1: w.p. $1-\delta$, for any $x \in [0,A]^d$

$$\xi' \stackrel{\text{def}}{=} \left|\widehat{f}(x) - f(x)\right| \leq r_N \frac{1}{A^d} = r_N \mathcal{U}_{[0,A]^d}.$$

Hoeffding's: w.p. $1-\delta$

$$\xi'' \stackrel{\text{def}}{=} \left\{ \left| \frac{A^d}{n} \sum_{i=1}^{n} f(X_i) - \int_{[0,A]^d} f(x) dx \right| \leq 2A^d c \sqrt{\frac{1}{N} \log(1/\delta)} \stackrel{\text{def}}{=} c_N \right\}$$

Is the sampling correct?

Theorem 1: w.p. $1-\delta$, for any $x \in [0, A]^d$

$$\xi' \stackrel{\text{def}}{=} \left|\widehat{f}(x) - f(x)\right| \leq r_N \frac{1}{A^d} = r_N \mathcal{U}_{[0,A]^d}.$$

Hoeffding's: w.p. $1-\delta$

$$\xi'' \stackrel{\text{def}}{=} \left\{ \left| \frac{A^d}{n} \sum_{i=1}^{n} f(X_i) - \int_{[0,A]^d} f(x)dx \right| \leq 2A^d c \sqrt{\frac{1}{N} \log(1/\delta)} \stackrel{\text{def}}{=} c_N \right\}$$

On, $\xi = \xi' \cap \xi''$, we have for our proposal and $8r_N \leq \int_{[0,A]^d} f(x)dx \stackrel{\text{def}}{=} m$

$$\widehat{g}^\star = \frac{\widehat{f} + r_N \mathcal{U}_{[0,A]^d}}{A^d/n \sum_{i=1}^{n} f(X_i) + r_N} \geq \frac{f}{\int_{[0,A]^d} f(x)dx + r_N + c_N}$$

$$\geq \frac{f}{\int_{[0,A]^d} f(x)dx} (1 - 4r_N/m)$$

Choice of empirical multiplication constant $\widehat{M}$

$$
\begin{aligned}
\frac{1}{1-4r_N/m} &= \frac{m}{m-4r_N} \\
&\leq \frac{A^d/N\sum_i f(X_i)+c_N}{A^d/N\sum_i f(X_i)-c_n-4r_N} \\
&\leq \frac{A^d/N\sum_i f(X_i)+r_N}{A^d/N\sum_i f(X_i)-5r_N} = \widehat{M}
\end{aligned}
$$

Choice of empirical multiplication constant $\widehat{M}$

$$\frac{1}{1-4r_N/m} = \frac{m}{m-4r_N}$$
$$\leq \frac{A^d/N\sum_i f(X_i)+c_N}{A^d/N\sum_i f(X_i)-c_n-4r_N}$$
$$\leq \frac{A^d/N\sum_i f(X_i)+r_N}{A^d/N\sum_i f(X_i)-5r_N} = \widehat{M}$$

$\widehat{M}\hat{g}^\star$ upperbounds f (under ξ)

Choice of empirical multiplication constant $\widehat{M}$

$$
\begin{aligned}
\frac{1}{1-4r_N/m} &= \frac{m}{m-4r_N} \\
&\leq \frac{A^d/N\sum_i f(X_i)+c_N}{A^d/N\sum_i f(X_i)-c_n-4r_N} \\
&\leq \frac{A^d/N\sum_i f(X_i)+r_N}{A^d/N\sum_i f(X_i)-5r_N} = \widehat{M}
\end{aligned}
$$

Sampling is correct whp.

How many accepted samples can we guarantee?

$$\widehat{M} = \frac{A^d/N\sum_i f(X_i) + r_N}{A^d/N\sum_i f(X_i) - 5r_N} \leq \frac{m + r_N + c_N}{m - 5r_N - c_N} \leq \frac{m + 2r_N}{m - 6r_N}.$$

How many accepted samples can we guarantee?

$$\widehat{M} = \frac{A^d/N \sum_i f(X_i) + r_N}{A^d/N \sum_i f(X_i) - 5r_N} \leq \frac{m + r_N + c_N}{m - 5r_N - c_N} \leq \frac{m + 2r_N}{m - 6r_N}.$$

On ξ, we get samples that are i.i.d. according to f, and $\widehat{n}$ will be a sum of Bernoulli random variables of parameter larger than

$$\frac{1}{\widehat{M}} \geq \frac{m - 6r_N}{m + 2r_N} \geq (1 - 6r_N/m)(1 - 4r_N/m) \geq 1 - 20r_N/m,$$

How many accepted samples can we guarantee?

$$\widehat{M} = \frac{A^d/N\sum_i f(X_i) + r_N}{A^d/N\sum_i f(X_i) - 5r_N} \leq \frac{m + r_N + c_N}{m - 5r_N - c_N} \leq \frac{m + 2r_N}{m - 6r_N}.$$

On ξ, we get samples that are i.i.d. according to f, and $\widehat{n}$ will be a sum of Bernoulli random variables of parameter larger than

$$\frac{1}{\widehat{M}} \geq \frac{m - 6r_N}{m + 2r_N} \geq (1 - 6r_N/m)(1 - 4r_N/m) \geq 1 - 20r_N/m,$$

$\widehat{n}$ is with probability larger than $1 - 3\delta$ lower bounded as

$$\widehat{n} \geq (n - N)\left(1 - 20r_N/m - 4\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

How many accepted samples can we guarantee?

$$\widehat{M} = \frac{A^d/N \sum_i f(X_i) + r_N}{A^d/N \sum_i f(X_i) - 5r_N} \leq \frac{m + r_N + c_N}{m - 5r_N - c_N} \leq \frac{m + 2r_N}{m - 6r_N}.$$

On ξ, we get samples that are i.i.d. according to f, and $\widehat{n}$ will be a sum of Bernoulli random variables of parameter larger than

$$\frac{1}{\widehat{M}} \geq \frac{m - 6r_N}{m + 2r_N} \geq (1 - 6r_N/m)(1 - 4r_N/m) \geq 1 - 20r_N/m,$$

$\widehat{n}$ is with probability larger than $1 - 3\delta$ lower bounded as

$$\widehat{n} \geq (n - N)\left(1 - 20r_N/m - 4\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Setting: $N = n^{\frac{2s+d}{3s+d}}$,

$$\widehat{n} \geq n\left[1 - K\log(nAd/\delta)^{\frac{s}{3s+d}} n^{-\frac{s}{3s+d}}\right]. \tag{1}$$

A bound on the acceptance rate

The asymptotic performace

Theorem 2

Under Theorem 1's assumptions and if $H_0 < H_C$, $8r_N \leq \int_{[0,A]^d} f(x)dx$. Then, for n large enough, we have with probability larger than $1-\delta$ that

$$\widehat{n} \geq n \left[1 - \mathcal{O}\left(\frac{\log\left(nAd/\delta\right)}{n} \right)^{\frac{s}{3s+d}} \right].$$

where $\widehat{n}$ is the number of i.i.d. samples generated by PRS.

A bound on the acceptance rate

The asymptotic performace

Theorem 2

Under Theorem 1's assumptions and if $H_0 < H_C$, $8r_N \leq \int_{[0,A]^d} f(x)dx$*. Then, for* n *large enough, we have with probability larger than* $1-\delta$ *that*

$$\widehat{n} \geq n\left[1-\mathcal{O}\left(\frac{\log\left(nAd/\delta\right)}{n}\right)^{\frac{s}{3s+d}}\right].$$

where $\widehat{n}$ *is the number of i.i.d. samples generated by PRS.*

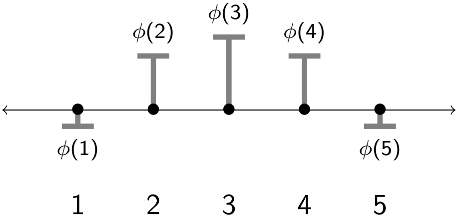
Convergence Rate ↑ with s

Convergence Rate ↓ with d

Competitor: A* sampling from Gumbel-Max trick

Gumbel-Max trick: $p(i) \propto \exp(\phi(i))$ for $i \in \{1, 2, 3, 4, 5\}$

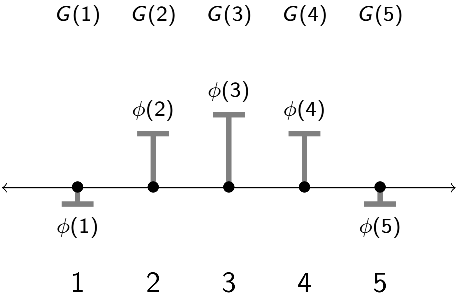
images from Chris J. Maddison



Competitor: A* sampling from Gumbel-Max trick

Gumbel-Max trick: $p(i) \propto \exp(\phi(i))$ for $i \in \{1, 2, 3, 4, 5\}$

images from Chris J. Maddison

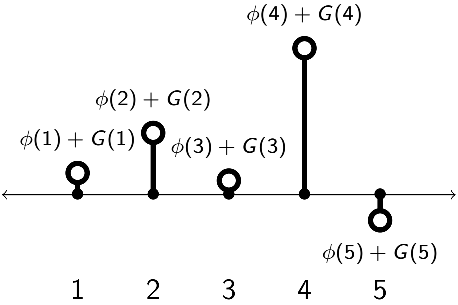


$G(i) \sim \text{Gumbel}(0)$ IID

Competitor: A* sampling from Gumbel-Max trick

Gumbel-Max trick: $p(i) \propto \exp\left(\phi\left(i\right)\right)$ for $i \in \{1, 2, 3, 4, 5\}$

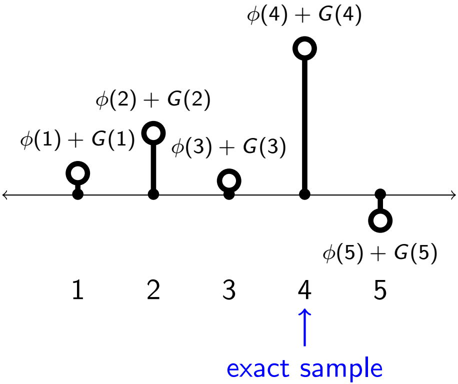
images from Chris J. Maddison



Competitor: A* sampling from Gumbel-Max trick

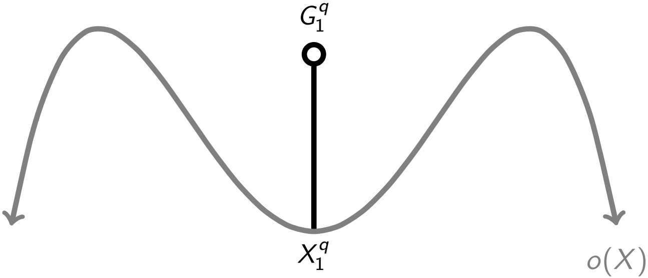
Gumbel-Max trick: $p(i) \propto \exp\left(\phi\left(i\right)\right)$ for $i \in \{1, 2, 3, 4, 5\}$

images from Chris J. Maddison



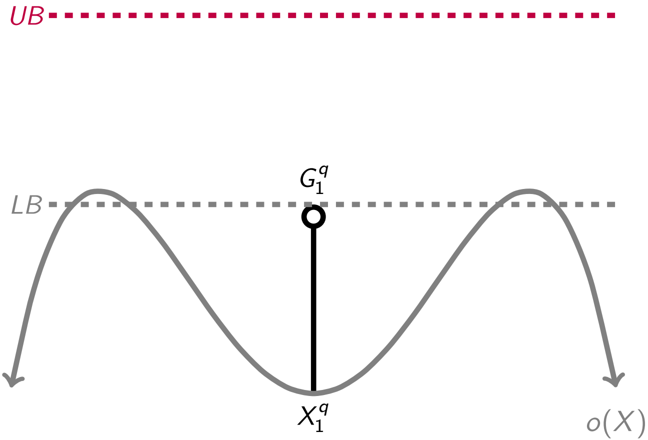
$A^\star$ sampling

Continuous Gumbel-Max trick: $f(x) \propto \exp\left(i(x) + o(x)\right)$



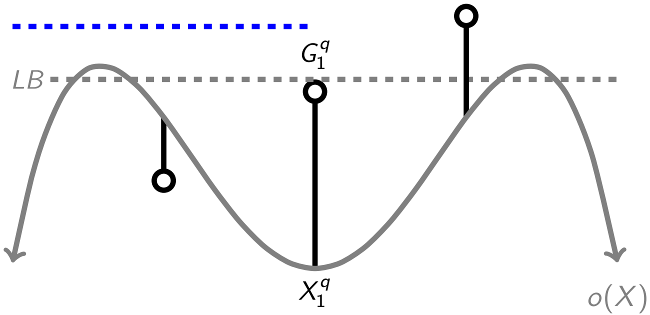
A* sampling

Continuous Gumbel-Max trick: $f(x) \propto \exp\left(i(x) + o(x)\right)$



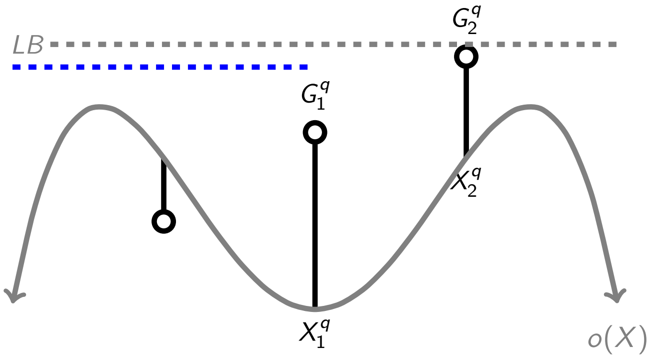
A* sampling

Continuous Gumbel-Max trick: $f(x) \propto \exp(i(x) + o(x))$



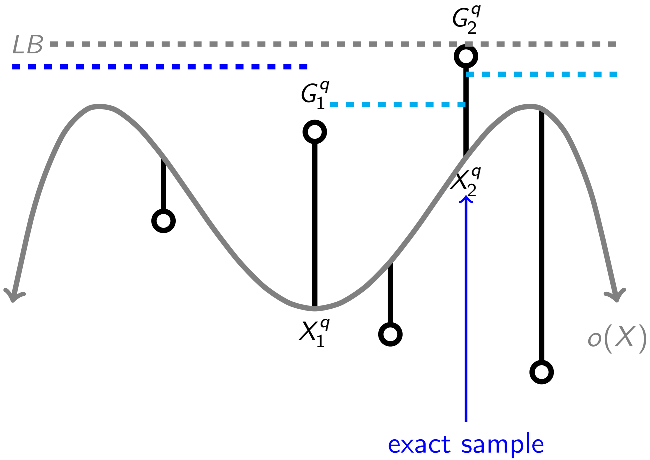
$A^\star$ sampling

Continuous Gumbel-Max trick: $f(x) \propto \exp\left(i(x) + o(x)\right)$



A* sampling

Continuous Gumbel-Max trick: $f(x) \propto \exp\left(i(x) + o(x)\right)$



A* sampling vs. PRS

– A* needs several calls to f to generate a sample

\+ PRS rejects (asymptotically) only a negligible number of samples with respect to n

A* sampling vs. PRS

– A* needs several calls to f to generate a sample

+ PRS rejects (asymptotically) only a negligible number of samples with respect to n

number of i.i.d. samples generated according to f per computation of f are better than the ones for A* sampling

A* sampling vs. PRS

- – A* needs several calls to f to generate a sample
- + PRS rejects (asymptotically) only a negligible number of samples with respect to n

number of i.i.d. samples generated according to f per computation of f are better than the ones for `A⋆ sampling`

- – A* needs a decomposition $f(x) \propto \exp(\phi(x))$, where $\phi(x) = i(x) + o(x)$
- + PRS learns it!

Scaling with d?

A* sampling vs. PRS

- $-$ A* needs several calls to f to generate a sample
- $+$ PRS rejects (asymptotically) only a negligible number of samples with respect to n

number of i.i.d. samples generated according to f per computation of f are better than the ones for A* sampling

- $-$ A* needs a decomposition $f(x) \propto \exp(\phi(x))$, where $\phi(x) = i(x) + o(x)$
- $+$ PRS learns it!

Scaling with d? Same.

$A^\star$ sampling vs. PRS

– $A^\star$ needs several calls to f to generate a sample

+ PRS rejects (asymptotically) only a negligible number of samples with respect to n

number of i.i.d. samples generated according to f per computation of f are better than the ones for $\texttt{A}^\star$ `sampling`

– $A^\star$ needs a decomposition $f(x) \propto \exp(\phi(x))$, where $\phi(x) = i(x) + o(x)$

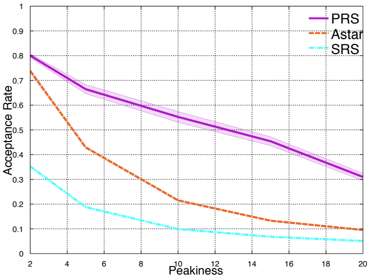
+ PRS learns it!

Scaling with d? Same.

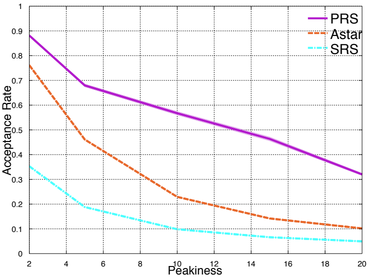
Experiments

Scaling with peakiness

$$f \propto \frac{e^{-x}}{(1+x)^a}, \text{ } a \text{ defines the peakiness level}$$



$n = 10^4$

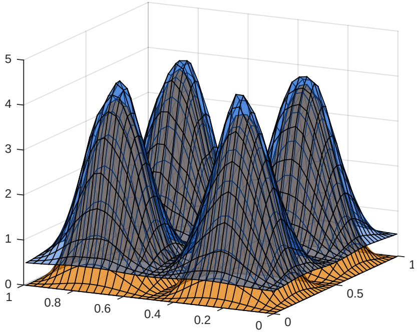


(b) $n = 10^5$

Figure: Acceptance rate *vs.* peakiness

Experiments

Two dimensional example



$n = 10^6$	*acceptance rate*	*standard deviation*
PRS	66.4%	0.45%
$A^\star$ sampling	76.1%	0.80%
SRS	25.0%	0.01%

Table: 2D example: Acceptance rates averaged over 10 trials

Experiments

The Clutter problem

$n = 10^5$, 1D	*acceptance rate*	*standard deviation*
PRS	79.5%	0.2%
A* sampling	89.4%	0.8%
SRS	17.6%	0.1%

$n = 10^5$, 2D	*acceptance rate*	*standard deviation*
PRS	51,0%	0.4%
A* sampling	56.1%	0.5%
SRS	2.10^{-3}%	10^{-5}%

Table: Clutter problem: Acceptance rates averaged over 10 trials

Discussion

Normalized distribution

If $\int_{[0,A]^d} f = 1$ then we can simplify the algorithm

$$\hat{g}^\star \stackrel{\text{def}}{=} \frac{1}{1+r_N}\left(\widehat{f} + r_N \mathcal{U}_{[0,A]^d}\right)$$

Discussion

Normalized distribution

If $\int_{[0,A]^d} f = 1$ then we can simplify the algorithm

$$\hat{g}^\star \stackrel{\text{def}}{=} \frac{1}{1+r_N}\left(\widehat{f} + r_N \mathcal{U}_{[0,A]^d}\right)$$

Case of a distribution with unbounded support

instead of uniformly sampling on $[0,A]^d$, we sample on a hypercube centered in 0 and of side length $\sqrt{\log(n)}$

Discussion

Normalized distribution

If $\int_{[0,A]^d} f = 1$ then we can simplify the algorithm

$$\hat{g}^\star \stackrel{\mathrm{def}}{=} \frac{1}{1+r_N}\left(\widehat{f} + r_N \mathcal{U}_{[0,A]^d}\right)$$

Case of a distribution with unbounded support

instead of uniformly sampling on $[0, A]^d$, we sample on a hypercube centered in 0 and of side length $\sqrt{\log(n)}$

Extensions for high dimensional cases (large d)

– when the mass of the distribution is localized in a few small subsets

Conclusion

+ PRS deals with a wide class of functions

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler

Conclusion

- + PRS deals with a wide class of functions
- + PRS has guarantees: asymptotically we accept everything (whp)
- + PRS is a **perfect** sampler
  - + (whp) the samples are iid (unlike MCMC)

Conclusion

- + PRS deals with a wide class of functions
- + PRS has guarantees: asymptotically we accept everything (whp)
- + PRS is a **perfect** sampler
  - + (whp) the samples are iid (unlike MCMC)
- + PRS's empirical performance is comparable to state of the art

Conclusion

- PRS deals with a wide class of functions
- PRS has guarantees: asymptotically we accept everything (whp)
- PRS is a **perfect** sampler
  - (whp) the samples are iid (unlike MCMC)
- PRS's empirical performance is comparable to state of the art
- We have an extension to densities with unbounded support

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
    + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

− PRS works only for small and moderate dimensions

Conclusion

- + PRS deals with a wide class of functions
- + PRS has guarantees: asymptotically we accept everything (whp)
- + PRS is a **perfect** sampler
  - + (whp) the samples are iid (unlike MCMC)
- + PRS's empirical performance is comparable to state of the art
- + We have an extension to densities with unbounded support

- − PRS works only for small and moderate dimensions
  - + in favorable cases, it can scale to high dimensions as well

Conclusion

- + PRS deals with a wide class of functions
- + PRS has guarantees: asymptotically we accept everything (whp)
- + PRS is a **perfect** sampler
  - + (whp) the samples are iid (unlike MCMC)
- + PRS's empirical performance is comparable to state of the art
- + We have an extension to densities with unbounded support

- – PRS works only for small and moderate dimensions
  - + in favorable cases, it can scale to high dimensions as well
- – It does not work well for peaky distributions (posteriors)

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
  + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

− PRS works only for small and moderate dimensions
  + in favorable cases, it can scale to high dimensions as well

− It does not work well for peaky distributions (posteriors)

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
    + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

– PRS works only for small and moderate dimensions
    + in favorable cases, it can scale to high dimensions as well

– It does not work well for peaky distributions (posteriors)

Extension 1: Iterative PRS by re-estimating f several times

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
  + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

– PRS works only for small and moderate dimensions
  + in favorable cases, it can scale to high dimensions as well

– It does not work well for peaky distributions (posteriors)

Extension 1: Iterative PRS by re-estimating f several times
Extension 2: Using the fact that the evaluations are noiselles

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
  + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

– PRS works only for small and moderate dimensions
  + in favorable cases, it can scale to high dimensions as well

– It does not work well for peaky distributions (posteriors)

Extension 1: Iterative PRS by re-estimating f several times
Extension 2: Using the fact that the evaluations are noiselles
Improved rate and lower bound: Follow-up work by Juliette Achdou and Alexandra Carpentier

Conclusion

+ PRS deals with a wide class of functions
+ PRS has guarantees: asymptotically we accept everything (whp)
+ PRS is a **perfect** sampler
  + (whp) the samples are iid (unlike MCMC)
+ PRS's empirical performance is comparable to state of the art
+ We have an extension to densities with unbounded support

– PRS works only for small and moderate dimensions
  + in favorable cases, it can scale to high dimensions as well

– It does not work well for peaky distributions (posteriors)

Extension 1: Iterative PRS by re-estimating f several times
Extension 2: Using the fact that the evaluations are noiselles
Improved rate and lower bound: Follow-up work by Juliette Achdou and Alexandra Carpentier

Thank you!

SequeL – Inria Lille

GdR ISIS