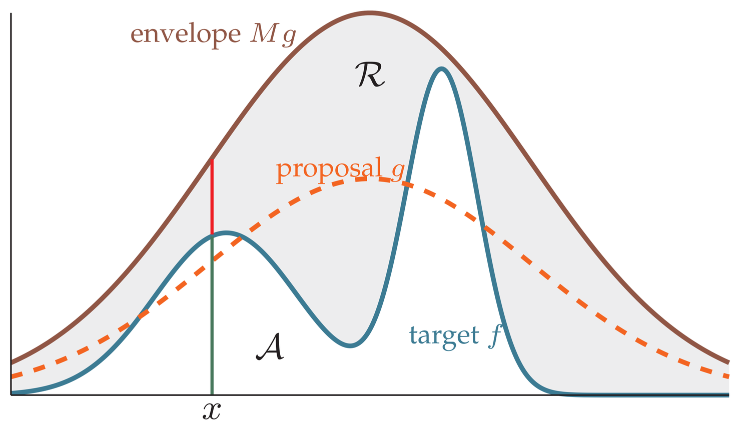
Pliable Rejection Sampling

AKRAM.ER-RAQABI@UMONTREAL.CA, MICHAL.VALKO@INRIA.FR, CARPENTIER@MATH.UNI-POTSDAM.DE, AND ODALRIC.MAILLARD@INRIA.FR

Simple 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)
Property: Smaller $\mathcal{R}$ $\implies$ fewer rejections (good!)



M verifies $f \le Mg$. The sampling algorithm:
- Sample x from g
- Accept x as a sample from f with probability $\frac{f(x)}{Mg(x)}$

Setting

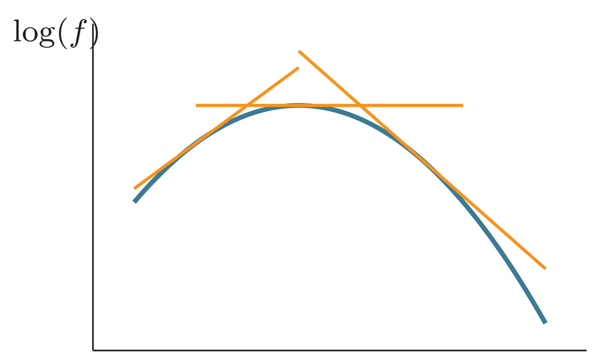
Let $d \ge 1$ and let f be a density on $\mathbb{R}^d$.

Question: Given a number n of requests to f, what is the number T of samples $Y_1, \dots, Y_T$ that one can generate such that they are i.i.d. according to f?

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

Can we increase the acceptance rate?

Adaptive Rejection Sampling (ARS) [Gilks and Wild 1992]
- The target f is assumed to be *log-concave* (unimodal)
- The enveloppe is made of tangents at a set of points $\mathcal{S}$
- At each rejection, the sample is added to $\mathcal{S}$

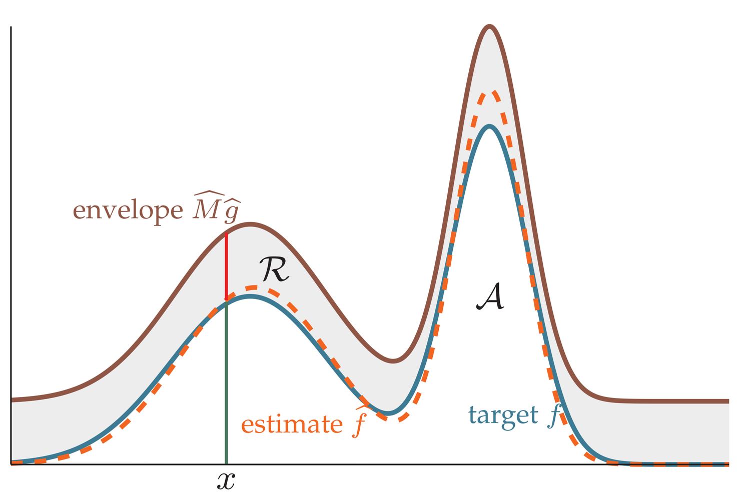


Adaptive Rejection Metropolis Sampling (ARMS)
[Gilks, Best and Tan 1995]
- Performs a Metropolis-Hastings step for each accepted sample (which correlates the samples)

Convex-Concave Adaptive Rejection Sampling
[Gorur and Tuh 2011]
- Decomposes the target as convex + concave
- Builds piecewise linear upper bounds (tangents, secant lines)

Pliable Rejection Sampling

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 $\widehat{g}$ ... and $\widehat{f}$!

Choice of the Estimate

Assumption 1 (on the density).
- *f defined on $[0, A]^d$ and bounded.*
- *It admits a Taylor expansion in any point up to some degree $0 < s \le 2$.*

Assumption 2 (on the kernel).
- *Let $K = \prod_{i=1}^d K_0$*
- *K_0 is a a density kernel : defined on $\mathbb{R}$, uniformly bounded, normalized and non-negative*
- *K_0 is ε-Hölder for some $\varepsilon > 0$*
- *K_0 is of degree 2, i.e:*

$$\int_{\mathbb{R}} x K_0(x)dx = 0 \quad \text{and} \quad \int_{\mathbb{R}} x^2 K_0(x)dx < \infty$$

Let $X_1, \dots, X_N \sim \mathcal{U}_{[0,A]^d}$. The (modified) kernel regression estimate is

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

K_0 Gaussian kernel $\Rightarrow$ $\widehat{f}$ is a Gaussian mixture!

Bounding the Gap

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

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

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

The Pliable Proposal

- 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

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

Algorithm: PRS

Parameters: s, n, δ, H_C
Initial sampling
Draw uniformly at random N samples on $[0, A]^d$ and evaluate f on them
Estimation of f
Estimate f by $\widehat{f}$ on these N samples (Equation 1)
Generating the samples
Sample $n - N$ samples from the *compact pliable proposal* $\widehat{g}^\star$
Perform rejection sampling on these samples using $\widehat{M}$ as a rejection constant to get $\widehat{n}$ samples
Output: Return the $\widehat{n}$ samples

Number of Accepted Samples

Theorem 2. *Under Theorem 1's assumptions and if:*
- $H_0 < H_C$
- $8r_N \le \int_{[0,A]^d} f(x)dx$

For n large enough, we have with probability larger than $1 - \delta$ that

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

Convergence Rate ↑ with smoothness

Convergence Rate ↓ with dimensionality

PRS Properties

- PRS deals with a wider class of functions and not necessarily normalized
- PRS has guarantees: asymptotically we accept everything (whp).
- PRS is a **perfect** sampler (whp) the samples are iid (unlike MCMC)
- PRS empirical performance is comparable to state of the art
  - PRS deals better with peakiness than `A* sampling`
  - in general PRS does not scale to high dimensions
- An extension to densities with unbounded support is provided

Some notes on (very) high dimensionality

Let the γ-support of f be

$$\text{Supp}_{f,\gamma} = \overline{\Lambda}_{f,\gamma} \quad \text{where} \quad \Lambda_{f,\gamma} \overset{\text{def}}{=} \{x \in \mathcal{D} : f(x) > \gamma\}$$

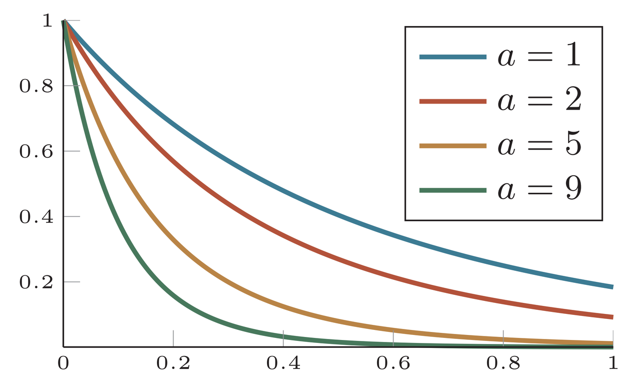
- In general, localizing the 0-support of f may require an exponential cost in d
- $\text{Supp}_{f,\gamma}$ is localizable with a less than exponential cost, if:
  - $f_{|\text{Supp}^c_{f,\gamma}}$ the restriction of f on the complementary of $\text{Supp}_{f,\gamma}$ is convex
  - One can evalute f and its gradient pointwise

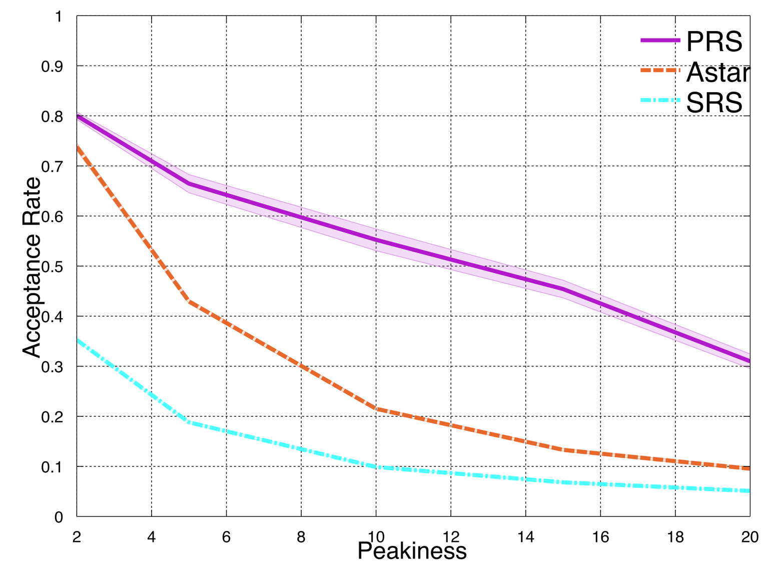
$\text{Supp}_{f,\gamma-\epsilon}$ localization cost $\mathcal{O}(d^2/\varepsilon^2)$

The trick: find a point x_0 in $\text{Supp}^c_{f,\gamma}$ and use standard gradient based optimization to find a maximum on $\partial\text{Supp}_{f,\gamma}$.

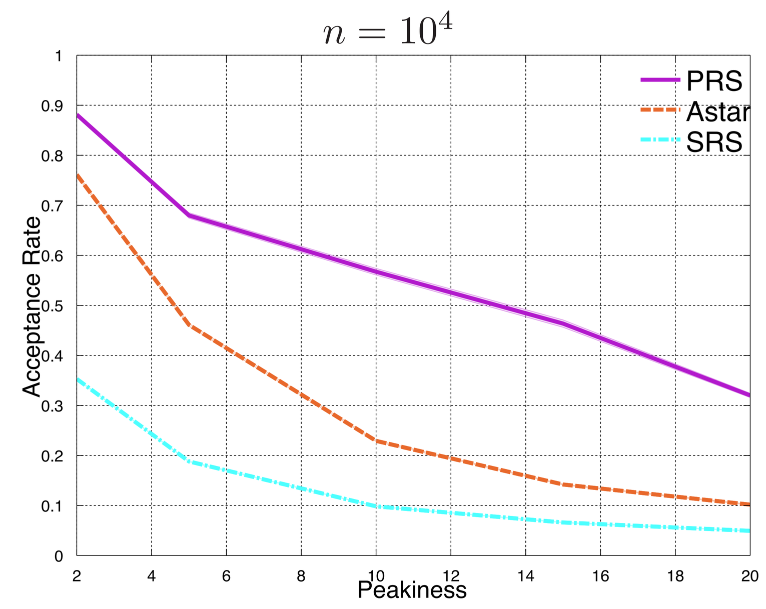
Experiments - Scaling with Peakiness

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





$n = 10^4$

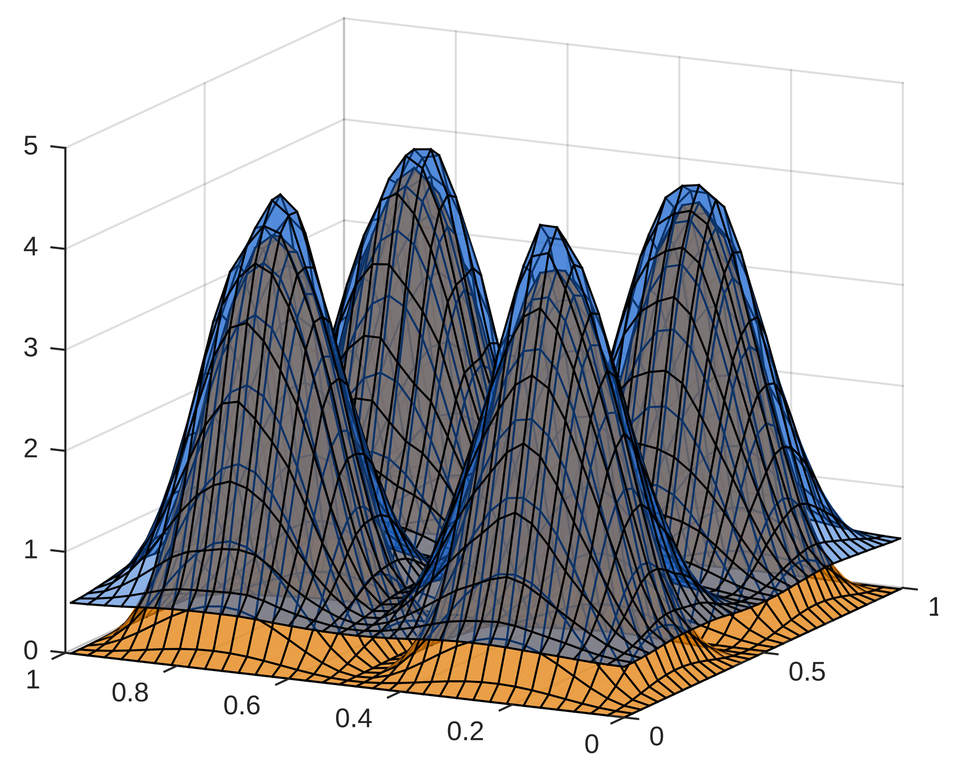


$n = 10^5$

Experiments - Multimodal Examples

A 2D example

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



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

The Clutter Problem [Thomas P. Minka, UAI '01]

- Consider K data points $(X_i)_{i=1}^K$ with half within [-5, -3]^d and half within [2, 4]^d
- These points are assumed to be generated from

$$p(x|\theta) = (1 - \pi)\mathcal{N}(x; \theta, I) + \pi\mathcal{N}(x; 0, 10I)$$

- We put a gaussian prior on the mean $p(\theta) = \mathcal{N}(\theta; 0, 100I)$
- The goal is to sample from $p(\theta|(X_i)_{i=1}^K) \propto p(\theta)\prod_{i=1}^K p(X_i|\theta)$

$n = 10^5$, 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}%

Extending this Work

Iterative version:
- PRS is a 2 step algorithm: estimation + RS
- Possibly improve the estimate on several steps
- Optimize the number of samples gathered between these "estimation update steps"