# Rejection Sampling for Weighted Densities by Majorization

**Andrew M. Raim**

Joint with **James A. Livsey** and **Kyle M. Irimata**
Center for Statistical Research and Methodology
U.S. Census Bureau

CSRM Seminar
June 8, 2023

# Disclaimer

This presentation is to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the author and not those of the U.S. Census Bureau.

# Overview

- This work considers a method to obtain proposal distributions in rejection sampling (von Neumann, 1951).
- The vertical strip method (aka “Ahrens' method”) provides one recipe for a proposal distribution.
- We revisit vertical strips for targets which are weighted densities.
- Can draw from some distributions which might otherwise be challenging. This includes several (univariate) examples from the literature to be discussed here:
  1. Polynomial-Normal distribution;
  2. Conway-Maxwell Poisson distribution;
  3. Posterior in a Gaussian Process with unknown variance parameter.
- There are many similarities here to the rejection sampling approach in Raim (2023) to the direct sampler (Walker et al., 2011). The methods in the present work are more straightforward.

# Target Distribution

- To generate variates from the target density

$$f(x) = f_0(x)/\psi, \quad \psi = \int_\Omega f_0(x) d\nu(x),$$

where

1. $\Omega$ is the support of $f$,
2. $\psi$ is a normalizing constant (may be intractable),
3. $\nu$ is a dominating measure.

# Rejection Sampling

- Let $h(x) = h_0(x)/a$ be a proposal distribution.
- Ideally, $h$ should be
  1. easy to evaluate,
  2. easy to draw variates from,
  3. its support of should include $\Omega$, and
  4. not too different from $f$.
- Find a "ratio adjustment factor" $M$ such that

$$\sup_{x \in \Omega} \frac{f_0(x)}{M \cdot h_0(x)} \leq 1.$$

  Ideally $M$ should be as small as possible.

- Standard rejection sampling algorithm:
  1. Draw $u$ from Uniform$(0, 1)$.
  2. Draw $x$ from proposal $h$.
  3. If $u \leq f_0(x)/\{M \cdot h_0(x)\}$, accept $x$ as a draw from $f$; otherwise return to Step 1.

  Repeat until desired number of draws are obtained, or bail out if there are too many rejections.

# Acceptance Probability

- The probability of accepting a proposed $x$, with accompanying $u$, is

$$\begin{aligned}
\mathrm{P}\left( U \leq \frac{f_0(X)}{M h_0(X)} \right) &= \mathrm{P}\left( U \leq \frac{\psi f(X)}{a M h(X)} \right) \\
&= \mathrm{E}_X \, \mathrm{E}_U \, \mathrm{I}\left( U \leq \frac{\psi f(X)}{a M h(X)} \;\middle|\; X \right) \\
&= \mathrm{E}_X \left[ \mathrm{P}\left( U \leq \frac{\psi f(X)}{a M h(X)} \;\middle|\; X \right) \right] \\
&= \mathrm{E}_X \left[ \frac{\psi f(X)}{a M h(X)} \right] \\
&= \int_\Omega \frac{\psi f(x)}{a M h(x)} h(x) d\nu(x) \\
&= \frac{\psi}{a M} \int_\Omega f(x) d\nu(x) \\
&= \frac{\psi}{a M}.
\end{aligned}$$

- The number of draws until one acceptance occurs is a random variable $S \sim \text{Geometric}(\psi / \{aM\})$. Expected value of $S$ is $\mathrm{E}(S) = aM/\psi$.

# Distribution of an Accepted Draw

- For a measureable $A \subseteq \Omega$, and $X \sim h$, we have

$$\begin{aligned}
\mathrm{P}(X \in A \mid \text{Accept}) &= \mathrm{P}(X \in A \mid U \leq f_0(X)/\{M \cdot h_0(X)\}) \\
&= \frac{\mathrm{P}\left(X \in A, U \leq \frac{f_0(X)}{M \cdot h_0(X)}\right)}{\mathrm{P}\left(U \leq \frac{f_0(X)}{M \cdot h_0(X)}\right)} \\
&= \frac{aM}{\psi} \mathrm{P}\left(X \in A, U \leq \frac{f_0(X)}{M \cdot h_0(X)}\right) \\
&= \frac{aM}{\psi} \int \mathrm{I}(x \in A) \left[ \int_0^{f_0(x)/\{M \cdot h_0(x)\}} du \right] h(x) d\nu(x) \\
&= \frac{aM}{\psi} \int \mathrm{I}(x \in A) \frac{f_0(x)}{M h_0(x)} h(x) d\nu(x) \\
&= \frac{aM}{\psi} \int \mathrm{I}(x \in A) \frac{\psi}{aM} \frac{f(x)}{h(x)} h(x) d\nu(x) \\
&= \int_A f(x) d\nu(x).
\end{aligned}$$

- Therefore, an accepted draw $x$ is from the target distribution.

# Formulating a Proposal

- The stock rejection sampling method does not include a recipe for $h$.
- With a poor choice of $h$, acceptances could be very rare so that the algorithm may not be practical.
- It may be nontrivial to design an efficient customized proposal. A few articles dedicated (at least partially) to this purpose are:
  1. the von Mises Fisher distribution (Ulrich, 1984; Wood, 1994);
  2. the Polynomial-Normal distribution (Evans and Swartz, 1998);
  3. the Bessel distribution (Devroye, 2002); and
  4. the Conway-Maxwell Poisson distribution (Chanialidis et al., 2018; Benson and Friel, 2021).

# Formulating a Proposal

- The Adaptive Rejection Sampling (ARS) method automatically adapts to the target using rejected draws, but requires that the target is log-concave (Gilks and Wild, 1992).
- The Adaptive Rejection Metropolis Sampling (ARMS) drops the log-concave restriction and uses a Metropolis step (Gilks et al., 1995).
  1. However, it produces a chain of non-independent draws.
  2. Proposal is not guaranteed to converge to the target with adaptations.
- The Independent Doubly Adaptive Rejection Metropolis Sampling (IA2RMS) algorithm addresses the ARMS convergence issue and reduces dependence (Martino et al., 2015).
- Another approach has been to relax log-concavity in ARS to something less restrictive. E.g., Evans and Swartz (1998) permit some transformation of the target to be concave.

# Formulating a Proposal

- The vertical strip method is a constructive way to make a proposal and does not require log-concavity.
- This method is discussed in Devroye (1986, Chapters II and VIII) and Martino et al. (2018, Chapter 3).
- Some recent works refer to this as “Ahrens method” based on Ahrens (1993) and Ahrens (1995).
- Karawatzki (2006) discusses vertical strips with multivariate targets.

# Vertical Strip Method

- Suppose the support is a bounded interval $\Omega = (a, b]$.
- Let $\alpha_0 \leq \cdots \leq \alpha_N$ be knots where $\alpha_0 \equiv a$ and $\alpha_N \equiv b$ are fixed.
- The intervals $\mathcal{D}_j = (\alpha_{j-1}, \alpha_j]$, $j = 1, \ldots, N$, partition the support.
- With $\overline{f}_{0j} = \max_{x \in \mathcal{D}_j} f_0(x)$, we have

$$f_0(x) \leq \sum_{j=1}^{N} \mathrm{I}(x \in \mathcal{D}_j) \cdot \overline{f}_{0j} \overset{\text{def}}{=} h_0(x).$$

- With normalizing constant $a = \sum_{\ell=1}^{N} \overline{\xi}_\ell$ and $\overline{\xi}_j = \overline{f}_{0j} \cdot (\alpha_j - \alpha_{j-1})$, the function $h_0(x)$ normalizes to a finite mixture of Uniforms:

$$h(x) = \sum_{j=1}^{N} \pi_j g_j(x), \quad \pi_j = \frac{\overline{\xi}_j}{\sum_{\ell=1}^{N} \overline{\xi}_\ell}, \quad g_j(x) = \frac{\mathrm{I}(x \in \mathcal{D}_j)}{\alpha_j - \alpha_{j-1}}.$$

- Take $h_0$ as an envelope and $h$ as a proposal for rejection sampling. We can use $M = \sup_{x \in \Omega} f_0(x)/h_0(x) \equiv 1$.
- Hörmann (2002) studies several knot selection strategies, such as "equally spaced" and "equal probabilities".

# Weighted Target Distributions

- Consider the density:

$$f(x) = f_0(x)/\psi, \quad f_0(x) = w(x)g(x), \quad \psi = \int_\Omega w(x)g(x)d\nu(x),$$

  where

  1. $\Omega$ is the support of $f$,
  2. $w(x) \geq 0$ is the "weight function",
  3. $g(x)$ is the "base distribution" density with $\Omega \subseteq \operatorname{supp} g$,
  4. $\psi$ is a normalizing constant (may be intractable),
  5. $\nu$ is a dominating measure.

- Many distributions encountered in practice have this form. E.g., in Bayesian methodology.
- Idea: partition the support and relax ("majorize") the weight function on each region.
  1. More flexible ✓
  2. More involved.
  3. More efficient?

# Weighted Vertical Strips

- Let $\overline{w}_j(x)$ be a function that dominates $w(x)$ on $\mathcal{D}_j$; i.e.,

$$w(x) \leq \overline{w}_j(x), \quad \text{for all } x \in \mathcal{D}_j.$$

- We have

$$f_0(x) = w(x)g(x) \leq \sum_{j=1}^{N} \mathrm{I}(x \in \mathcal{D}_j) \cdot \overline{w}_j(x)g(x) \stackrel{\text{def}}{=} h_0(x).$$

- Let $\overline{\xi}_j = \mathrm{E}[\overline{w}_j(T)\,\mathrm{I}(T \in \mathcal{D}_j)]$ with $T \sim g$.
- Normalizing function $h_0(x)$ with $a = \sum_{\ell=1}^{N} \overline{\xi}_\ell$ yields

$$h(x) = \sum_{j=1}^{N} \pi_j g_j(x), \quad \pi_j = \frac{\overline{\xi}_j}{\sum_{\ell=1}^{N} \overline{\xi}_\ell}, \quad g_j(x) = \frac{\overline{w}_j(x)g(x) \cdot \mathrm{I}(x \in \mathcal{D}_j)}{\mathrm{E}[\overline{w}_j(T)\,\mathrm{I}(T \in \mathcal{D}_j)]}.$$

- As before, $h_0$ serves as an envelope for rejection sampling, $h$ as a proposal, and $M = \sup_{x \in \Omega} f_0(x)/h_0(x) \equiv 1$.

# Rejection Probability

- Probability of rejecting a proposed draw is $1 - \psi / a$.
- This can be reduced by decomposing into more regions and/or taking better choices of $\overline{w}_j(x)$.
- When $\psi$ is intractable, the following bound may be useful.
- Suppose $\underline{w}_j(x)$ is another function such that $\underline{w}_j(x) \leq w(x)$ for all $x \in \mathcal{D}_j$, and let $\underline{\xi}_j = \mathrm{E}[\underline{w}_j(T) \, \mathrm{I}(T \in \mathcal{D}_j)]$.
- We have

$$
\begin{aligned}
1 - \frac{\psi}{a} &= 1 - \frac{1}{a} \sum_{j=1}^{N} \int_{\mathcal{D}_j} w(x) g(x) d\nu(x) \\
&\leq 1 - \frac{1}{a} \sum_{j=1}^{N} \int_{\mathcal{D}_j} \underline{w}_j(x) g(x) d\nu(x) \\
&= \frac{\sum_{j=1}^{N} (\overline{\xi}_j - \underline{\xi}_j)}{\sum_{j=1}^{N} \overline{\xi}_j} = 1 - \frac{\sum_{j=1}^{N} \underline{\xi}_j}{\sum_{j=1}^{N} \overline{\xi}_j}.
\end{aligned}
$$

# Choice of Weight Function

- There is some flexibility to define $w(x)$ versus $g(x)$, and to choose $\overline{w}_j(x)$.

- If $q(x)$ is another positive function,

$$f(x) \propto w(x)g(x) = \underbrace{w(x)q(x)}_{\tilde{w}(x)} \underbrace{\frac{1}{q(x)}g(x)}_{\propto \tilde{g}(x)}.$$

- Some examples are given later where this kind of refactoring is useful.

- For $\overline{w}_j(x)$, we would like:
  1. $\overline{w}_j(x)$ not too much larger than $w(x)$;
  2. easy to compute $\overline{\xi}_j = \mathrm{E}[\overline{w}_j(T)\, \mathrm{I}(T \in \mathcal{D}_j)]$;
  3. easy to draw from $g_j(x) = \frac{\overline{w}_j(x)g(x) \cdot \mathrm{I}(x \in \mathcal{D}_j)}{\mathrm{E}[\overline{w}_j(T)\, \mathrm{I}(T \in \mathcal{D}_j)]}$.

- Note that vertical strips is a special case of weighted vertical strips with $w(x) = f_0(x)$ and $g(x)$ is Uniform($\Omega$). (Assuming $\Omega$ is bounded).

# Choice of Weight Function

- We will focus on a specific choice for the remainder of the talk:

$$\overline{w}_j = \max_{x \in \mathcal{D}_j} w(x), \quad \underline{w}_j = \min_{x \in \mathcal{D}_j} w(x).$$

- These may be computed analytically in some problems, but we will use numerical optimization.
- Let $\overline{\xi}_j = \overline{w}_j \, \mathrm{P}(T \in \mathcal{D}_j)$ and $\underline{\xi}_j = \underline{w}_j \, \mathrm{P}(T \in \mathcal{D}_j)$ with $T \sim g$.
- Normalizing $h_0(x)$ with $a = \sum_{\ell=1}^N \overline{\xi}_\ell$ yields

$$h(x) = \sum_{j=1}^N \pi_j g_j(x), \quad \pi_j = \frac{\overline{\xi}_j}{\sum_{\ell=1}^N \overline{\xi}_\ell}, \quad g_j(x) = \frac{g(x) \, \mathrm{I}(x \in \mathcal{D}_j)}{\mathrm{P}(T \in \mathcal{D}_j)}.$$

- This $h$ is often not difficult to work with in the univariate case. Suppose $G$ and $G^-$ are the CDF and quantile functions for $g$.
  1. $\mathrm{P}(T \in \mathcal{D}_j) = G(\alpha_j) - G(\alpha_{j-1})$.
  2. A draw $x$ from $g_j$ is obtained as $G^-\big(\{G(\alpha_j) - G(\alpha_{j-1})\}u + G(\alpha_{j-1})\big)$, with $u$ drawn from Uniform$(0, 1)$.

# Adapting the Proposal

- We decompose $\Omega$ into $\mathcal{D}_1, \ldots, \mathcal{D}_N$ before sampling, with $N$ prespecified.
- It is also possible to adapt using rejected draws, but we currently do not.
- With $N_0$ current regions, split one of the regions.
    1. Draw an index

$$j = \begin{cases} 1, & \text{with probability} \propto \overline{\xi}_1 - \underline{\xi}_1, \\ & \vdots \\ N_0, & \text{with probability} \propto \overline{\xi}_{N_0} - \underline{\xi}_{N_0}. \end{cases}$$

    2. Bifrucate region $j$ at its midpoint.

  Repeat to get $N$ regions or until rejection probability (or its upper bound) is sufficiently small.

- Special handling is needed for intervals where:
    1. one or both limits are infinite—to find a suitable bifrucation point; or
    2. where support is discrete—so that each region contains at least one support point.

# Example: Polynomial-Normal Distribution

# Polynomial-Normal Distribution

- The Polynomial-Normal distribution (Evans and Swartz, 1994, 1998) has density

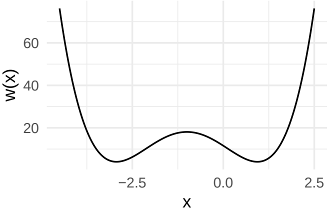
$$f(x) = \frac{\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} \cdot \prod_{\ell=1}^{m} (x - \lambda_\ell)(x - \bar{\lambda}_\ell)}{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-s^2/2\} \left[ \prod_{\ell=1}^{m} (s - \lambda_\ell)(s - \bar{\lambda}_\ell) \right] ds}, \quad x \in \mathbb{R},$$

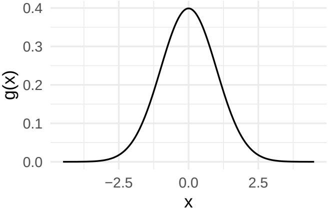
  based on a non-negative polynomial of degree $2m$ where $(\lambda_\ell, \bar{\lambda}_\ell)$ are pairs of roots which are complex conjugates.

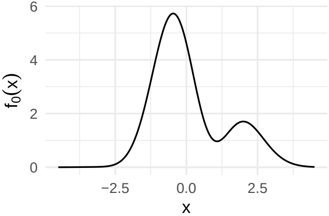
- To make this a concrete example, let $m = 2$ with $\lambda_1 = 1 + 0.5i$ and $\lambda_2 = -3 + 0.5i$.

# Rejection Sampler

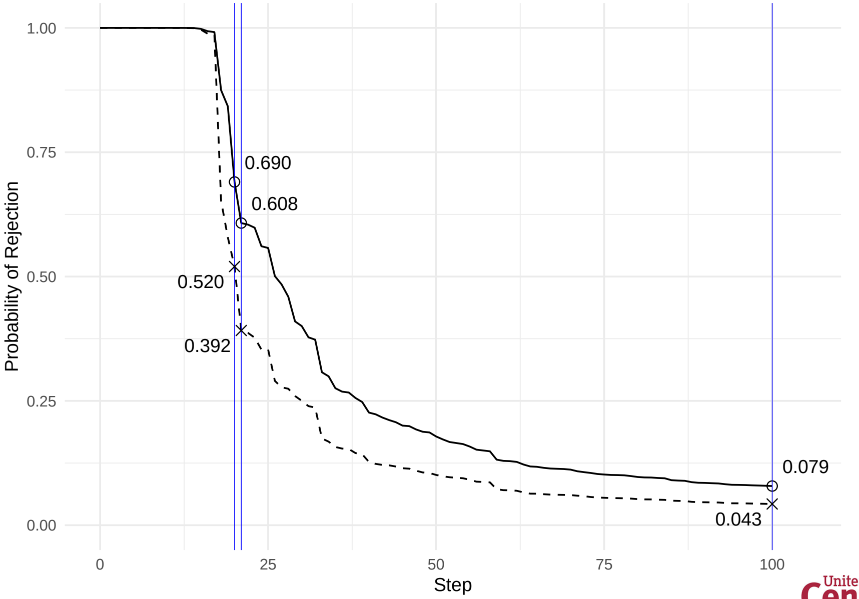
$$f_0(x) = \underbrace{\frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\}}_{g(x)} \cdot \underbrace{\prod_{\ell=1}^{m} (x - \lambda_\ell)(x - \bar{\lambda}_\ell)}_{w(x)}, \quad x \in \mathbb{R}$$



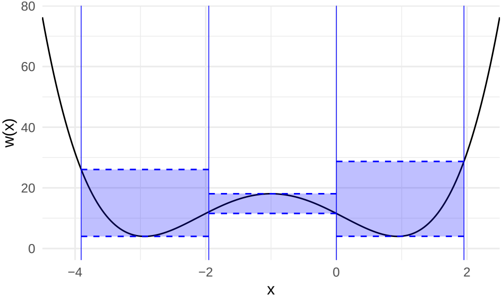


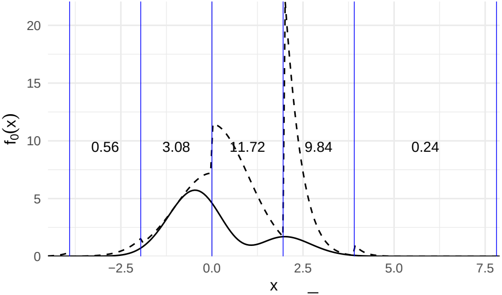


# Adapting the Proposal



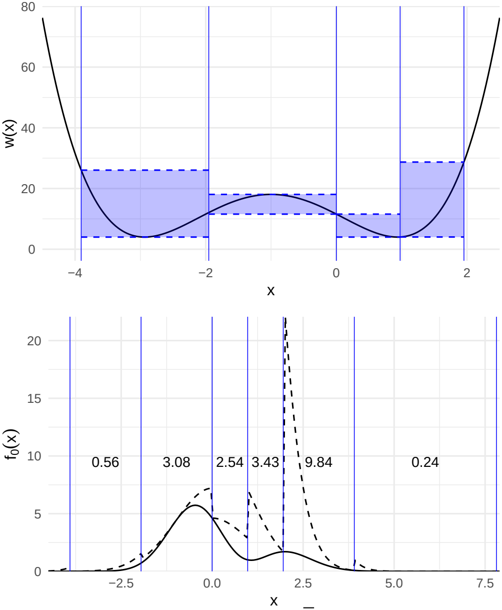
# Proposal with $N = 20$





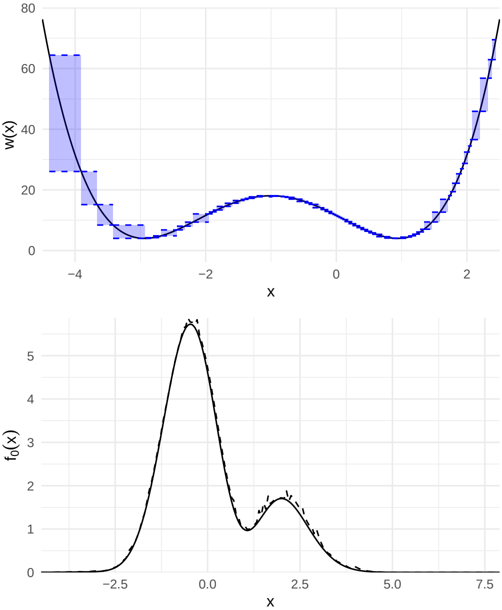
Displayed value is $\overline{\xi}_j - \underline{\xi}_j$.

# Proposal with $N = 21$

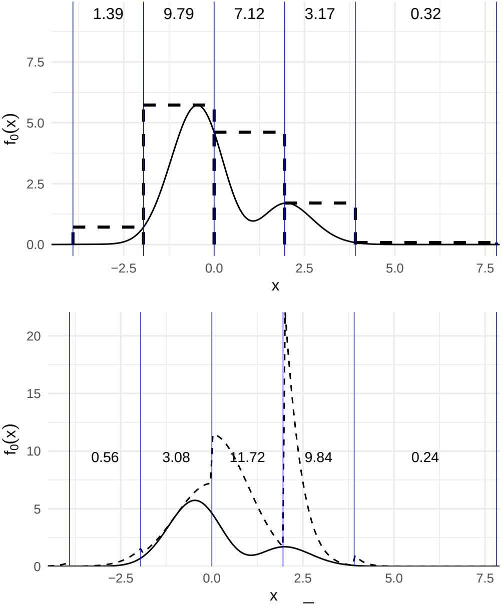


Displayed value is $\overline{\xi}_j - \underline{\xi}_j$.

# Proposal with $N = 100$



# Vertical Strip Proposal with $N = 20$



Displayed value is $\overline{\xi}_j - \underline{\xi}_j$.

# Normalizing Constant

- The rejection sampler can provide an approximation of normalizing constant $\psi$ with $a$.
- The probability of rejecting one proposed draw, $1 - \psi/a \equiv \frac{a-\psi}{a}$, can be interpreted as a relative error.
- From the construction of $a$, we have $a - \psi \geq 0$. Combining this with the upper bound,

$$0 \leq \frac{a - \psi}{a} \leq \frac{\sum_{j=1}^{N}(\overline{\xi}_j - \underline{\xi}_j)}{\sum_{j=1}^{N} \overline{\xi}_j}.$$

- Can also show that

$$|\,\mathrm{P}(Y \in A) - \mathrm{P}(X \in A)| \leq \frac{a - \psi}{a},$$

for $X \sim f$, $Y \sim h$, and any measureable $A \subseteq \Omega$.

# Normalizing Constant for Polynomial-Normal

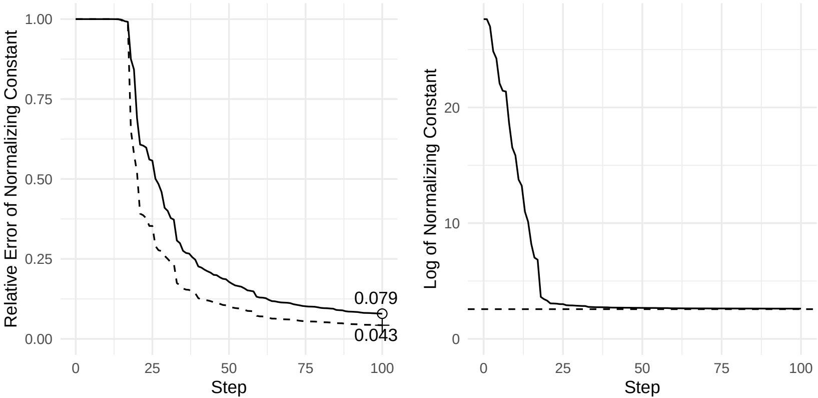
- Evans and Swartz (1994) note that the normalizing constant can be computed with quadrature, or by expanding the polynomial and calculating Normal moments.
- With quadrature,

$$\begin{aligned}\psi &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \left[ \prod_{\ell=1}^{m} (x - \lambda_\ell)(x - \bar{\lambda}_\ell) \right] dx \\ &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-z^2} \phi(z) dz \\ &= \frac{1}{\sqrt{\pi}} \sum_{\ell=1}^{m+1} \alpha_\ell \phi(x_\ell),\end{aligned}$$

  where $\alpha_1, \ldots, \alpha_{m+1}$ are Gauss-Hermite quadrature weights, $x_1, \ldots, x_{m+1}$ are corresponding nodes, and $\phi(x) = \sqrt{2} \cdot \prod_{\ell=1}^{m} (x - \lambda_\ell)(x - \bar{\lambda}_\ell)$.

- We get exact equality because $\phi(x)$ is a polynomial of degree $2m$.
- Quadrature weights and nodes can be obtained, for example, using the statmod package in R (Smyth, 2005).

# Normalizing Constant for Polynomial-Normal



- $\psi = 13.0625$.
- With $N = 100$ regions, $a = 13.6448$ and $(a - \psi)/a = 0.043$.

# Example: Conway-Maxwell Poisson

# Conway-Maxwell Poisson

- The Conway-Maxwell Poisson (CMP) distribution has become popular for modeling count data which may exhibit over- and/or underdispersion, relative to Poisson.
- The monograph by Sellers (2023) gives an overview of CMP and a number of recent developments.
- The R package COMPoissonReg (Raim and Sellers, 2022) implements basic CMP distribution functions and regression.

# CMP Distribution

- A random variable $X$ with distribution CMP$(\lambda, \nu)$ has probability mass function (pmf)

$$f(x) = \frac{\lambda^x}{(x!)^\nu Z(\lambda, \nu)}, \quad x = 0, 1, 2, \ldots, \quad Z(\lambda, \nu) = \sum_{x=0}^{\infty} \frac{\lambda^x}{(x!)^\nu}$$

  where $\lambda \geq 0$ and $\nu \geq 0$.

- The CMP$(\lambda, \nu)$ family includes some cases of interest.
  1. When $\nu = 1$, it corresponds to Poisson$(\lambda)$. Here variance and mean are both $\lambda$.
  2. When $\nu < 1$, it becomes overdispersed so that the variance is larger than the mean. At the extreme $\nu = 0$, it corresponds to Geometric$(1 - \lambda)$.
  3. When $\nu > 1$, it becomes underdispersed so that the variance is smaller than the mean. As $\nu \to \infty$, it becomes Bernoulli$(\lambda/(1 + \lambda))$.

# Sampling from CMP

- Generating variates from CMP is non-trivial because of $Z(\lambda, \nu)$.

- It does not have a closed form and its magnitude can vary wildly with $\lambda$ and $\nu$. The mass of the distribution can shift accordingly.

- For example, let $\lambda = 2$.
  1. If $\nu = 1$, then $Z(\lambda, \nu) = e^2$ and $\mathsf{E}(X) = 2$.
  2. If $\nu = 0.05$, $Z(\lambda, \nu) \approx \exp(52{,}437.76)$ and $\mathsf{E}(X) = 1{,}048{,}585$.

# Sampling from CMP

- Motivating works by Chanialidis et al. (2018) and Benson and Friel (2021) use rejection sampling.
    1. Bayesian analysis of CMP parameters.
    2. Exchange algorithm: Metropolis sampler with data augmentation to avoid computing intractable normalizing $Z(\lambda, \nu)$ constant.
    3. Requires method to draw exactly from CMP.
    4. They develop custom rejection sampling algorithms.

# Sampling from CMP

- COMPoissonReg works by either truncating or approximating $Z(\lambda, \nu)$.
- If $\lambda^{-1/\nu}$ is small, use an approximation (Shmueli et al., 2005)

$$Z(\lambda, \nu) = \frac{\exp(\nu \lambda^{1/\nu})}{\lambda^{(\nu-1)/2\nu} (2\pi)^{(\nu-1)/2} \nu^{1/2}} \left\{ 1 + O(\lambda^{-1/\nu}) \right\}.$$

- Otherwise, truncate the series $Z(\lambda, \nu)$ to a finite sum $Z^{(m)}(\lambda, \nu) = \sum_{r=0}^{m} \frac{\lambda^r}{(r!)^\nu}$.
- The remainder

$$Z(\lambda, \nu) - Z^{(m)}(\lambda, \nu) = \sum_{r=m+1}^{\infty} \frac{\lambda^r}{(r!)^\nu}$$

  can be bounded above by a convergent geometric series; $m$ can then be chosen large enough to achieve a desired tolerance.
- The approximated $Z(\lambda, \nu)$ is used to do a brute force calculation of the quantile function `qcmp(p, lambda, nu)`.
- The function `rcmp` generates a variate using `qcmp(u, lambda, nu)` with $u$ generated from Uniform$(0, 1)$.

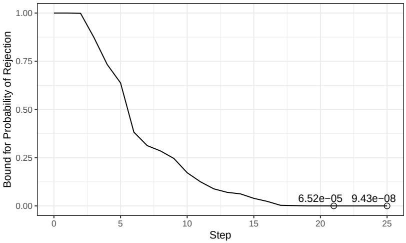
# Rejection Sampler: Underdispersion Case

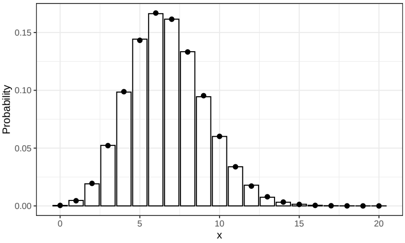
- For the case $\nu \geq 1$, let $g$ be the pmf of Geometric$(1/\{1+\lambda\})$ so that

$$f(x) \propto \frac{\lambda^x}{(x!)^\nu} = \underbrace{\left(\frac{\lambda}{1+\lambda}\right)^x \frac{1}{1+\lambda}}_{g(x)} \underbrace{(1+\lambda)^{x+1} \frac{\lambda^x}{(x!)^\nu}}_{w(x)}.$$

- Let $\lambda = 10$ and $\nu = 1.2$.

- Sampler with $N = 21$ regions rejected 5 proposed draws to obtain 100,000 variates (rejection rate 0.005%).





# Rejection Sampler: Overdispersion Case

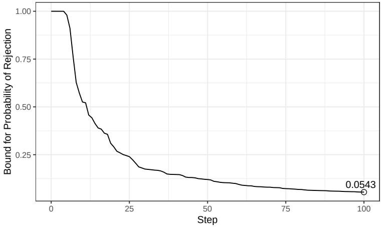
- For $\nu < 1$, Geometric$(1/\{1+\lambda\})$ may be an inefficient base because its mass can be practically disjoint from CMP$(\lambda, \nu)$.

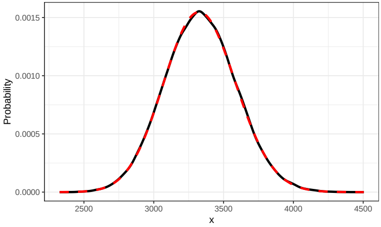
- Here let $\mu = \lambda^{1/\nu}$ and

$$f(x) \propto \frac{\mu^{\nu x}}{(x!)^{\nu}} = \underbrace{\left(\frac{\mu}{1+\mu}\right)^{x} \frac{1}{1+\mu}}_{g(x)} \underbrace{(1+\mu)^{x+1} \frac{\mu^{x(\nu-1)}}{(x!)^{\nu}}}_{w(x)}.$$

- For $\lambda = 1.5$ and $\nu = 0.05$.

- Sampler with $N = 101$ regions rejected 2,922 proposed draws to obtain 100,000 variates (rejection rate 2.84%).

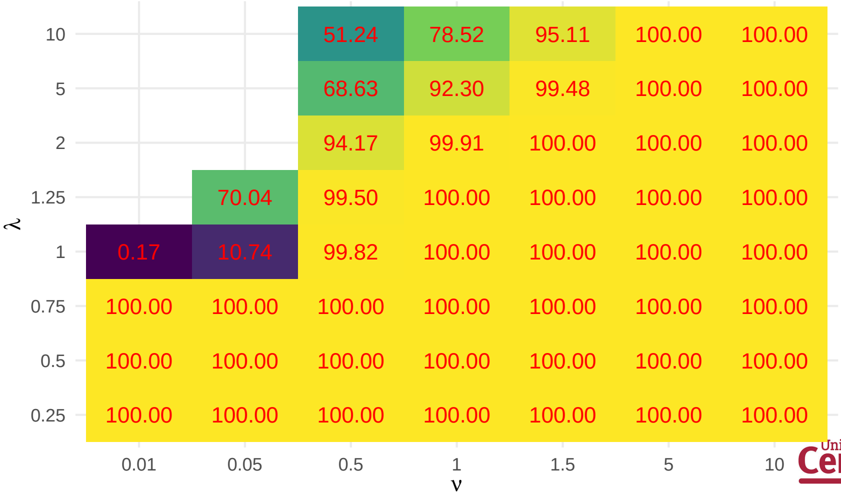




# Acceptance Study

A brief study of acceptance % with $R = 100{,}000$ and $N = 10$:

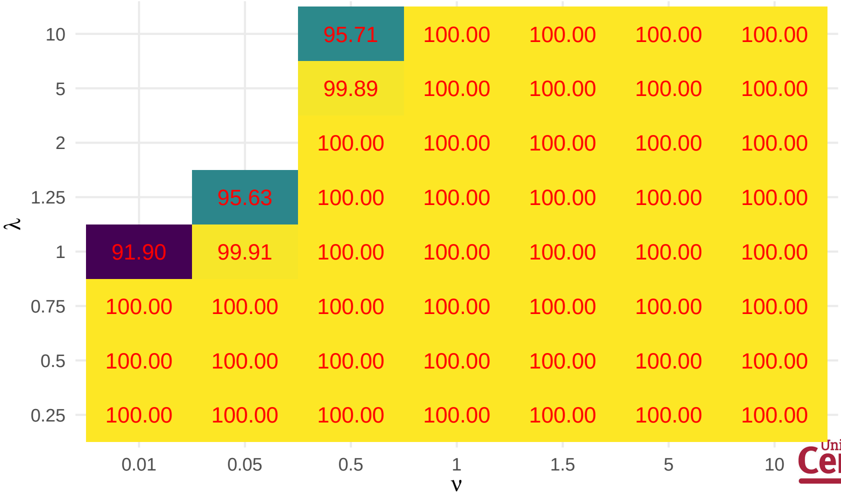
- $\lambda \in \{0.25, 0.5, 0.75, 1, 1.25, 2, 5, 10\}$,
- $\nu \in \{0.01, 0.05, 0.5, 1, 1.5, 5, 10\}$,
- where $\lambda^{1/\nu} \leq 50,000$.



# Acceptance Study

A brief study of acceptance % with $R = 100{,}000$ and $N = 50$:

- $\lambda \in \{0.25, 0.5, 0.75, 1, 1.25, 2, 5, 10\}$,
- $\nu \in \{0.01, 0.05, 0.5, 1, 1.5, 5, 10\}$,
- where $\lambda^{1/\nu} \leq 50,000$.



| λ \ ν | 0.01 | 0.05 | 0.5 | 1 | 1.5 | 5 | 10 |
|---|---|---|---|---|---|---|---|
| 10 | | | 95.71 | 100.00 | 100.00 | 100.00 | 100.00 |
| 5 | | | 99.89 | 100.00 | 100.00 | 100.00 | 100.00 |
| 2 | | | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 1.25 | | 95.63 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 1 | 91.90 | 99.91 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 0.75 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 0.5 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| 0.25 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |

# Example: Gaussian Process Regression

# Gaussian Process Regression

- Suppose $\mu(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is a function whose form may be unknown.

- Data are $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $y_i$ is a noisy observation of $\mu(\boldsymbol{x}_i)$.

- Consider the GP model

$$y_i = \mu(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2), \quad i = 1, \ldots, n,$$
$$\mu \sim \text{GP}(0, k(\cdot, \cdot)), \quad \sigma^2 \sim \text{Gamma}(a_\sigma, b_\sigma),$$

shape $a_\sigma$, rate $b_\sigma$, and covariance kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\{-\frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\}$.

- Likelihood portion of the model in vector form is

$$\boldsymbol{y} = \mu(\boldsymbol{X}) + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \text{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \quad \mu(\boldsymbol{X}) \sim \text{N}\big(\boldsymbol{0}, k(\boldsymbol{X}, \boldsymbol{X})\big),$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \; \mu(\boldsymbol{X}) = \begin{bmatrix} \mu(\boldsymbol{x}_1) \\ \vdots \\ \mu(\boldsymbol{x}_n) \end{bmatrix}, \; k(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \ddots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix}.$$

- Using the proposed rejection sampler, we can draw exactly from the posterior distribution $[\sigma^2 \mid \boldsymbol{y}]$ without MCMC.

# Rejection Sampler

- Use transformation to avoid repeating matrix operations in the sampler.
- Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $k(\boldsymbol{X}, \boldsymbol{X})$, with $\boldsymbol{\Lambda} = \mathrm{Diag}(\lambda_1, \ldots, \lambda_n)$, so that

$$\begin{aligned}\mathrm{Var}(\boldsymbol{y} \mid \sigma^2) &= \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}) \\ &= \boldsymbol{U}[\sigma^2 \boldsymbol{I} + \boldsymbol{\Lambda}]\boldsymbol{U}^\top.\end{aligned}$$

- Transform the marginal likelihood of $\boldsymbol{y} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}))$ to $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$ where $z_i \overset{\text{ind}}{\sim} \mathrm{N}(0, \sigma^2 + \lambda_i)$.
- Let weight function be the unnormalized posterior with respect to $\boldsymbol{z}$:

$$\log w(\sigma^2) = -\frac{1}{2}\sum_{i=1}^{n} \log(\sigma^2 + \lambda_i) - \frac{1}{2}\sum_{i=1}^{n} \frac{z_i^2}{\sigma^2 + \lambda_i} \\ + (a_\sigma - 1)\log \sigma^2 - b_\sigma \sigma^2.$$

- Take base distribution $g$ as the density of $\mathrm{Uniform}(0, 1000)$.
- A similar of formulation of the rejection sampler can be used with other priors on $\sigma^2$ and other covariance kernels (with fixed hyperparameters).

# Posterior Predictive Distribution

- Let $\boldsymbol{X}_0$ be (potentially new) inputs:

$$\boldsymbol{X}_0 = \begin{bmatrix} \boldsymbol{x}_{01}^\top \\ \vdots \\ \boldsymbol{x}_{0n_0}^\top \end{bmatrix}, \quad \text{so that} \quad \mu(\boldsymbol{X}_0) = \begin{bmatrix} \mu(\boldsymbol{x}_{01}) \\ \vdots \\ \mu(\boldsymbol{x}_{0n0}) \end{bmatrix}.$$

- To sample from the posterior predictive distribution $[\mu(\boldsymbol{X}_0) \mid \boldsymbol{y}]$.
  1. Draw $\sigma^{2(r)}$, $r = 1, \ldots, R$, from the posterior.
  2. Draw $\mu(\boldsymbol{X}_0)$ from $[\mu(\boldsymbol{X}_0) \mid \sigma^2, \boldsymbol{y}]$ for each $\sigma^2 = \sigma^{2(r)}$.

- The distribution $[\mu(\boldsymbol{X}_0) \mid \sigma^2, \boldsymbol{y}]$ can be obtained as $\text{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, where

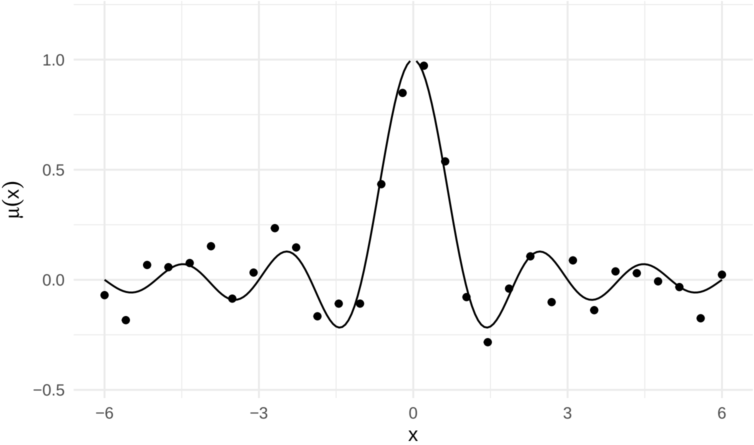
$$\begin{aligned} \boldsymbol{\mu}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1} \boldsymbol{y}, \\ \boldsymbol{\Sigma}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X}_0) - k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_0). \end{aligned}$$

- This is be obtained from

$$[\boldsymbol{y}, \mu(\boldsymbol{X}_0) \mid \sigma^2] \sim \text{N}\left( \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}) & k(\boldsymbol{X}, \boldsymbol{X}_0) \\ k(\boldsymbol{X}_0, \boldsymbol{X}) & k(\boldsymbol{X}_0, \boldsymbol{X}_0) \end{bmatrix} \right).$$

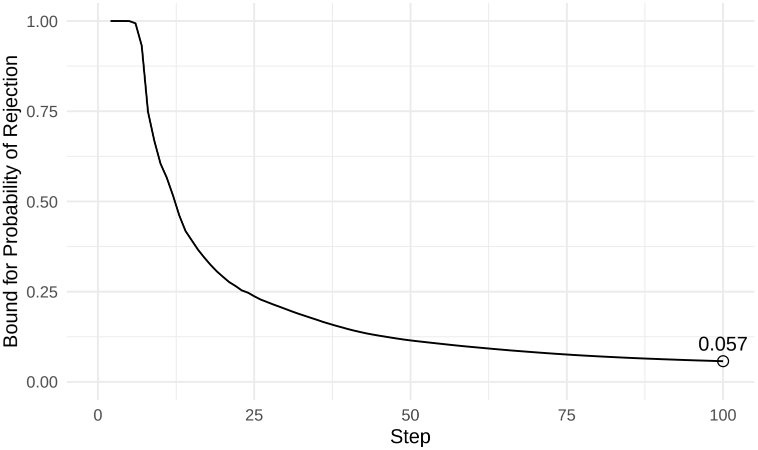
# Example

- To learn about the sinc function $\mu(x) = \text{sin}(\pi x)/\{\pi x\}$.
- Observe $n = 25$ outcomes $y_i = \mu(x_i) + \epsilon_i$ with $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, with
  1. $\sigma^2 = 0.1^2$,
  2. $x_i$ on an evenly spaced grid in $[-6, 6]$,
- Choose hyperparameters $a_\sigma = 2$ and $b_\sigma = 1/2$.



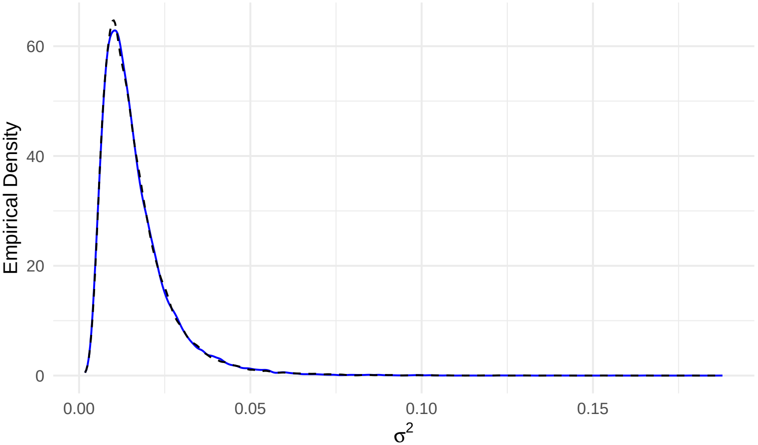
# Adapting the Proposal

- With $N = 101$ regions, 1,502 proposed draws were rejected to obtain $R = 50{,}000$ (rejection rate 2.92%)



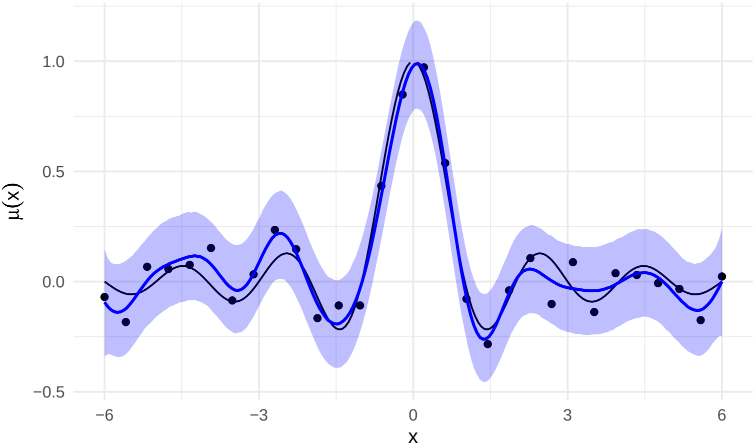
# Posterior Draws

Empirical distribution of $R$ draws from rejection sampler (solid blue) to $R$ draws computed via Stan (Carpenter et al., 2017) with No U-Turn sampler (dashed black).



# Posterior Predictive Results

Posterior predictive mean of $\mu(x)$ (blue curve) for $x$ on a fine grid on $[-6, 6]$, and associated 95% pointwise interval from 0.025 and 0.975 quantiles (blue shaded area).



# Spatial Linear Regression

- The spatial linear regression model presented in Chapter 6 of Banerjee et al. (2015) is an application of the GP.

- Suppose $\boldsymbol{x}_i$ are locations on a spatial domain with fixed covariate $\boldsymbol{s}(\boldsymbol{x}_i) \in \mathbb{R}^m$ and observation $y_i$, $i = 1, \ldots, n$, and

$$
\begin{aligned}
&y_i = \boldsymbol{s}(\boldsymbol{x}_i)^\top \boldsymbol{\beta} + \zeta(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \\
&\zeta \sim \mathrm{GP}(0, k(\cdot, \cdot)), \quad \boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}), \quad \sigma^2 \sim \mathrm{Gamma}(a_\sigma, b_\sigma).
\end{aligned}
$$

- With kernel $k(\cdot, \cdot)$ completely specified, we can draw from the exact posterior using the proposed rejection sampler.

- The R package spBayes (Finley et al., 2007) considers a fully conjugate variation of this model with $\sigma^2$ fixed. Also, full Bayesian treatments of more general variants with MCMC via Metropolis-Hastings.

# Spatial Linear Regression

- Marginally, $[\boldsymbol{y} \mid \sigma^2] \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} + \sigma_\beta^2 \boldsymbol{S}\boldsymbol{S}^\top + k(\boldsymbol{X}, \boldsymbol{X}))$, where $\boldsymbol{S}$ is a matrix with $\boldsymbol{s}(\boldsymbol{x}_i)$ as the $i$th row.

- Let $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be the spectral decomposition of $\sigma_\beta^2 \boldsymbol{S}\boldsymbol{S}^\top + k(\boldsymbol{X}, \boldsymbol{X})$, and consider the posterior with respect to data $\boldsymbol{z} = \boldsymbol{U}^\top \boldsymbol{y}$ where $z_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2 + \lambda_i)$ as before.

- Draws of $\beta$ can be recovered from $[\beta \mid \sigma^2, \boldsymbol{y}] \propto [\boldsymbol{y} \mid \beta, \sigma^2] \cdot [\beta]$ using conjugacy of $\mathrm{N}(\boldsymbol{y} \mid \boldsymbol{S}\beta, \sigma^2 \boldsymbol{I} + k(\boldsymbol{X}, \boldsymbol{X}))$ and $\mathrm{N}(\beta \mid \boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$.

- Draws of $\zeta(\boldsymbol{X}_0)$ from posterior predictive distribution $[\zeta(\boldsymbol{X}_0) \mid \boldsymbol{y}]$ may be obtained using $[\zeta(\boldsymbol{X}_0) \mid \beta,\ \sigma^2, \boldsymbol{y}] \equiv \mathrm{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$,

$$\begin{aligned}
\boldsymbol{\mu}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1}(\boldsymbol{y} - \boldsymbol{S}\beta), \\
\boldsymbol{\Sigma}_0 &= k(\boldsymbol{X}_0, \boldsymbol{X}_0) - k(\boldsymbol{X}_0, \boldsymbol{X})[k(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2 \boldsymbol{I}]^{-1} k(\boldsymbol{X}, \boldsymbol{X}_0).
\end{aligned}$$

# Conclusions

- We have revisited the vertical strip method of rejection sampling.
- Some additional flexibility is possible for weighted distributions.
- We presented several examples from the literature.
- Such rejection samplers may be useful within Gibbs samplers for multivariate targets.
    1. An example is in Bayesian modeling of noisy tabulations for disclosure avoidance.
    2. Distribution of noise mechanism and regression model may not be conjugate.
    3. Rejection sampling can be used as a step within a Gibbs sampler (Raim, 2021; Irimata et al., 2022).
- Several applications of the rejection sampler involving (joint) multivariate targets are currently being considered.
- Our adaptation method is fast and easy to compute, but improvements in efficiency (reduced rejection rates for smaller $N$) are possible.

Thank you!

✉ andrew.raim@census.gov

# References I

J H. Ahrens. Sampling from general distributions by suboptimal division of domains. *Grazer Mathematische Berichte*, 319:20, 1993.

J.H. Ahrens. A one-table method for sampling from continuous and discrete distributions. *Computing*, 54(20):127–146, 1995.

S. Banerjee, B. P. Carlin, and A. E. Gelfand. *Hierarchical modeling and analysis for spatial data*. CRC Press, 2nd edition, 2015.

Alan Benson and Nial Friel. Bayesian Inference, Model Selection and Likelihood Estimation using Fast Rejection Sampling: The Conway-Maxwell-Poisson Distribution. *Bayesian Analysis*, 16(3):905–931, 2021.

Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Charalampos Chanialidis, Ludger Evers, Tereza Neocleous, and Agostino Nobile. Efficient Bayesian inference for COM-Poisson regression models. *Statistics and Computing*, 23: 595–608, 2018.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer, 1986.

Luc Devroye. Simulating Bessel random variables. *Statistics & Probability Letters*, 57(3): 249–257, 2002.

M. Evans and T. Swartz. Distribution theory and inference for polynomial-normal densities. *Communications in Statistics—Theory and Methods*, 23(4):1123–1148, 1994.

# References II

M. Evans and T. Swartz. Random variable generation using concavity properties of transformed densities. *Journal of Computational and Graphical Statistics*, 7(4): 514–528, 1998.

Andrew O. Finley, Sudipto Banerjee, and Bradley P. Carlin. spBayes: An R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, 19(4):1–24, 2007.

W. R. Gilks and P. Wild. Adaptive rejection sampling for Gibbs sampling. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 41(2):337–348, 1992.

W. R. Gilks, N. G. Best, and K. K. C. Tan. Adaptive rejection Metropolis sampling within Gibbs sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 44 (4):455–472, 1995.

W. Hörmann. A note on the performance of the "Ahrens algorithm". *Computing*, 69(1): 83–89, 2002.

Kyle M. Irimata, Andrew M. Raim, Ryan Janicki, James A. Livsey, and Scott H. Holan. Evaluation of Bayesian hierarchical models of differentially private data based on an approximate data model. Research Report Series: Statistics #2022-05, Center for Statistical Research and Methodology, U.S. Census Bureau, 2022. URL `https://www.census.gov/library/working-papers/2022/adrm/RRS2022-05.html`.

Roman Karawatzki. The multivariate ahrens sampling method. WorkingPaper 30, Department of Statistics and Mathematics, WU Vienna University of Economics and Business, 2006.

# References III

Luca Martino, Jesse Read, and David Luengo. Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. *IEEE Transactions on Signal Processing*, 63(12):3123–3138, 2015.

Luca Martino, David Luengo, and Joaquín Míguez. *Accept–Reject Methods*, pages 65–113. Springer, 2018.

Andrew M. Raim. Direct sampling in Bayesian regression models with additive disclosure avoidance noise. Research Report Series: Statistics #2021-01, Center for Statistical Research and Methodology, U.S. Census Bureau, 2021. URL `https://www.census.gov/library/working-papers/2021/adrm/RRS2021-01.html`.

Andrew M. Raim. Direct sampling with a step function. *Statistics and Computing*, 33(1), 2023.

Andrew M. Raim and Kimberly F. Sellers. COMPoissonReg: Usage, the normalizing constant, and other computational details. Research Report Series: Computing #2022-01, Center for Statistical Research and Methodology, U.S. Census Bureau, 2022. URL `https://www.census.gov/library/working-papers/2022/adrm/RRC2022-01.html`.

Kimberly F. Sellers. *The Conway-Maxwell-Poisson Distribution*. Cambridge University Press, 2023.

Galit Shmueli, Thomas P. Minka, Joseph B. Kadane, Sharad Borle, and Peter Boatwright. A useful distribution for fitting discrete data: revival of the Conway-Maxwell-Poisson distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54 (1):127–142, 2005.

# References IV

Gordon K. Smyth. Numerical integration. In P. Armitage and T. Colton, editors, *Encyclopedia of Biostatistics*, pages 3088–3095. Wiley, 2005.

Gary Ulrich. Computer generation of distributions on the m-sphere. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 33(2):158–163, 1984.

John von Neumann. Various techniques in connection with random digits. In A.S. Householder, G.E. Forsythe, and H.H. Germond, editors, *Monte Carlo Methods*, National Bureau of Standards Applied Mathematics Series, pages 36–38. U.S. Government Printing Office, Washington, DC, 1951.

Stephen G. Walker, Purushottam W. Laud, Daniel Zantedeschi, and Paul Damien. Direct sampling. *Journal of Computational and Graphical Statistics*, 20(3):692–713, 2011.

Andrew T. A. Wood. Simulation of the von Mises Fisher distribution. *Communications in Statistics—Simulation and Computation*, 23(1):157–164, 1994.