Direct Sampling in Bayesian Hierarchical Models for Privacy Protected Data

Andrew M. Raim

Center for Statistical Research and Methodology
U.S. Census Bureau

2021 International Conference on Advances in Interdisciplinary Statistics and Combinatorics

Disclaimer

This article is released 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 revisits the direct sampler proposed by Walker, Laud, Zantedeschi, and Damien (2011).

- Raim (2021b) proposes some customizations to improve reliability in some particular settings of interest.

- We will demonstrate in an application.
  1. Analysis of data released under differential privacy (DP).
  2. Incorporate added variability from the privacy protection mechanism in a Gibbs sampler.
  3. This work was motivated by an ongoing project to model 2020 Decennial Census data with DP (Abowd, 2018; Garfinkel et al., 2018).
  4. Modeling teammates at Census Bureau: Scott Holan, Kyle Irimata, Ryan Janicki, and James Livsey.

Weighted Densities

- To draw from a weighted density

$$f(x) = \frac{w(x)g(x)}{\psi}, \quad x \in \Omega.$$

- $\psi = \int_\Omega w(x)g(x)d\nu(x)$ is the normalizing constant.
- Ω is the support of random variable $x \sim f(x)$.
- $\nu(\cdot)$ is a dominating measure so that x may be discrete or continuous.
- f can be considered a modified version of a base distribution g. The weight function $w : \Omega \to [0, \infty)$ emphasizes or deemphasizes parts of the space.

Direct Sampling Idea

- Weighted density $f(x) = w(x)g(x)/\psi \cdot \mathrm{I}(x \in \Omega)$.
  1. Let $c = \sup_{x \in \Omega} w(x)$.
  2. Let $A_u = \{x \in \Omega : w(x) > uc\}$ for $u \in [0, 1]$.

- **Objective**: augment a random variable u so that $[x, u]$ is easier to draw than x. Especially, avoid computing ψ.

- Assume that $[u \mid x] \sim \text{Uniform}(0, w(x)/c)$, so that

$$f(u \mid x) = \frac{c}{w(x)} \mathrm{I}(0 < u < w(x)/c) = \frac{c}{w(x)} \mathrm{I}(x \in A_u).$$

- The joint density of $[x, u]$ is then

$$f(x, u) = \frac{c}{\psi} g(x) \mathrm{I}(x \in A_u).$$

- The marginal density of u is then

$$p(u) = \frac{c}{\psi} \mathrm{P}(A_u), \quad u \in [0, 1], \quad \text{where } \mathrm{P}(A_u) = \int \mathrm{I}(x \in A_u) g(x) d\nu(x).$$

- The distribution of $[x \mid u]$ is then

$$f(x \mid u) = \frac{g(x)}{\mathrm{P}(A_u)} \mathrm{I}(x \in A_u).$$

Direct Sampling Idea

- Now we can take a draw from $[x, u]$ using

$$u \sim p(u) = \frac{c}{\psi} \mathrm{P}(A_u), \quad x \sim f(x \mid u) = \frac{g(x)}{\mathrm{P}(A_u)} \mathrm{I}(x \in A_u).$$

Direct Sampler Assumptions

- f is univariate.
- w is unimodal, so that:
  a. we can identify the maximum value c.
  b. $A_u = \{x \in \Omega : w(x) > uc\}$ is an interval with endpoints $\{x_1(u), x_2(u)\}$.
- CDF, quantiles, and exact draws for g can readily be computed.
- These operations may be invoked many times, so ideally they can be computed with little work.

Drawing from $[x \mid u]$

- With a unimodal w and $A_u = (x_1(u), x_2(u))$ and

$$\begin{aligned} \mathrm{P}(A_u) &= \int_{(x_1(u), x_2(u))} g(x) d\nu(x) \\ &= G(x_2(u)-) - G(x_1(u)), \end{aligned}$$

  where G is the CDF of g.

- The density of $[x \mid u]$ is

$$f(x \mid u) = \frac{g(x)}{\mathrm{P}(A_u)} \mathrm{I}(x \in A_u) = \frac{g(x) \, \mathrm{I}(x_1(u) < x < x_2(u))}{G(x_2(u)-) - G(x_1(u))}.$$

  The associated CDF $F(x \mid u)$ and quantile function $F^{-}(\varphi \mid u)$ can also be obtained from G and G^{-}.

- An exact draw from $f(x \mid u)$ can be obtained via the inverse CDF method: draw $v \sim \text{Uniform}(0, 1)$ and take $x = F^{-}(v \mid u)$.

About $p(u)$

A few important features about the density

$$p(u) = \frac{c}{\psi} \mathrm{P}(A_u) = \frac{c}{\psi} \int \mathrm{I}(x \in A_u) g(x) d\nu(x), \quad u \in [0,1]$$

are:

- $\mathrm{P}(A_u)$ is monotonically nonincreasing in u.
- $A_0 \equiv \operatorname{supp} w$ so that $\mathrm{P}(A_0) = \int_\Omega \mathrm{I}(w(x) > 0) g(x) d\nu(x)$.
- A_1 is an empty set so that $\mathrm{P}(A_1) = 0$.

Walker et al. (2011) draw from $p(u)$ using an approximation by Bernstein polynomials on $[0,1]$.

Focusing the Support of $p(u)$

Use a bisection search to find an interval $[u_L, u_H]$ containing the “descent” of $\mathrm{P}(A_u)$.

- u_L is the smallest $u \in [0, 1]$ such that $\mathrm{P}(A_u) < \mathrm{P}(A_0)$.
- u_H is the smallest $u \in [0, 1]$ such that $\mathrm{P}(A_u) = 0$.

Step Function

- Let $u_0 < \cdots < u_N$ be knot points with $u_0 \equiv u_L$ and $u_N \equiv u_H$.
- To approximate the unnormalized $\mathrm{P}(A_u)$, consider the function

$$h^*(u) = \mathrm{P}(A_{u_0}) \cdot \mathrm{I}(0 \leq u < u_0) + \sum_{j=0}^{N-1} \mathrm{P}(A_{u_j}) \cdot \mathrm{I}(u_j \leq u < u_{j+1}).$$

- A density is obtained using $h(u) = h^*(u)/a$ with

$$a = \int_0^1 h^*(u) du = \mathrm{P}(A_{u_0}) \cdot u_0 + \sum_{j=0}^{N-1} \mathrm{P}(A_{u_j}) \cdot (u_{j+1} - u_j),$$

- Expressions for the CDF and quantile function of h can also be obtained, and the quantile function can be used to generate draws.

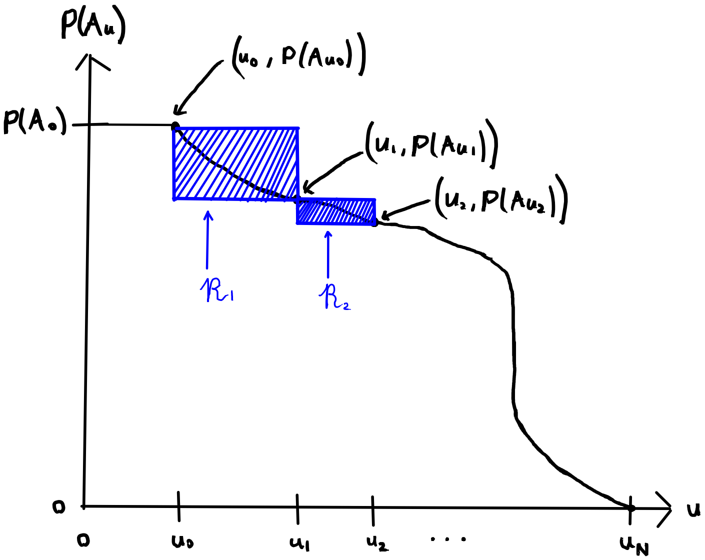
Bound for Approximation Error

- We can bound the total variation distance between the h and p distributions.
- Let $\mathcal{R}_j$ represent the rectangle in $\mathbb{R}^2$ whose upper-left point is $(u_{j-1}, \mathrm{P}(A_{u_{j-1}}))$ and lower-right point is $(u_j, \mathrm{P}(A_{u_j}))$.
- The area of $\mathcal{R}_j$ is $|\mathcal{R}_j| = \left[\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})\right](u_j - u_{j-1})$.

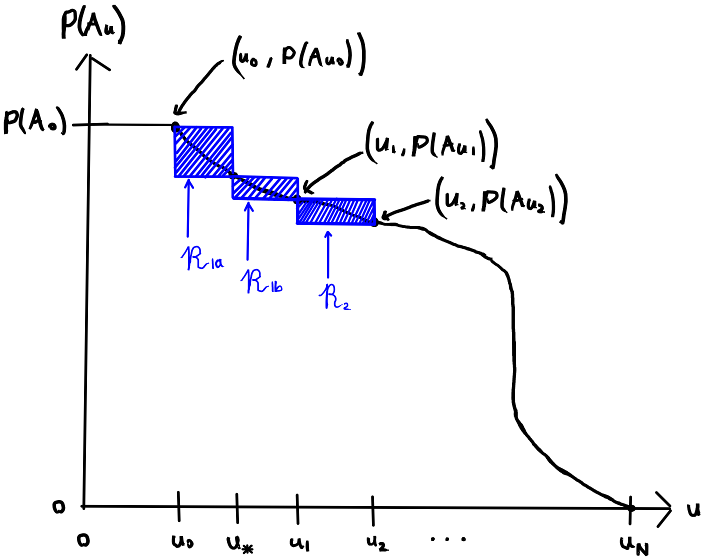
Result: Total variation distance bound.

$$d_{\mathrm{TV}}(h, p) \leq \frac{c}{\psi} \sum_{j=1}^{N} |\mathcal{R}_j|.$$

Bound for Approximation Error



Bound for Approximation Error



Knot Selection

- Our bound motivates selecting knots $u_1, \ldots, u_{N-1}$ sequentially to reduce the largest $|\mathcal{R}_j|$. This motivates the following algorithm.

Small Rectangles Algorithm.

```
Let $u^{(0)} = u_L$, and $u^{(1)} = u_H$.
for $i = 1, \ldots, N-1$ do
    Let $u_0 < \ldots < u_i$ be sorted $u^{(0)}, \ldots, u^{(i)}$.
    Let $|\mathcal{R}_j| = \{\mathrm{P}(A_{u_{j-1}}) - \mathrm{P}(A_{u_j})\}(u_j - u_{j-1})$ for $j = 1, \ldots, i$.
    Let $j^* = \operatorname*{argmax}_{j=1,\ldots,i} |\mathcal{R}_j|$.
    Let $u^{(i+1)} = \mathrm{mid}(u_{j^*-1}, u_{j^*})$.
Let $u_0 < \ldots < u_N$ be sorted $u^{(0)}, \ldots, u^{(N)}$.
return $(u_0, \ldots, u_N)$.
```

- The cost of this over equally-spaced knots $u_j = u_L + (j/N)(u_H - u_L)$ is increased computation.

- To avoid repeated sorting of the $|\mathcal{R}_j|$'s, we can use a priority queue.

Accept-Reject Algorithm

- The step function h^* was constructed so that $h^*(u) \geq \mathrm{P}(A_u)$ for all $u \in [0, 1]$.
- Therefore h^* may be used as an envelope for rejection sampling to take exact draws from $p(u)$.
- Taking $v \sim \text{Uniform}(0, 1)$, the candidate $u \sim h(u)$ is accepted as a draw from $p(u)$ if $v < \frac{\mathrm{P}(A_u)}{h^*(u)}$. Otherwise, repeat.

Result: An upper bound for probability of a rejection.

$$\mathrm{P}\left(v \geq \frac{\mathrm{P}(A_u)}{h^*(u)} \right) \leq \frac{c}{\psi} \sum_{j=1}^{N} |\mathcal{R}_j|.$$

- A rejected u may be added to the knot points to improve the envelope, at the cost of more bookkeeping.

Sequential Sampling

- Suppose $\Omega \subseteq \mathbb{R}^k$ and we want to draw $\boldsymbol{x}$ from

$$f(\boldsymbol{x}) = \frac{w(\boldsymbol{x})g(\boldsymbol{x})}{\psi}, \quad \boldsymbol{x} \in \Omega, \quad \psi = \int_\Omega w(\boldsymbol{x})g(\boldsymbol{x})d\nu(\boldsymbol{x})$$

- Suppose we can factorize

$$\begin{aligned} g(\boldsymbol{x}) &= g_1(x_1)g_2(x_2 \mid x_1)\cdots g_k(x_k \mid x_1, \ldots x_{k-1}), \\ w(\boldsymbol{x}) &= w_1(x_1)w_2(x_2 \mid x_1)\cdots w_k(x_k \mid x_1, \ldots x_{k-1}), \end{aligned}$$

  so that each $g_j(x_j \mid x_1, \ldots x_{j-1})$ is a density, and $w_j(x_j \mid x_1, \ldots x_{j-1})$ is a corresponding weight function.

- We may sample from

$$f_j(x_j \mid x_1, \ldots x_{j-1}) = w_j(x_j \mid x_1, \ldots x_{j-1})g_j(x_j \mid x_1, \ldots x_{j-1})$$

  sequentially, for $j = 1, \ldots, k$, via direct sampler (or another method) to obtain a draw $\boldsymbol{x}$.

An Application

- Differential privacy (DP) has become increasingly popular for its ability to mathematically bound risks of unwanted disclosure in the released data (Dwork and Roth, 2014).
- A basic setting for DP involves an agency wishing to release statistics based on sensitive data.
- Bernstein and Sheldon (2018) provide a simple-but-nontrivial setting where the direct sampler may be used within a Gibbs sampler.
- Under Sufficient Statistic Perturbation (SSP), privacy protection is applied to the sufficient statistics under an assumed model (e.g. Foulds et al., 2016; Bernstein and Sheldon, 2018, 2019).
- The privacy protection mechanism is known to users. Using it in a data analysis may complicate the analysis but ignoring it can distort results (Gong, 2020).

Non-sensitive Data

- Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $x_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2)$.
- A Gibbs sampler can be formulated with the sufficient statistics $t_{\text{sse}}(\boldsymbol{x}) = \sum_{i=1}^n (x_i - \bar{x})^2$ and $t_{\text{mean}}(\boldsymbol{x}) = \bar{x}$ in lieu of $\boldsymbol{x}$.
- Assuming prior $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$, we have

$$\begin{aligned} f(\mu \mid \sigma^2, \boldsymbol{x}) &\propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \\ &= \exp\left\{ -\frac{1}{2\sigma^2} [t_{\text{sse}} + n(t_{\text{mean}} - \mu)^2] \right\} \exp\left\{ -\frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right\} \\ &\propto \text{N}(\vartheta, 1/\omega^2), \end{aligned}$$

  with $\omega^2 = n\sigma^{-2} + \sigma_0^{-2}$ and $\vartheta = \omega^{-2}(\sigma^{-2} t_{\text{mean}} + \sigma_0^{-2}\mu_0)$.

- Also,

$$\begin{aligned} f(\sigma^2 \mid \mu, \boldsymbol{x}) &\propto (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} (\sigma^2)^{-a_\sigma - 1} e^{-b_\sigma/\sigma^2} \\ &\propto \text{IG}\left(a_\sigma + \frac{n}{2}, b_\sigma + \frac{1}{2}[t_{\text{sse}} + n(t_{\text{mean}} - \mu)^2] \right). \end{aligned}$$

Release for Sensitive Data

- If the data $\boldsymbol{x}$ are sensitive, releasing $\boldsymbol{t}(\boldsymbol{x})$ without protection represents a leak of privacy.
- Definitions.
  1. Privacy loss budget $\epsilon > 0$ which represents a tolerance for loss of privacy.
  2. Two datasets $\boldsymbol{x}, \boldsymbol{x}'$ are neighbors if $x_i \neq x_i'$ for exactly one i (following Bernstein and Sheldon, 2018).
  3. The sensitivity of statistic $\boldsymbol{t}(\boldsymbol{x})$ is $\Delta_{\boldsymbol{t}} = \max \|\boldsymbol{t}(\boldsymbol{x}) - \boldsymbol{t}(\boldsymbol{x}')\|_1$ s.t. $\boldsymbol{x}, \boldsymbol{x}'$ are neighbors.
- A Laplace mechanism is

$$M_{\text{Lap}}(\boldsymbol{x} \mid \boldsymbol{t}, \epsilon) = \boldsymbol{t}(\boldsymbol{x}) + \boldsymbol{\xi}, \quad \xi_1, \ldots, \xi_k \overset{\text{iid}}{\sim} \text{Lap}(0, \Delta_{\boldsymbol{t}}/\epsilon).$$

- The mechanism M_{Lap} above satisfies ϵ-differential privacy:

$$\text{P}[M(\boldsymbol{x}) \in B] \leq e^{\epsilon}\, \text{P}[M(\boldsymbol{x}') \in B]$$

for all $B \subseteq \text{range}(M)$ and any neighboring datasets $\boldsymbol{x}, \boldsymbol{x}'$.

Release for Sensitive Data

- Sensitivity $\Delta_{\boldsymbol{t}}$ is infinite if data are unbounded, so we take a censoring approach similar to Bernstein and Sheldon (2018).
- Agency chooses an $[a, b]$ so that the extreme x_i are excluded from mean and SSE statistics.
- Consider a censored likelihood:

$$f(\boldsymbol{x} \mid \boldsymbol{\theta}) = \Big[\Phi(a \mid \boldsymbol{\theta})\Big]^{s_\ell} \Big[1 - \Phi(b \mid \boldsymbol{\theta})\Big]^{s_u} \times \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{s_c} \exp\left\{-\frac{1}{2\sigma^2}[s_{\text{sse}} + s_c(s_{\text{mean}} - \mu)^2]\right\}.$$

- Now there are four sufficient statistics:

$$s_{\text{mean}}(\boldsymbol{x}) = \frac{1}{s_c(\boldsymbol{x})} \sum_{i \in \mathcal{I}_c(\boldsymbol{x})} x_i, \quad s_{\text{sse}}(\boldsymbol{x}) = \sum_{i \in \mathcal{I}_c(\boldsymbol{x})} (x_i - s_{\text{mean}}(\boldsymbol{x}))^2,$$

$$s_\ell(\boldsymbol{x}) = \sum_{i=1}^n \mathrm{I}(x_i < a), \quad s_u(\boldsymbol{x}) = \sum_{i=1}^n \mathrm{I}(x_i > b),$$

with $\mathcal{I}_c(\boldsymbol{x}) = \{i : x_i \in [a, b]\}$ and $s_c(\boldsymbol{x}) = n - s_\ell(\boldsymbol{x}) - s_u(\boldsymbol{x})$.

Release for Sensitive Data

- Based on

$$\boldsymbol{s}(\boldsymbol{x}) = \Big(s_{\text{mean}}(\boldsymbol{x}), s_{\text{sse}}(\boldsymbol{x}), s_{\ell}(\boldsymbol{x}), s_{u}(\boldsymbol{x})\Big),$$

  we can obtain an upper bound $\Delta_{\boldsymbol{s}}$ for the sensitivity.

Release for Sensitive Data

To summarize:

- Agency collects sensitive data $\boldsymbol{x} = (x_1, \ldots, x_n)$.
- Agency selects a, b, and ϵ. This determines $\Delta_{\boldsymbol{s}}$ (or an upper bound).
- Agency releases $\boldsymbol{z}$ whose elements are

$$\begin{aligned} z_{\text{mean}} &= s_{\text{mean}}(\boldsymbol{x}) + \xi_{\text{mean}}, \\ z_{\text{sse}} &= s_{\text{sse}}(\boldsymbol{x}) + \xi_{\text{sse}}, \\ z_{\ell} &= s_{\ell}(\boldsymbol{x}) + \xi_{\ell}, \\ z_{u} &= s_{u}(\boldsymbol{x}) + \xi_{u}, \end{aligned}$$

  where $\xi_{\text{mean}}, \xi_{\text{sse}}, \xi_{\ell}, \xi_{u} \overset{\text{iid}}{\sim} \text{Lap}(0, \Delta_{\boldsymbol{s}}/\epsilon)$.

- Analyst now wants inference on parameter $\boldsymbol{\theta} = (\mu, \sigma^2)$ based on released data $\boldsymbol{z}$, along with knowledge of n, a, b, and ϵ.

Analysis of Sensitive Data

- To formulate a Gibbs sampler, consider the augmented data model

$$\begin{aligned} &\boldsymbol{z} = \boldsymbol{s}(\boldsymbol{x}) + \boldsymbol{\xi}, \quad \xi_j \overset{\text{iid}}{\sim} \text{Lap}(0, \Delta_{\boldsymbol{s}}/\epsilon), \\ &x_i \overset{\text{iid}}{\sim} \text{N}(\mu, \sigma^2), \end{aligned}$$

with prior $\mu \sim \text{N}(\mu_0, \sigma_0^2)$ and $\sigma^2 \sim \text{IG}(a_\sigma, b_\sigma)$.

Analysis of Sensitive Data

- Our strategy will be to take repeated draws from the following conditionals:
  1. $[\boldsymbol{s}, \boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{z}] = [\boldsymbol{x} \mid \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{z}][\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{z}]$.
  2. $[\mu \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}, \sigma^2]$
  3. $[\sigma^2 \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}, \mu]$
- $[\mu \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}, \sigma^2]$ and $[\sigma^2 \mid \boldsymbol{x}, \boldsymbol{s}, \boldsymbol{z}, \mu]$ are Normal and Inverse Gamma draws, respectively, similar to the non-sensitive data setting.
- $[\boldsymbol{x} \mid \boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{z}]$ involves drawing the censored elements of $\boldsymbol{x}$ from truncated $\mathrm{N}(\mu, \sigma^2)$ distributions: s_ℓ are truncated to $(-\infty, a)$ and s_u are truncated to (b, ∞).
- $[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{z}]$ is where direct sampling can be used.

Analysis of Sensitive Data

- To sample from $[\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{z}] \propto [\boldsymbol{z} \mid \boldsymbol{s}][\boldsymbol{s} \mid \boldsymbol{\theta}]$, we need to determine $[\boldsymbol{s} \mid \boldsymbol{\theta}]$.

- We have that

$$[s_\ell, s_u, s_c \mid \boldsymbol{\theta}] \sim \text{Mult}_3(n, q_\ell(\boldsymbol{\theta}), q_u(\boldsymbol{\theta}), q_c(\boldsymbol{\theta}))$$

  denoting $s_c = n - s_\ell - s_u$, $q_\ell(\boldsymbol{\theta}) = \text{P}(x_1 < a)$, $q_u(\boldsymbol{\theta}) = \text{P}(x_1 > b)$, and $q_c(\boldsymbol{\theta}) = \text{P}(x_1 \in [a, b])$.

- For the remaining elements of $\boldsymbol{s}$, we have

$$[s_\text{mean} \mid s_\ell, s_u, \boldsymbol{\theta}] \sim \text{N}(\mu, \sigma^2 / s_c), \quad [s_\text{sse} \mid s_\ell, s_u, \boldsymbol{\theta}] \sim \sigma^2 \chi^2_{s_c - 1}.$$

Analysis of Sensitive Data

- Therefore, we may write

$$
\begin{aligned}
f(\boldsymbol{s} \mid \boldsymbol{\theta}, \boldsymbol{z}) = {} & f_{\text{Lap}}(z_{\text{sse}} \mid s_{\text{sse}}, \Delta_{\boldsymbol{s}}/\epsilon) f(s_{\text{sse}} \mid s_\ell, s_u, \boldsymbol{\theta}) \\
& \times f_{\text{Lap}}(z_{\text{mean}} \mid s_{\text{mean}}, \Delta_{\boldsymbol{s}}/\epsilon) f(s_{\text{mean}} \mid s_\ell, s_u, \boldsymbol{\theta}) \\
& \times f_{\text{Lap}}(z_\ell \mid s_\ell, \Delta_{\boldsymbol{s}}/\epsilon) f(s_\ell \mid s_u, \boldsymbol{\theta}) \\
& \times f_{\text{Lap}}(z_u \mid s_u, \Delta_{\boldsymbol{s}}/\epsilon) f(s_u \mid \boldsymbol{\theta}),
\end{aligned}
$$

- We may now draw $\boldsymbol{s}$ sequentially in four steps.

Sampling the Sufficient Statistics I

- To sample from our target conditional:
  1. Sample s_u from
  $$f(s_u \mid \boldsymbol{\theta}, \boldsymbol{z}) \propto f_{\mathrm{Lap}}(z_u - s_u \mid 0, \Delta_{\boldsymbol{s}}/\epsilon) f_{\mathrm{Bin}}(s_u \mid n, q_u(\boldsymbol{\theta})).$$
  2. Sample s_ℓ from
  $$f(s_\ell \mid s_u, \boldsymbol{\theta}, \boldsymbol{z}) \propto f_{\mathrm{Lap}}(z_\ell - s_\ell \mid 0, \Delta_{\boldsymbol{s}}/\epsilon) f_{\mathrm{Bin}}\left(s_\ell \mid n - s_u, \frac{q_\ell(\boldsymbol{\theta})}{1 - q_u(\boldsymbol{\theta})}\right).$$
- Support is finite in steps 1 and 2, so we may sample with standard method: compute unnormalized probabilities, normalize, and draw via discrete distribution.

Sampling the Sufficient Statistics II

- To sample from our target conditional:
  3. Sample ξ_{mean} from weighted density

$$f(\xi_{\text{mean}} \mid s_\ell, s_u, \boldsymbol{\theta}, \boldsymbol{z}) \propto \underbrace{f_{\text{Lap}}(\xi_{\text{mean}} \mid 0, \Delta_{\boldsymbol{s}}/\epsilon)}_{g(\xi_{\text{mean}})} \underbrace{f_{\text{N}}(z_{\text{mean}} - \xi_{\text{mean}} \mid \mu, \sigma^2/s_c)}_{w(\xi_{\text{mean}})}$$

  with $s_c = n - s_\ell - s_u$ and let $s_{\text{mean}} = z_{\text{mean}} - \xi_{\text{mean}}$.

  4. Sample ξ_{sse} from

$$\begin{aligned} &f(\xi_{\text{sse}} \mid s_{\text{mean}}, s_\ell, s_u, \boldsymbol{\theta}, \boldsymbol{z}) \\ &\quad \propto f_{\text{Lap}}(\xi_{\text{sse}} \mid 0, \Delta_{\boldsymbol{s}}/\epsilon)(z_{\text{sse}} - \xi_{\text{sse}})^{\frac{s_c - 1}{2} - 1} e^{-(z_{\text{sse}} - \xi_{\text{sse}})/2\sigma^2} \\ &\quad \propto \underbrace{f_{\text{Lap}}(\xi_{\text{sse}} \mid 0, \Delta_{\boldsymbol{s}}/\epsilon)}_{g(\xi_{\text{sse}})} \underbrace{f_{\text{Gamma}}(z_{\text{sse}} - \xi_{\text{sse}} \mid (s_c - 1)/2, 2\sigma^2)}_{w(\xi_{\text{sse}})} \end{aligned}$$

  and let $s_{\text{sse}} = z_{\text{sse}} - \xi_{\text{sse}}$.

- Direct sampling is useful in steps 3 and 4.

Laplace Base Distribution

- CDF and quantile functions of $\text{Lap}(0, \lambda)$ are respectively

$$G(\xi \mid \lambda) = \frac{1}{2} + \frac{1}{2}\,\text{sgn}(\xi)[1 - e^{-|\xi|/\lambda}], \quad \text{and}$$
$$G^{-}(\varphi \mid \lambda) = -\lambda\,\text{sgn}\left(\varphi - \frac{1}{2}\right)\log\left(1 - 2\left|\varphi - \frac{1}{2}\right|\right).$$

- Draws from $\text{Lap}(0, \lambda)$ may be taken using inverse CDF method.

Normal Weight Function

- Normal weight function is proportional to a Normal density,

$$w(\xi) = \exp\left\{ -\frac{1}{2\tau^2} [z_{\text{mean}} - \xi - \vartheta]^2 \right\},$$

with $\vartheta = \mu$ and $\tau^2 = \sigma^2 / s_c$.

- The maximum value of $w(\xi)$ is $c = 1$, attained at $\xi^* = z_{\text{mean}} - \mu$.
- The set $A_u = \{\xi \in \Omega : w(\xi) > uc\}$ is an interval with endpoints

$$\{\xi_1(u), \xi_2(u)\} = z_{\text{mean}} - \vartheta \pm \sqrt{-2\tau^2 \log(cu)}.$$

Gamma Weight Function

- Gamma weight function is

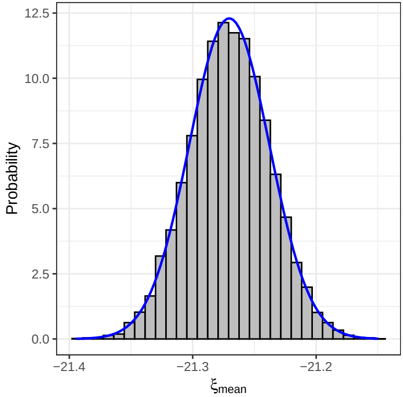
$$w(\xi) = \frac{(z_{\text{sse}} - \xi)^{\alpha-1} e^{-(z_{\text{sse}}-\xi)/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \, \mathrm{I}(\xi < z_{\text{sse}}),$$

  with $\alpha = (s_c - 1)/2$ and $\beta = 2\sigma^2$.

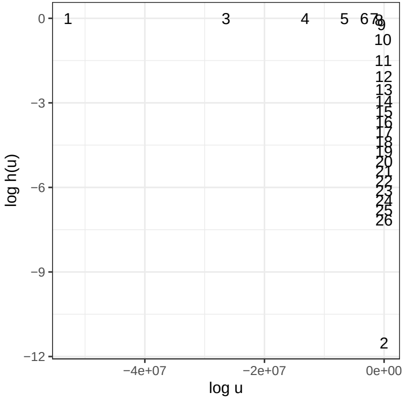
- The maximum value of $w(\xi)$, attained at $\xi^* = z_{\text{sse}} - \beta(\alpha - 1)$, is

$$w(\xi^*) = \frac{(\alpha-1)^{\alpha-1} e^{-(\alpha-1)}}{\Gamma(\alpha)\beta}.$$

- The set $A_u = \{\xi \in \Omega : w(\xi) > uc\}$ is an interval. Endpoints are computed using numerical root finding.



(a) Draws from the target density (blue curve).

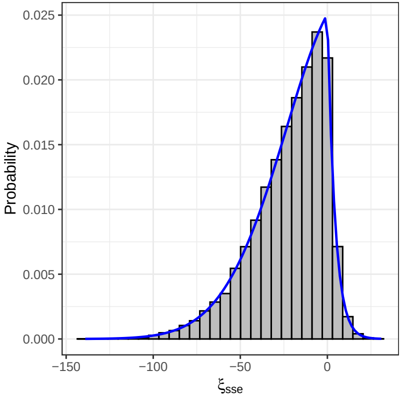


(b) Step function $h(u)$ on log-log scale with $N = 25$. Numbers indicate order in which knots were added.

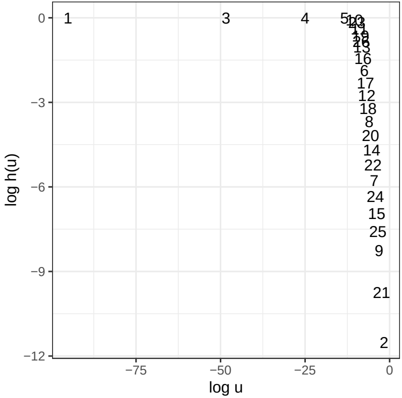
Sample based on 20,000 draws from target density

$$f(\xi_{\text{mean}} \mid s_\ell = 23, s_u = 27, \mu = 0, \sigma^2 = 1, z_{\text{mean}} = -21.27)$$

using direct sampler with accept-reject. Here, $n = 1000$, $a = -2$, $b = 2$, $\epsilon = 2$.



(a) Draws from the target density (blue curve).



(b) Step function $h(u)$ on log-log scale with $N = 25$. Numbers indicate order in which knots were added.

Sample is based on 20,000 draws from target density

$$f(\xi_{\text{sse}} \mid s_\ell = 23, s_u = 27, \mu = 0, \sigma^2 = 1, z_{\text{sse}} = 778.56)$$

using direct sampler with accept-reject. Here, $n = 1000$, $a = -2$, $b = 2$, $\epsilon = 2$.

Conclusions

- We discussed some extensions to the direct sampler from Raim (2021b).
- We reviewed an application to Gibbs sampling in SSP framework.
    1. To draw s_{mean}: Laplace base distribution and Normal weight function.
    2. To draw s_{sse}: Laplace base distribution and Gamma weight function.
- R and Rcpp code is on Github (Raim, 2021a).
    1. Object-oriented and functional programming design.
    2. Care is required to operate on floating point numbers with extreme magnitudes.
- Other additive privacy mechanisms yield similar direct sampling steps.
    1. Gaussian mechanism (Dwork and Roth, 2014, Appendix A).
    2. Double Geometric mechanism (Ghosh et al., 2012).
    3. Discrete Gaussian mechanism (Canonne et al., 2020).
- Larger scale computational studies of Gibbs sampler are in progress.

Thank You!

Andrew M. Raim
andrew.raim@census.gov

Andrew Raim. *Direct Sampling*, 2021a. R package version 0.1.0. https://github.com/andrewraim/DirectSampling.

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, 2021b. https://www.census.gov/library/working-papers/2021/adrm/RRS2021-01.html.

References I

John M. Abowd. The U.S. Census Bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2867, New York, NY, USA, 2018. Association for Computing Machinery.

Garrett Bernstein and Daniel Sheldon. Differentially private bayesian inference for exponential families. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 2924–2934, Red Hook, NY, USA, 2018. Curran Associates Inc.

Garrett Bernstein and Daniel R Sheldon. Differentially private Bayesian linear regression. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 525–535. Curran Associates, Inc., 2019.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. The discrete Gaussian for differential privacy, 2020. URL `https://arxiv.org/abs/2004.00010`.

Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc, 2014.

References II

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, pages 192–201, Arlington, Virginia, USA, 2016. AUAI Press.

Simson L. Garfinkel, John M. Abowd, and Sarah Powazek. Issues encountered deploying differential privacy. In *Proceedings of the 2018 Workshop on Privacy in the Electronic Society*, WPES'18, pages 133–137, New York, NY, USA, 2018. Association for Computing Machinery.

Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. *SIAM Journal on Computing*, 41(6): 1673–1693, 2012.

Ruobin Gong. Transparent privacy is principled privacy, 2020. URL `https://arxiv.org/abs/2006.08522`.

Andrew Raim. *Direct Sampling*, 2021a. URL `https://github.com/andrewraim/DirectSampling`. R package version 0.1.0.

References III

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, 2021b. URL `https://www.census.gov/library/working-papers/2021/adrm/RRS2021-01.html`.

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