Amortized Rejection Sampling in Universal Probabilistic Programming

THE UNIVERSITY OF BRITISH COLUMBIA

Saeid Naderiparizi, Adam Ścibior, Andreas Munk, Mehrdad Ghadiri, Atılım Güneș Baydin,
Bradley Gram-Hansen, Christian Schroeder de Witt, Robert Zinkov, Philip Torr, Tom Rainforth, Yee Whye Teh, Frank Wood

Georgia Tech

UNIVERSITY OF OXFORD

Introduction

- Rejection sampling is widely used in implementing complex generative models.
- Inference in probabilistic programs that contain unbounded loops (e.g. rejection sampling) is hard.
- We address the problem of efficient amortized importance-sampling-based inference, in particular Inference Compilation (IC) [1], in such models.
- We prove naive application of IC can produce importance weights with unbounded variance.
- We propose Amortized Rejection Sampling (ARS), an importance sampling procedure that produces unbiased expectations for programs that contain rejection sampling loops.
- We prove ARS does not introduce infinite variance in its handling of rejection sampling loops.
- We implement ARS in PyProb [2, 3] in a way that requires minimal modifications to user code.

(a) Original program

$x \sim p(x)$
for $k \in \mathbb{N}^+$ **do**
  $z^k \sim p(z|x)$
  if $c(x, z^k)$ **then**
    $z = z^k$
    break
observe$(y, p(y|z,x))$

(b) Inference compilation

$x \sim q(x|y)$
$w \leftarrow \frac{p(x)}{q(x|y)}$
for $k \in \mathbb{N}^+$ **do**
  $z^k \sim q(z|x,y)$
  $w^k \leftarrow \frac{p(z^k|x)}{q(z^k|x,y)}$
  $w \leftarrow w\, w^k$
  if $c(x, z^k)$ **then**
    $z = z^k$
    break
$w_{\text{IC}} \leftarrow wp(y|z,x)$

(c) Equivalent to above

$x \sim p(x)$
$z \sim p(z|x, c(x,z))$
observe$(y, p(y|z,x))$

(d) ARS

$x \sim q(x|y)$
$w \leftarrow \frac{p(x)}{q(x|y)}$
$z \sim q(z|x,y,c(x,z))$
$w \leftarrow w\,\frac{p(z|x,c(x,z))}{q(z|x,y,c(x,z))}$
$w_{\text{C}} \leftarrow w\, p(y|z,x)$

IC weights

$$w_{\text{IC}} = \frac{p(x)}{q(x|y)} p(y|x,z) \prod_{k=1}^{L} \frac{p(z^k|x)}{q(z^k|x,y)}$$

Theorem: Under some mild conditions if the following holds then the variance of w_{IC} is infinite.

$$\mathbb{E}_{z\sim q(z|x,y)}\left[\frac{p(z|x)^2}{q(z|x,y)^2}(1-p(A|x,z))\right] \geq 1$$

where A is the event of $c(x,z)$ being satisfied.

Collapsed weights

$$w_{\text{C}} = \frac{p(x)}{q(x|y)} \frac{p(z|x,A)}{q(z|x,y,A)} p(y|x,z)$$

- $\mathbb{E}[w_{\text{IC}}] = \mathbb{E}[w_{\text{C}}]$ but rejection sampling loops do not cause infinite variance in w_{C}.
- Unfortunately, we cannot directly compute w_{C} due to the intractable term $\frac{p(z|x,A)}{q(z|x,y,A)}$.

Amortized Rejection Sampling (ARS)

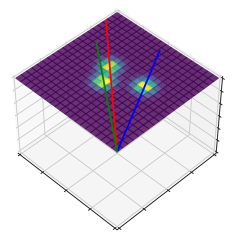
$$w_{\text{C}} = \frac{p(x)}{q(x|y)} \frac{p(z|x)}{q(z|x,y)} p(y|x,z) \frac{q(A|x,y)}{p(A|x)}$$

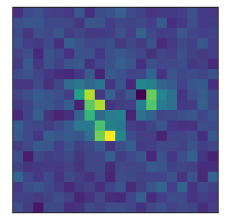
- $q(A|x,y)$ is the probability of exiting the rejection sampling loop under the proposal.
- $p(A|x)$ is the probability of exiting the rejection sampling loop in the original probabilistic program.
- We use Monte Carlo sampling to get unbiased estimates of $q(A|x,y)$ and $\frac{1}{p(A|x)}$.

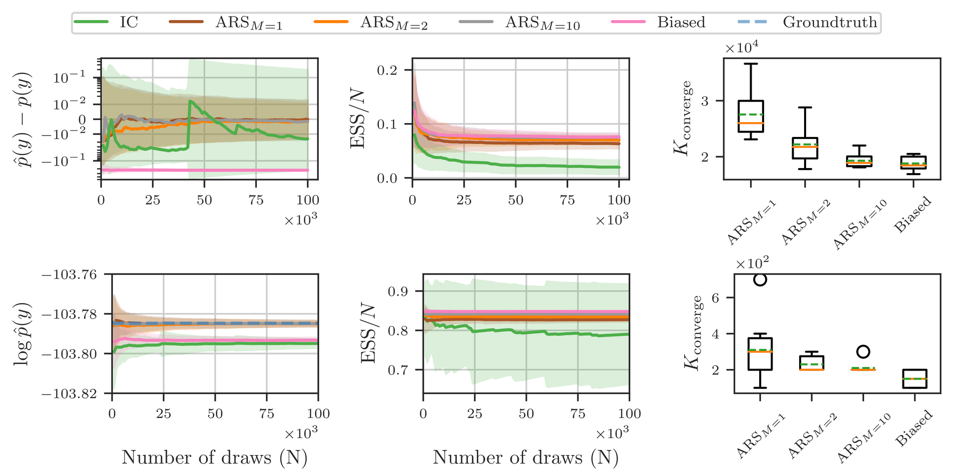
Experiments

Marsaglia and Mini-Sherpa

for $k \in \mathbb{N}^+$ **do**
  $a_k \sim \text{Uniform}(-1,1)$
  $b_k \sim \text{Uniform}(-1,1)$
  $s = a_k^2 + b_k^2$
  if $s < 1$ **then**
    $a = a_k$
    $b = b_k$
    break
$\mu = a\sqrt{\frac{-2\log(s)}{s}}$
observe$(y_1, \mathcal{N}(\mu, \sigma^2))$
observe$(y_2, \mathcal{N}(\mu, \sigma^2))$





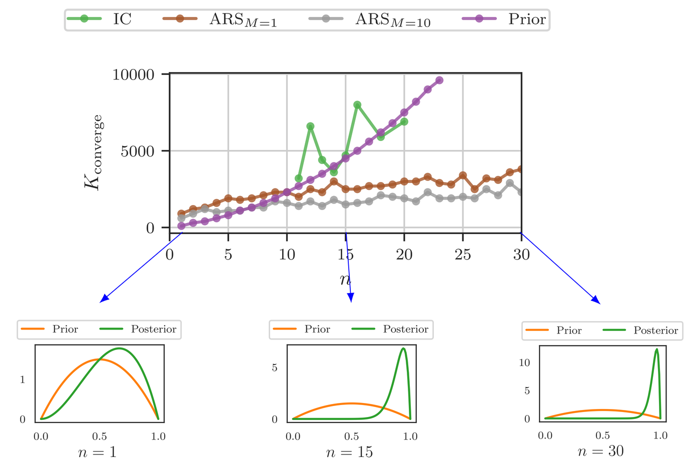


(Top) Marsaglia. (Bottom) Mini-SHERPA

Beta-Bernoulli

- We compare ARS with "Prior," a baseline proposed in [2].
- This baseline uses the prior as proposal for the variables within rejection sampling loops, ignoring the learned proposal.

for $k \in \mathbb{N}^+$ **do**
  $x_k \sim \text{Uniform}(0,1)$
  $u_k \sim \text{Uniform}(0,1)$
  if $c(x_k, u_k)$ **then**
    $x = x_k$
    $u = u_k$
    break
for $i \in [1,n]$ **do**
  observe(True, Ber(x))



- Prior converges quicker than ARS if the posterior and prior are close.
- As the difference between prior and posterior grows, ARS quickly outperforms Prior.

Algorithm

$x \sim q(x|y)$
$w \leftarrow \frac{p(x)}{q(x|y)}$
for $k \in \mathbb{N}^+$ **do**
  $z^k \sim q(z|x,y)$
  if $c(x, z^k)$ **then**
    $z = z^k$
    break
$w \leftarrow w\,\frac{p(z|x)}{q(z|x,y)}$
$K \leftarrow 0$
for $i \in 1, \ldots N$ **do**
  $z'_i \leftarrow q(z|x,y)$
  $K \leftarrow K + c(z,x)$
for $j \in 1, \ldots M$ **do**
  for $l \in \mathbb{N}^+$ **do**
    $z''_{j,l} \leftarrow q(z|x,y)$
    if $c(x, z''_{j,l})$ **then**
      $T_j \leftarrow l$
      break
$T \leftarrow \frac{1}{M}\sum_{j=1}^{M} T_j$
$w \leftarrow w\,\frac{KT}{N}$
$w \leftarrow w\,p(y|z,x)$

- $\frac{K}{N}$ estimates $q(A|x,y)$
- T estimates $\frac{1}{p(A|x)}$

Implementation

We introduce two new functions to tag the beginning and end of rejection sampling loops.

Original

```
x = sample(P_x)
while True:


    z = sample(P_z(x))
    if c(x, z):

        break
observe(P_y(x,z), y)
return x, z
```

Annotated

```
x = sample(P_x)
while True:
    rs_start()
    z = sample(P_z(x))
    if c(x, z):
        rs_end()
        break
observe(P_y(x,z), y)
return x, z
```

References

[1] Tuan Anh Le, Atılım Güneş Baydin, and Frank Wood. Inference compilation and universal probabilistic programming. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1338–1348, Fort Lauderdale, FL, USA, 2017. PMLR.

[2] Atilim Gunes Baydin, Lukas Heinrich, Wahid Bhimji, Bradley Gram-Hansen, Gilles Louppe, Lei Shao, Kyle Cranmer, Frank Wood, et al. Efficient probabilistic inference in the quest for physics beyond the standard model. In *Thirty-second Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Atılım Güneş Baydin, Lei Shao, Wahid Bhimji, Lukas Heinrich, Lawrence Meadows, Jialin Liu, Andreas Munk, Saeid Naderiparizi, Bradley Gram-Hansen, Gilles Louppe, et al. Etalumis: Bringing probabilistic programming to scientific simulators at scale. In *the International Conference for High Performance Computing, Networking, Storage and Analysis (SC '19)*, 2019. doi: 10.1145/3295500.3356180.