Adaptive Pareto Smoothed Importance Sampling

William Ruth

Joint work with Payman Nickchi

Introduction

- Importance sampling
- Measuring performance
- Improving performance
  - Modifications
  - Optimization

Importance Sampling

- Need to compute an expected value
  - $\mathbb{E}_F \varphi(X)$
- Can't do the sum/integral
- Monte Carlo approximation
  - Simulating from F might be hard

Importance Sampling

- Introduce "proposal distribution", G:

$$\begin{aligned}\mathbb{E}_F \varphi(X) &= \mathbb{E}_G \left[\varphi(X) \cdot \frac{f(X)}{g(X)} \right] \\ &= \mathbb{E}_G \left[\varphi(X) \cdot w(X) \right]\end{aligned}$$

- G can be nearly anything*
  - *Some choices will be better than others

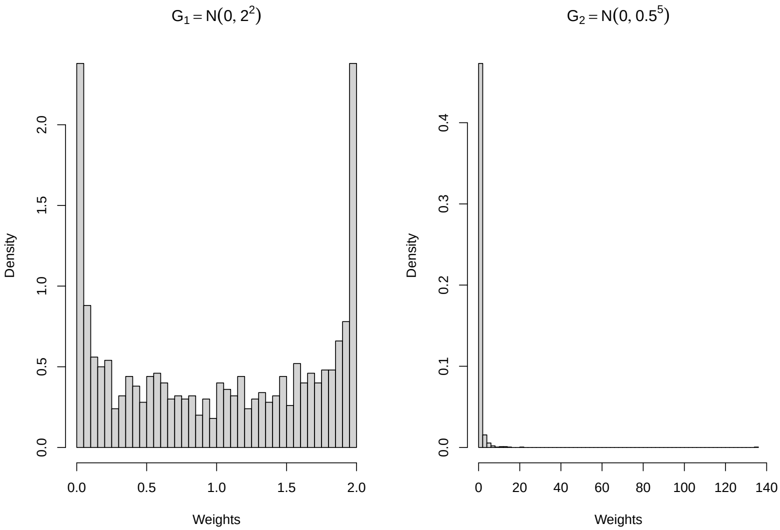
Example: Mystery Target

- f unknown, but can be evaluated
- $\varphi(X) = X^2$

- Try some proposals:
  - $G_1 \sim N(0, 2^2)$
  - $G_2 \sim N(0, 0.6^2)$

- Use $M = 1000$ samples from proposal
  - $\hat{\mathbb{E}}_1 = 0.99$, $\hat{\text{SD}} = 1.97$
  - $\hat{\mathbb{E}}_2 = 1.10$, $\hat{\text{SD}} = 2.32$

Example: Mystery Target



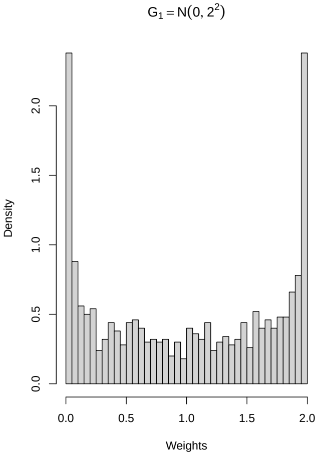
Importance Sampling

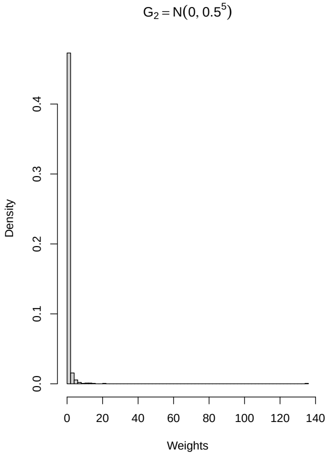
- G_1 weights look fine
- G_2 weights dominated by one large value

- We can make this difference precise
- "Effective Sample Size":

$$ESS = \frac{\left[\sum_i w(X_i)\right]^2}{\sum_i w(X_i)^2}$$

Example: Mystery Target





$ESS_1 \approx 662$

$ESS_2 \approx 54$

- Problem: Low ESS $\rightarrow$ hard to estimate means
- But ESS is based on means
  - (Chatterjee and Diaconis, 2018)

Improving IS

- Choose a good proposal
- Modify large weights
  - Truncated IS
  - Pareto Smoothed IS

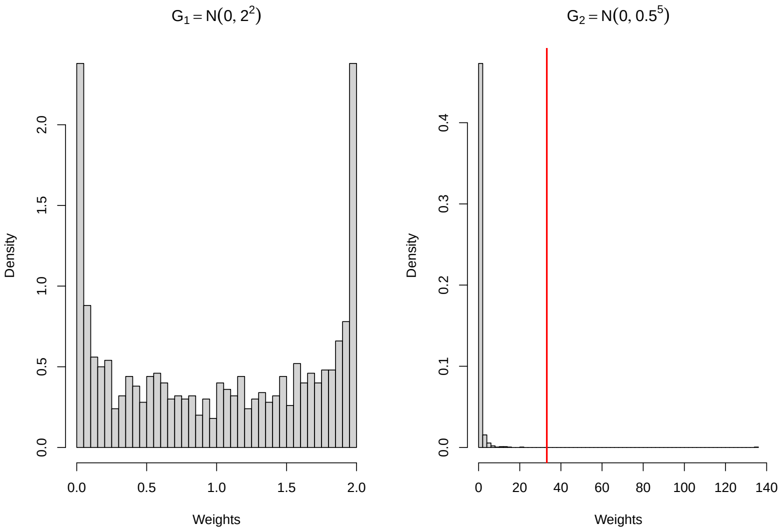
Improving IS

- Truncated Importance Sampling:
    - (Ionides, 2008)

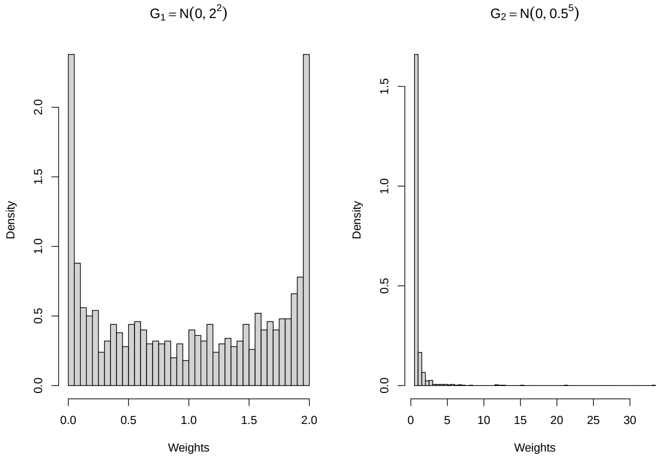
1. Choose a threshold
2. Apply hard thresholding to any large weights

- Still consistent for the target

Example: Mystery Target



Example: Mystery Target



$$ESS_1 \approx 662$$
$$ESS_1^{(\mathrm{trunc})} \approx 662$$

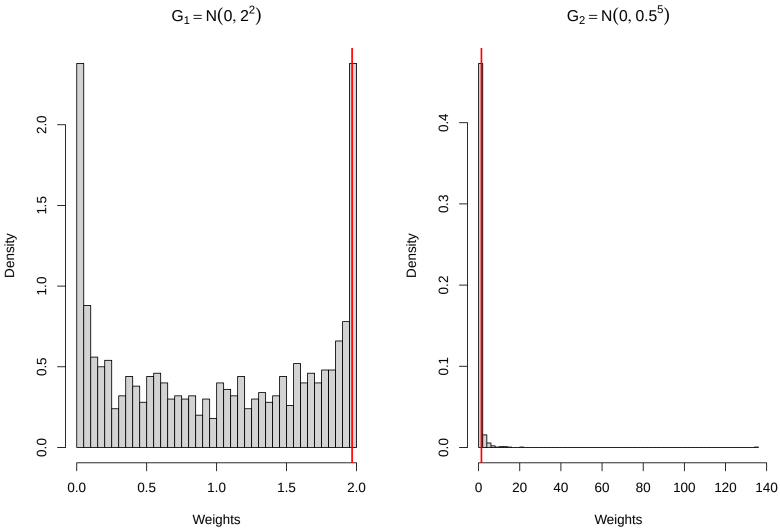
$$ESS_2 \approx 54$$
$$ESS_2^{(\mathrm{trunc})} \approx 245$$

Improving IS

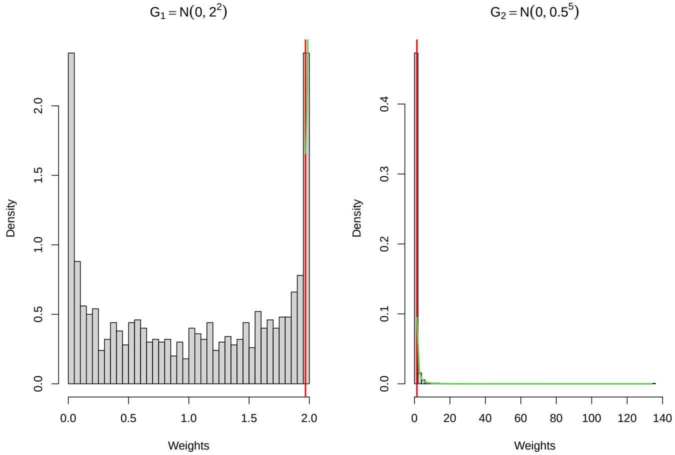
- Pareto Smoothed Importance Sampling:
  - (Vehtari et al., 2024)

1. Choose a threshold
   - Weights above threshold represent tail of their dist.
2. Approximate tail with Generalized Pareto Dist.
   - Fit GPD to weights above threshold
   - (Zhang and Stephens, 2009)
3. Replace large weights with quantiles of fitted GPD

Example: Mystery Target

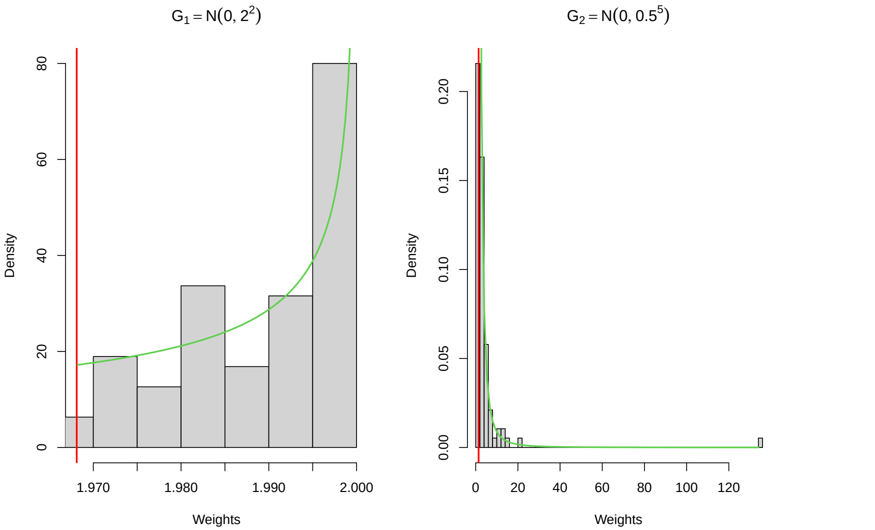


Example: Mystery Target

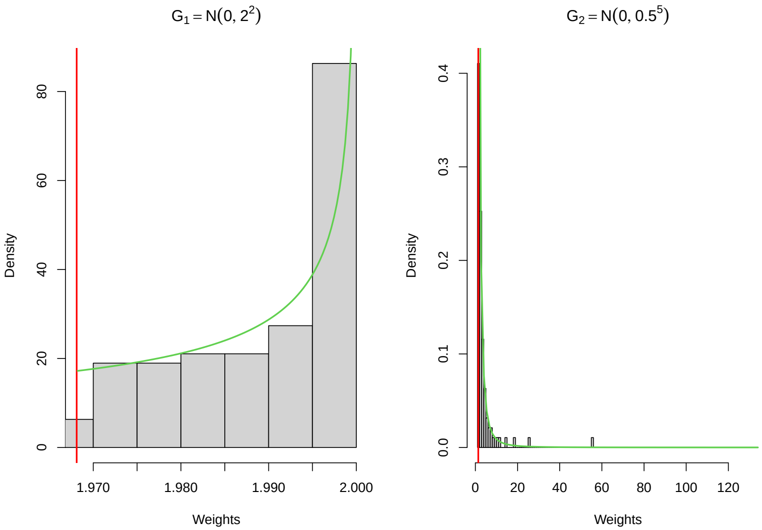


$$\hat{k}_1 \approx -1.81 \qquad \hat{k}_2 \approx 0.72$$

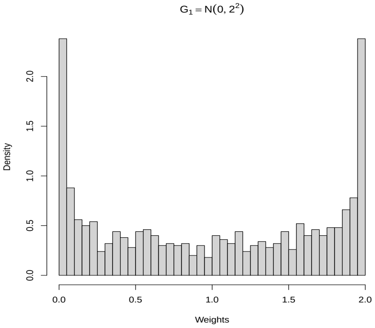
Example: Mystery Target

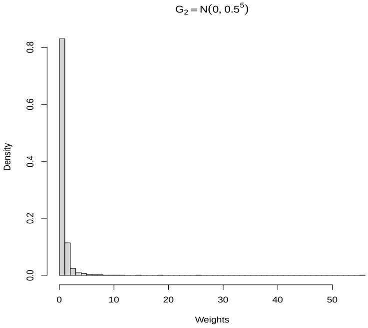


Example: Mystery Target



Example: Mystery Target





$$ESS_1 \approx 662$$
$$ESS_1^{(\text{trunc})} \approx 662$$
$$ESS_1^{(\text{PS})} \approx 662$$

$$ESS_2 \approx 54$$
$$ESS_2^{(\text{trunc})} \approx 245$$
$$ESS_2^{(\text{PS})} \approx 160$$

Adaptive IS

- Alternative approach: directly optimize ESS
- Adaptive Importance Sampling:
  - (Akyildiz and Míguez, 2021)

1. Choose a (parametric) family of proposals
2. Iteratively update the proposal to maximize ESS

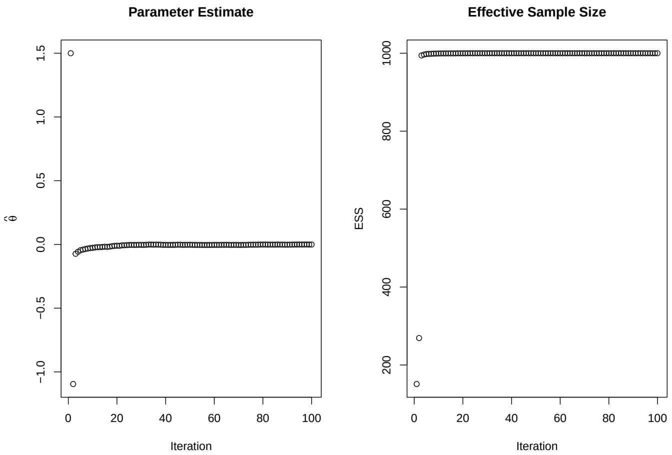
Stochastic Approximation

- Actually, we want to maximize a population-level analog: ESS^*
- If we had ESS^*, we would do gradient ascent
  - $\theta_{k+1} = \theta_k + \alpha \nabla ESS^*(\theta_k)$
- Instead, do gradient ascent on ESS
  - $\hat{\theta}_{k+1} = \hat{\theta}_k + \alpha_k \nabla ESS(\hat{\theta}_k)$

Stochastic Approximation

- Originally developed for root finding with noise
  - (Robbins and Monro, 1951)
- Quickly adapted for optimization
  - Use noisy evaluations for finite difference
  - (Kiefer and Wolfowitz, 1952)
- Very well developed theory
- Stochastic gradient descent

Example: Mystery Target



$\hat{\theta}^{(ESS)}_{\text{end}} \approx -8 \times 10^{-4}$ $ESS_{\text{end}} \approx 1000 - (7 \times 10^{-4})$

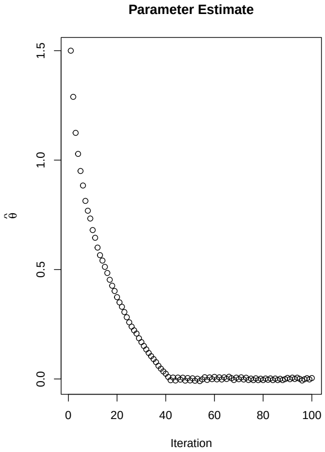
Our Method

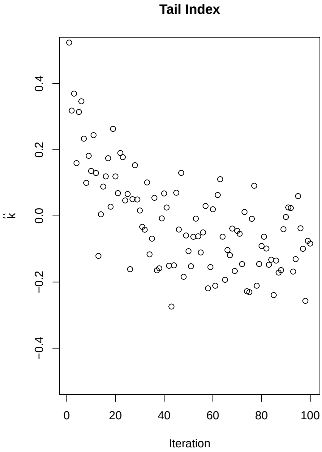
- Recall: Be careful using IS means to diagnose IS
- Vehtari et al. give an alternative
  - Shape parameter of fitted tail distribution, $\hat{k}$

Our Method

- Use diagnostic as objective function
- Apply stochastic approximation to minimize $\hat{k}$
  - More precisely, $k(\theta)$

Example: Mystery Target



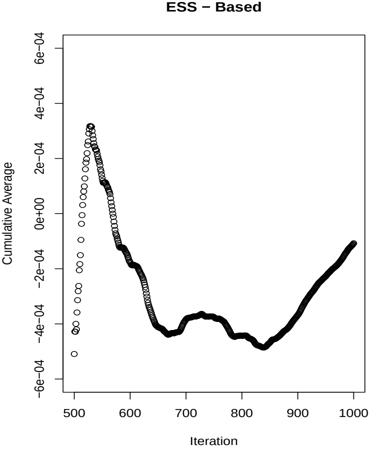


$$\hat{\theta}^{(PS)}_{\text{end}} \approx 4 \times 10^{-3}$$

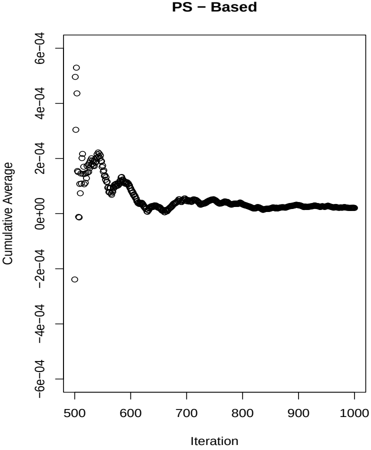
Example: Mystery Target

- Performance tends to be better if we average all the estimates
- Call this $\bar{\theta}$

Example: Mystery Target



$$\bar{\theta}_{\text{end}}^{(ESS)} \approx -1 \times 10^{-4}$$



$$\bar{\theta}_{\text{end}}^{(PS)} \approx 2 \times 10^{-5}$$

Another Example

Recap

- Importance sampling and extensions
  - Truncation
  - Pareto Smoothing
- Diagnostics for importance sampling
  - Effective sample size
  - Pareto tail index
- Adaptive importance sampling
  - Stochastic approximation

Thank You

Some References

Akyildiz, Ö. D. and Míguez, J. (2021). Convergence rates for optimized adaptive importance samplers. *Statistics and Computing*, 31(12).

Chatterjee, S. and Diaconis, P. (2018). The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2).

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2).

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3).

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3).

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2022). Pareto smoothed importance sampling. *ArXiv*.

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2024). Pareto smoothed importance sampling. *Journal of Machine Learning Research*, 25(72).

Zhang, J. and Stephens, M. A. (2009). A new and efficient estimation method for the generalized Pareto distribution. *Technometrics*, 51(3).