Adaptive Pareto Smoothed Importance Sampling

William Ruth

Joint work with Payman Nickchi

RichCon
2024

Introduction

- Importance Sampling
- Measuring performance
- Improving performance
  - Modifications
  - Optimization

Importance Sampling

- Need to compute an expected value
  - $\mathbb{E}_F \varphi(X) = \int \varphi(x) f(x) dx$
- Can't do the integral

- Monte Carlo approximation:
  - Simulate from F
  - Average over simulations

- $\hat{\mathbb{E}} = \sum_i \frac{\varphi(X_i)}{M}, \ X_i \overset{\text{iid}}{\sim} F$

Importance Sampling

- Simulating from F might be hard
- "Multiply by 1":

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

Importance Sampling

$$\mathbb{E}_F \varphi(X) = \mathbb{E}_G \left[\varphi(X) \cdot w(X) \right]$$

- G can be nearly anything*
    - *Some choices will be better than others
- Simulate from G to estimate $\mathbb{E}_G \left[\varphi(X) \cdot w(X) \right]$
    - By extension, estimate $\mathbb{E}_F \varphi(X)$

$$\hat{\mathbb{E}} = \sum_i \frac{\varphi(X_i) \cdot w(X_i)}{M}, X_i \overset{\text{iid}}{\sim} G$$

Example: Mystery Target

- f unknown, but can be evaluated
- Try some proposals:
  - $G_1 \sim N(0, 1)$
  - $G_2 \sim N(2, 1)$
- Use $M = 1000$ samples from proposal
  - $\hat{\mathbb{E}}_1 =$
  - $\hat{\mathbb{E}}_2 =$

Example: Mystery Target

Histograms of weights

Importance Sampling

- Can we quantify this difference?
  - Yes!
- "Effective Sample Size":

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

$$1 \leq ESS \leq M$$
$$M \geq \hat{\rho} \geq 1$$

Example: Mystery Target

Histograms of weights with ESS

Importance Sampling

- Problem: Low ESS $\rightarrow$ can't estimate means
- But ESS *is* a mean
  - (Chatterjee and Diaconis, 2018)

Improving IS

- Large variance in weights is bad
  - Reduce variance
  - Shrink large weights
- Truncated IS
- Pareto Smoothed IS

Improving IS

- Truncated Importance Sampling:
  - (Ionides, 2008)

1. Choose a threshold
2. Set any weights above threshold equal to threshold

Example: Mystery Target

Histograms of weights with threshold

Example: Mystery Target

Histograms of truncated weights

Example: Mystery Target

Histograms of truncated weights with before and after ESS

Improving IS

- Pareto Smoothed Importance Sampling:
    - (Vehtari et al., 2022)

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

Histograms of weights with threshold

Example: Mystery Target

Histograms of weights with threshold and fitted GPD density above threshold

Example: Mystery Target

Histograms of smoothed weights

Example: Mystery Target

Histograms of smoothed weights with ESS for raw, truncated and smoothed weights

Adaptive IS

- Modifications are nice, but require creativity
- Alternative: directly optimize ESS

- Adaptive Importance Sampling:
  - (Akyildiz and Míguez, 2021)

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

Adaptive IS

- Recall:

$$ESS = \frac{M}{\sum_i w(X_i)^2} =: \frac{M}{\hat{\rho}}$$

- Want to maximize a population-level analog
  - Equivalently, minimize $\rho = \mathbb{E}_G \left[w(X)^2 \right]$

- We only get ESS, $\hat{\rho}$
- Noisy version of the function we want to optimize

Stochastic Approximation

- If we had ρ, do gradient descent
- $\theta_{k+1} = \theta_k - \alpha \nabla \rho(\theta_k)$

- Instead, do gradient descent on $\hat{\rho}$
- $\hat{\theta}_{k+1} = \hat{\theta}_k - \alpha_k \nabla \hat{\rho}(\hat{\theta}_k)$

- Stochastic Approximation
    - (Robbins and Monro, 1951)

Stochastic Approximation

- Have to choose $\{\alpha_k\}$ carefully
- May not have $\nabla \hat{\rho}$
    - Finite difference approximation
    - (Kiefer and Wolfowitz, 1952)
- Improve performance by cumulative averaging

Example: Mystery Target

- Trajectory of $\hat{\theta}$
- Trajectory of ESS

- Values of above at convergence

Our Method

- Remember Chatterjee and Diaconis
  - Be careful using IS means to diagnose IS
- Vehtari et al. give an alternative
  - Shape parameter of fitted tail distribution, $\hat{k}$
  - "Tail Index"
- Theoretical and empirical support for $\hat{k}$ as diagnostic
  - Smaller is better

Our Method

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

Example: Mystery Target

- Trajectory of $\hat{\theta}$
- Trajectory of $\hat{k}$ and k

- Values of above at convergence

- Big reveal!

Recap

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

Acknowledgements

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*.

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