Bayesian Coresets: Revisiting the Nonconvex Optimization Perspective

Jacky Y. Zhang[1] Rajiv Khanna[2] Anastasios Kyrillidis[3] Oluwasanmi Koyejo[1]
[1]University of Illinois at Urbana-Champaign [2]University of California at Berkeley [3]Rice University

Scan Me for Full Paper

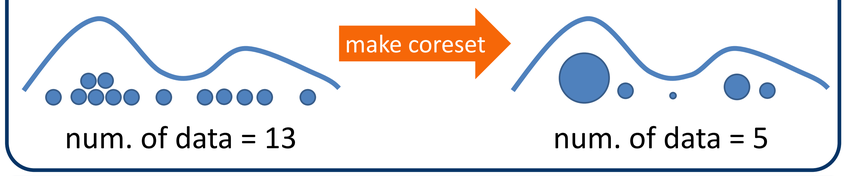
Motivation

Problem: Bayesian inference on large datasets can be impractical.

Question

Can we quickly find a small coreset (weighted subset) to summarize the whole dataset, improving scalability for Bayesian inference?

Illustration:



To Construct Bayesian Coresets

Settings:

- Given *i.i.d.* data samples $X = \{x_i\}_{i=1}^n$, and a probability model parametrized by $\theta \in \mathbb{R}^D$.
- The posterior distribution is

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)}$$

(likelihood; prior; constant)

Notations:

- $w := [w_1, \dots, w_n]^\top \in \mathbb{R}_+^n$ is a non-negative vector.
- $w \in C_k := \{w: ||w||_0 \le k\}$ is a k-sparse vector
- $w \in C_k \cap \mathbb{R}_+^n$ defines a coreset of size k.

($C_k \cap \mathbb{R}_+^n$: nonconvex set)

To find a Bayesian coreset of size k ($k < n$):

- The coreset needs to approximate the likelihood.
- The full-dataset log-likelihood $\mathcal{L}(\theta) := \log p(X|\theta)$

$$\mathcal{L}(\theta) = \sum_{i=1}^n \log p(x_i|\theta) = \sum_{i=1}^n \mathcal{L}_i(\theta)$$

- The coreset log-likelihood

$$\mathcal{L}_w(\theta) := \sum_{i=1}^n w_i \mathcal{L}_i(\theta),$$

- To minimize the distance between $\mathcal{L}$ and $\mathcal{L}_w$:

Bayesian Coreset Construction

$$\operatorname{argmin}_w \; f(w) := \mathbb{E}_{\theta\sim\pi}\left||\mathcal{L}(\theta) - \mathcal{L}_w(\theta)|\right|^2 \quad (1)$$
$$s.t. \qquad w \in C_k \cap \mathbb{R}_+^n$$

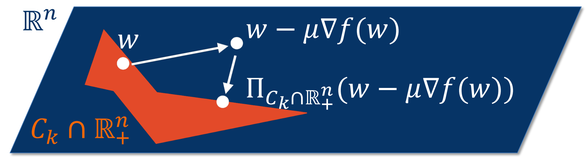
- π can be a cheap approximation of the posterior.

The Proposed Method

Goal: to solve the nonconvex optimization problem (1) for Bayesian coreset construction.

Core Idea: Iterative Hard Thresholding (IHT).

- Hard threshold operator $\Pi_{C_k\cap\mathbb{R}_+^n}(w)$: choose the k largest elements of w, and set the rest to be 0.
- Illustration of IHT with step size μ. Iteratively do:



Proposed Algorithms: Automated Accelerated IHT (A-IHT and A-IHT II). Key components are described below.

A-IHT

IHT + step size selection + active subspace exploration + momentum

A-IHT II

IHT + step size selection + active subspace exploration + momentum + debias step

Theoretical Analysis

Time Complexity: w.r.t. dataset size n, coreset size k.

- A-IHT and A-IHT II are $O(n \log k)$.
- Recent work GIGA [1] and SparseVI [2] are $O(n\,k)$.

Convergence Analysis:

- Standard assumption: Restricted Isometry Property (RIP), *i.e.*, loosely speaking convexity + smoothness.

Convergence Theorem

With the RIP assumption, A-IHT converges to the optimal solution linearly under certain condition.

References

[1] Campbell et al. Bayesian coreset construction via greedy iterative geodesic ascent. ICML'18
[2] Campbell et al. Sparse Variational Inference: Bayesian Coresets from Scratch. NeurIPS'19

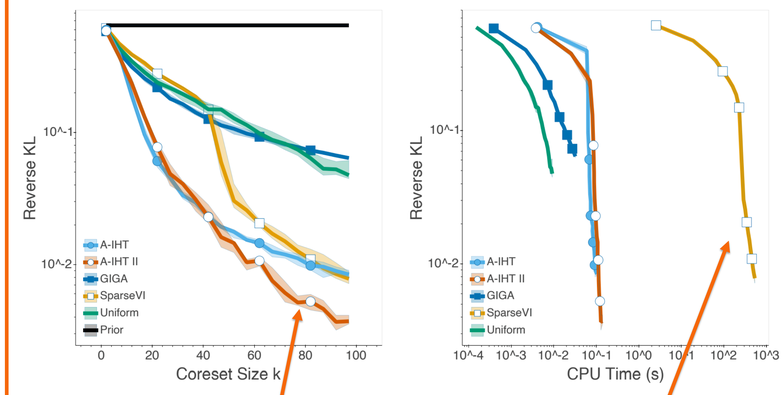
Experiments

Task: Bayesian Logistic Regression
Baselines: Random, GIGA [1], and SparseVI [2].
Datasets:

- The *reduced phishing* dataset (Include SparseVI): dataset size $n = 500$; parameter dimension $D = 10$.
- The *original phishing* dataset (No SparseVI): dataset size $n = 11055$; parameter dimension $D = 68$.

Evaluation: KL divergence between true posterior and coreset posterior.

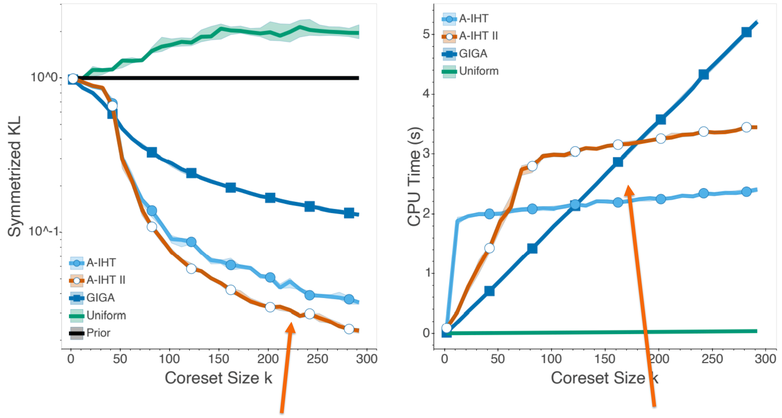
Results with the *reduced phishing* dataset



A-IHT & A-IHT II find better coresets

SparseVI costs $\times 10^4$ more time

Results with the *original phishing* dataset



A-IHT & A-IHT II find better coresets

GIGA starts to cost more time at coreset size $k \approx 2\%$ of the dataset.