Bayesian Coresets: Revisiting the Nonconvex Optimization Perspective Appendix

Appendix Contents

A Automated Accelerated IHT with De-bias Step

In the main text, we mention that Algorithm 2 can be boosted better in practice using de-bias steps. Here we present the algorithm with de-bias step, as shown in Algorithm 3.

Like Automated Accelerated IHT, Algorithm 3 also starts with active subspace expansion, *i.e.*, line 3 & 4. As $\mathcal{Z} = \text{supp}(z_t) = \text{supp}(w_{t-1}) \cup \text{supp}(w_t)$ is a $2k$-sparse index set, the expanded index set $\mathcal{S}$ is a $3k$-sparse index set that is the union of the support of three elements, *i.e.*,

$$\mathcal{S} = \text{supp}(w_{t-1}) \cup \text{supp}(w_t) \cup \text{supp}(\Pi_{\mathcal{C}_k \setminus \mathcal{Z}}(\nabla f(z_t))).$$

We note that, with a little abuse of notation, we use $\mathcal{Z}$ to denote both the support set $\mathcal{Z} \subset [n]$, and the subspace restricted by the support, *i.e.*, $\{x \in \mathbb{R}^n \mid \text{supp}(x) \subseteq \mathcal{Z}\}$, depending on the context.

Algorithm 3 Automated Accelerated IHT - II (A-IHT II)

input Objective $f(w) = \|y - \Phi w\|_2^2$; sparsity k

1: $t = 0$, $z_0 = 0$, $w_0 = 0$
2: **repeat**
3: $\mathcal{Z} = \text{supp}(z_t)$
4: $\mathcal{S} = \text{supp}(\Pi_{\mathcal{C}_k \setminus \mathcal{Z}}(\nabla f(z_t))) \cup \mathcal{Z}$ where $|\mathcal{S}| \leq 3k$ {active subspace expansion}
5: $\widetilde{\nabla}^{(1)} = \nabla f(z_t)\big|_{\mathcal{S}}$
6: $\mu_t^{(1)} = \arg\min_\mu f(z_t - \mu \widetilde{\nabla}^{(1)}) = \frac{\|\widetilde{\nabla}^{(1)}\|_2^2}{2\|\Phi \widetilde{\nabla}^{(1)}\|_2^2}$ {step size selection}
7: $x_t = \Pi_{\mathcal{C}_k \cap \mathbb{R}_+^n}\left(z_t - \mu_t^{(1)} \nabla f(z_t)\right)$ {projected gradient descent}
8: $\widetilde{\nabla}^{(2)} = \nabla f(x_t)\big|_{\text{supp}(x)}$
9: $\mu_t^{(2)} = \arg\min_\mu f(x_t - \mu \widetilde{\nabla}^{(2)}) = \frac{\|\widetilde{\nabla}^{(2)}\|_2^2}{2\|\Phi \widetilde{\nabla}^{(2)}\|_2^2}$ {step size selection}
10: $w_{t+1} = \Pi_{\mathbb{R}_+^n}(x_t - \mu_t^{(2)} \widetilde{\nabla}^{(2)})$ {de-bias step}
11: $\tau_{t+1} = \arg\min_\tau f(w_{t+1} + \tau(w_{t+1} - w_t)) = \frac{\langle y - \Phi w_{t+1}, \Phi(w_{t+1} - w_t)\rangle}{2\|\Phi(w_{t+1} - w_t)\|_2^2}$
12: $z_{t+1} = w_{t+1} + \tau_{t+1}(w_{t+1} - w_t)$ {momentum step}
13: $t = t + 1$
14: **until** Stop criteria met
15: **return** w_t

The subspace corresponding to this index set $\mathcal{S}$ is a subspace that the algorithm considers as potential to achieve low loss within. Therefore, in the next step, we perform projected gradient descent in this expanded subspace. Note that we use $\nabla f(\cdot)\big|_{\mathcal{S}}$ to denote a sparse subset $\mathcal{S}$ of the gradient, *i.e.*, setting the i^{th} entry of $\nabla f(\cdot)$ to 0 if $i \notin \mathcal{S}$.

The projected gradient descent step consists of three sub-steps, *i.e.*, step size selection (line 6), gradient descent (line 7), and projection to non-negative k-sparse restricted domain (line 7). The step size selection is performed by an exact line search to obtain a good step size automatically. The projection step (line 7) is where we do "hard thresholding" to obtain a k-sparse solution x_t. As mentioned before, this projection step can be done optimally in the sense of ℓ_2-norm by choosing the k-largest non-negative elements.

Then, we come to the key difference between Algorithm 2 and Algorithm 3, *i.e.*, the de-bias step at line 8, 9 & 10. With additional de-bias steps, we adjust the solution k-sparse solution x_t inside its own sparse space, *i.e.*, the space corresponding to $\text{supp}(x_t)$, such that a better k-sparse solution is found. After computing the gradient (line 8), another exact line search is performed (line 9). By gradient descent and imposing the non-negativity constraint (line 10), we have the solution w_{t+1} for this iteration.

Lastly, the momentum step (line 11 & 12) is the same as Algorithm 2. We select the momentum term as the minimizer of the objective: $\tau_{t+1} = \arg\min_\tau f(w_{t+1} + \tau(w_{t+1} - w_t))$, and then apply the momentum to our solutions w_{t+1} and w_t as $z_{t+1} = w_{t+1} + \tau_{t+1}(w_{t+1} - w_t)$ to capture memory in the algorithm. Momentum can offer faster convergence rate for convex optimization (Nesterov, 1983).

B Theoretical Analysis

In this section, we provide a detailed theoretical analysis that is abstracted in the main paper due to space limitation. All of the proofs are defer to section C for clarity. To begin with, let us show that all of the projection operators used in our algorithms can be done optimally and efficiently.

Given an index set $\mathcal{S} \subseteq [n]$, the projection of w to the subspace with support $\mathcal{S}$ is $\Pi_{\mathcal{S}}(w)$, which can be done optimally by setting $w_{\mathcal{S}^c} = 0$, where $\mathcal{S}^c$ denotes the complement of $\mathcal{S}$. We note that, with a little abuse of notation, we use $\mathcal{S}$ to denote both the support set $\mathcal{S} \subset [n]$, and the subspace restricted by the support, *i.e.*, $\{x \in \mathbb{R}^n \mid \text{supp}(x) \subseteq \mathcal{S}\}$. The projection to non-negative space, *i.e.*, $\Pi_{\mathbb{R}_+^n}(w)$, can also be done optimally and efficiently by setting the negative entries to zero. Moreover, $\Pi_{\mathcal{C}_k}$ is shown to be optimal by simply picking the top k largest (in absolute value) entries. It is also the case for $\Pi_{\mathcal{C}_k \cap \mathbb{R}_+^n}(w)$, where it can be done by picking the top k

largest non-negative entries. The optimality for the above projections is in terms of Euclidean distance.

Let us show the optimality for $\Pi_{\mathcal{C}_k \cap \mathbb{R}^n_+}(w)$. Given a k-sparse support $\mathcal{S}$, the optimal projection of $w \in \mathbb{R}^n$ to its restricted sparsity space intersecting the non-negative orthant is $w' = \Pi_{\mathcal{S} \cap \mathbb{R}^n_+}(w)$. We can see that for entry $i \in [n]$, $w'_i = w_i$ if $i \in \mathcal{S}$ and $w_i \geq 0$, and $w'_i = 0$ otherwise. Therefore, the distance between w and its projection to $\mathcal{S} \cap \mathbb{R}^n_+$ is $\|w' - w\|_2^2 = \|w\|_2^2 - \sum_{i \in S, w_i > 0} w_i^2$. As $\Pi_{\mathcal{C}_k \cap \mathbb{R}^n_+}(w) = \min_{\mathcal{S}:|\mathcal{S}| \leq k} \Pi_{\mathcal{S} \cap \mathbb{R}^n_+}(w)$, we can see that it is the support with k largest w_i that has the least distance. Therefore, simply picking top k largest non-negative entries gives the optimal projection.

We give the convergence analysis for our main algorithm Automated Accelerated IHT in Algorithm 2. One standard assumption about the objective is required for the theory to begin, *i.e.,* RIP property, which is a normal assumption in IHT context, reflecting convexity and smoothness of the objective in some sense (Khanna & Kyrillidis, 2018; Kyrillidis & Cevher, 2014). We note that the assumption is not necessary but is sufficient. For example, if the number of samples required to exactly construct $\hat{g}$ is less than the coreset size ($a_k = 0$ in RIP), so that the system becomes underdetermined, then local minima can be global one achieving zero-error without the RIP. On the other hand, when the number of samples goes to infinity, RIP ensures the eigenvalues of the covariance matrix, $cov[\mathcal{L}_i(\theta), \mathcal{L}_j(\theta)]$ where $\theta \sim \hat{\pi}$, are lower and upper bounded. It is an active area of research in random matrix theory to quantify RIP constants *e.g.* see (Baraniuk et al., 2008).

Assumption 1 (Restricted Isometry Property)**.** *Matrix Φ in the objective function satisfies the RIP property,* i.e.*, for $\forall w \in \mathcal{C}_k$*

$$\alpha_k \|w\|_2^2 \leq \|\Phi w\|_2^2 \leq \beta_k \|w\|_2^2.$$

It is known that there are connections between RIP and restricted strong convexity and smoothness assumptions (Chen & Sanghavi, 2010); thus our results could potentially generalized for different convex $f(\cdot)$ functions.

Leading to our main theorem, some useful technical properties are presented. An useful observation is that, for any set $\mathcal{S} \subseteq [n]$, the projection operator $\Pi_{\mathcal{S}} : \mathbb{R}^n \to \mathbb{R}^n$ is in fact a linear operator in the form of a diagonal matrix

$$\Pi_{\mathcal{S}} = \{\text{diag}(\delta_i)\}_{i=1}^n,$$

where δ_i is an indicator function: $\delta_i = 1$ if $i \in \mathcal{S}$, and $\delta_i = 0$ otherwise. This leads to our first lemma.

Lemma 1. *Supposing Φ satisfies the RIP assumption, given a sparse set $\mathcal{S} \subseteq [n]$ and $|\mathcal{S}| \leq k$, for $\forall w \in \mathbb{R}^n$ it holds that*

$$\alpha_k \|\Pi_{\mathcal{S}} w\|_2 \leq \|\Pi_{\mathcal{S}} \Phi^\top \Phi \Pi_{\mathcal{S}} w\|_2 \leq \beta_k \|\Pi_{\mathcal{S}} w\|_2.$$

Lemma 1 reveals a property of the eigenvalues of $\Pi_{\mathcal{S}} \Phi^\top \Phi \Pi_{\mathcal{S}}$, which leads to the following lemma that bounds an iterated projection using the RIP property.

Lemma 2. *Supposing Φ satisfies the RIP assumption, given two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [n]$ and $|\mathcal{S}_1 \cup \mathcal{S}_2| \leq k$, for $\forall w \in \mathbb{R}^n$ it holds that*

$$\|\Pi_{\mathcal{S}_1} \Phi^\top \Phi \Pi_{\mathcal{S}_1 \setminus \mathcal{S}_2} \Pi_{\mathcal{S}_2} w\|_2 \leq \tfrac{\beta_k - \alpha_k}{2} \cdot \|\Pi_{\mathcal{S}_2} w\|_2.$$

Armed with above two lemmas, we are ready to prove convergence for Automated Accelerated IHT (Algorithm 2). A key observation is that solution w_{t+1} found by Algorithm 2 is derived by the following two steps:

$$\{w_t, w_{t-1}\} \xRightarrow[\text{line 9}]{①} z_t \xRightarrow[\text{line 7}]{②} w_{t+1}.$$

Procedure ① is a momentum step, with momentum size chosen automatically; procedure ② aims for exploration in an expanded subspace spanned by a $3k$-sparse subset $\mathcal{S}$, and projecting to k-sparse non-negative subspace.

We break down the proof into two parts. Denoting the optimal solution as

$$w^\star = \underset{w \in \mathcal{C}_k \cap \mathbb{R}^n_+}{\arg\min} \|y - \Phi w\|_2^2,$$

we propose the following two lemmas for the two steps respectively.

Lemma 3. *For procedure ①, the following iterative invariant holds.*

$$\|z_t - w^\star\|_2 \leq |1+\tau_t| \cdot \|w_t - w^\star\|_2 + |\tau_t| \cdot \|w_{t-1} - w^\star\|_2.$$

For the second procedure, we consider the actual step size μ_t automatically chosen by the algorithm. Noting that $|\mathrm{supp}(\widetilde{\nabla}_t)| \leq 3k$, according to RIP we can see that the step size $\mu_t = \frac{\|\widetilde{\nabla}_t\|_2^2}{2\|\Phi\widetilde{\nabla}_t\|_2^2}$ is bounded as

$$\frac{1}{2\beta_{3k}} \leq \mu_t \leq \frac{1}{2\alpha_{3k}}.$$

Therefore, using the Lemma 1 and Lemma 2, one can prove the following lemma.

Lemma 4. *For procedure ②, the following iterative invariant holds.*

$$\|w_{t+1} - w^\star\|_2 \leq \rho\|z_t - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

where $\rho = \left(2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\} + \frac{\beta_{4k}-\alpha_{4k}}{\alpha_{3k}}\right)$*, and* $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2$ *is the optimal objective value.*

Combining the above two lemmas leads to our main convergence analysis theorem.

Theorem 1 (Restated)**.** *In the worst case scenario, with Assumption 1, the solutions path find by Automated Accelerated IHT (Algorithm 2) satisfy the following iterative invariant.*

$$\|w_{t+1} - w^\star\|_2 \leq \rho|1+\tau_t| \cdot \|w_t - w^\star\|_2 + \rho|\tau_t| \cdot \|w_{t-1} - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

where $\rho = \left(2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\} + \frac{\beta_{4k}-\alpha_{4k}}{\alpha_{3k}}\right)$*, and* $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2$ *is the optimal objective value.*

The theorem provides an upper bound invariant among consecutive iterates of the algorithm. To have better sense of convergence rate, we assume the optimal solution achieves $\|\epsilon\|_2 = 0$. Theorem 1 then implies

$$\|w_{t+1} - w^\star\|_2 \leq \rho(1+|\tau_t|)\|w_t - w^\star\|_2 + \rho|\tau_t| \cdot \|w_{t-1} - w^\star\|_2.$$

Given the above homogeneous recurrence, we can solve for the following corollary that shows linear convergence of the proposed algorithm under given conditions.

Corollary 1 (Restated)**.** *Given the iterative invariant as stated in Theorem 1, and assuming the optimal solution achieves* $\|\epsilon\|_2 = 0$*, the solution found by Algorithm 2 satisfies:*

$$f(w_{t+1}) - f(w^\star) \leq \phi^t \left(\frac{\beta_{2k}}{\alpha_{2k}} f(w_1) + \frac{\rho\tau\beta_{2k}}{\phi\alpha_k} f(w_0)\right),$$

where $\phi = (\rho(1+\tau) + \sqrt{\rho^2(1+\tau)^2 + 4\rho\tau})/2$ *and* $\tau = \max_{i\in[t]} |\tau_i|$*. It is sufficient to show linear convergence to the global optimum, when* $\phi < 1$*, or equivalently* $\rho < 1/(1+2\tau)$*.*

C Proofs

This section provides proofs for the theoretical results presented in the previous section. For the sake of good readability, the lemma/theorem to be proven is also restated preceding its proof.

C.1 Proof of Lemma 1

Lemma 1 (Restated)**.** *Supposing* Φ *satisfies the RIP assumption, given a sparse set* $\mathcal{S} \subseteq [n]$ *and* $|\mathcal{S}| \leq k$*, for* $\forall w \in \mathbb{R}^n$ *it holds that*

$$\alpha_k\|\Pi_{\mathcal{S}} w\|_2 \leq \|\Pi_{\mathcal{S}}\Phi^\top\Phi\Pi_{\mathcal{S}} w\|_2 \leq \beta_k\|\Pi_{\mathcal{S}} w\|_2.$$

Proof. Recall that $\Pi_{\mathcal{S}}$ is a linear operator that projects a vector $w \in \mathbb{R}^n$ to sparse restricted set with support $\mathcal{S}$ by simply setting $w_i = 0$ for each $i \notin \mathcal{S}$. As a result, for a k-sparse set $\mathcal{S}$, $\Pi_{\mathcal{S}} w$ is a k-sparse vector. Given that $\Phi \in \mathbb{R}^{m \times n}$ satisfies RIP property, for $\forall w \in \mathbb{R}^n$, it holds that

$$\alpha_k \|\Pi_{\mathcal{S}} w\|_2^2 \leq \|\Phi \Pi_{\mathcal{S}} w\|_2^2 \leq \beta_k \|\Pi_{\mathcal{S}} w\|_2^2. \tag{4}$$

Let us denote $b = \Phi \Pi_{\mathcal{S}} w$, and $\langle \cdot, \cdot \rangle$ as standard Euclidean inner product. With regular linear algebra manipulation, the following stands:

$$\begin{aligned}
\|\Pi_{\mathcal{S}} \Phi^\top b\|_2^2 &= \max_{x \in \mathbb{R}^n : \|x\|_2 = 1} \left(\langle \Pi_{\mathcal{S}} \Phi^\top b, x \rangle \right)^2 \\
&= \max_{x \in \mathbb{R}^n : \|x\|_2 = 1} \left(b^\top \Phi \Pi_{\mathcal{S}} x \right)^2 \\
&= \max_{x \in \mathbb{R}^n : \|x\|_2 = 1} \left(\langle b, \Phi \Pi_{\mathcal{S}} x \rangle \right)^2 \\
&= \max_{x \in \mathbb{R}^n : \|x\|_2 = 1} \left(\langle \Phi \Pi_{\mathcal{S}} w, \Phi \Pi_{\mathcal{S}} x \rangle \right)^2,
\end{aligned} \tag{5}$$

where the second equality is due to the fact that $\Pi_{\mathcal{S}}$ is symmetric, *i.e.*, $(\Pi_{\mathcal{S}} \Phi^\top b)^\top = b^\top \Phi \Pi_{\mathcal{S}}$.

Letting $x^\star$ be the solution of (5), we have the upper bound of (5):

$$(5) = \left(\langle \Phi \Pi_{\mathcal{S}} w, \Phi \Pi_{\mathcal{S}} x^\star \rangle \right)^2 \leq \|\Phi \Pi_{\mathcal{S}} w\|_2^2 \cdot \|\Phi \Pi_{\mathcal{S}} x^\star\|_2^2,$$

where the inequality is by Cauchy-Schwarz inequality applying on inner product.

On the other hand, the lower bound can be obtained by removing the maximizing operator and setting $x = \Pi_{\mathcal{S}} w / \|\Pi_{\mathcal{S}} w\|_2$, as follows. Denoting $x' = \Pi_{\mathcal{S}} w / \|\Pi_{\mathcal{S}} w\|_2$, we have,

$$(5) \geq \left(\langle \Phi \Pi_{\mathcal{S}} w, \Phi \Pi_{\mathcal{S}} x' \rangle \right)^2 = \|\Phi \Pi_{\mathcal{S}} w\|_2^2 \cdot \|\Phi \Pi_{\mathcal{S}} x'\|_2^2,$$

where the last equality is due to that $\Pi_{\mathcal{S}} w$ and x' are parallel.

Applying (4) to the above upper bound and lower bound, it follows that

$$\begin{aligned}
(5) &\leq \|\Phi \Pi_{\mathcal{S}} w\|_2^2 \cdot \|\Phi \Pi_{\mathcal{S}} x^\star\|_2^2 \leq \beta_k \|\Pi_{\mathcal{S}} w\|_2^2 \cdot \beta_k \|\Pi_{\mathcal{S}} x^\star\|_2^2, \\
(5) &\geq \|\Phi \Pi_{\mathcal{S}} w\|_2^2 \cdot \|\Phi \Pi_{\mathcal{S}} x'\|_2^2 \geq \alpha_k \|\Pi_{\mathcal{S}} w\|_2^2 \cdot \alpha_k \|\Pi_{\mathcal{S}} x'\|_2^2.
\end{aligned} \tag{6}$$

Noting that $x^\star$ is an unit-length vector, and the projection $\Pi_{\mathcal{S}}$ is done by setting elements to zero, we can see that $\|\Pi x^\star\|_2 \leq 1$. As $x' = \Pi_{\mathcal{S}} w / \|\Pi_{\mathcal{S}} w\|_2$ has already been a sparse vector in the restricted space by $\mathcal{S}$, we can see that $\|\Pi_{\mathcal{S}} x'\|_2 = \|x'\|_2 = 1$. Plugging them in (6), it holds that

$$\alpha_k^2 \|\Pi_{\mathcal{S}} w\|_2^2 = \alpha_k \|\Pi_{\mathcal{S}} w\|_2^2 \cdot \alpha_k \|\Pi_{\mathcal{S}} x'\|_2^2 \leq (5) \leq \beta_k \|\Pi_{\mathcal{S}} w\|_2^2 \cdot \beta_k \|\Pi_{\mathcal{S}} x^\star\|_2^2 \leq \beta_k^2 \|\Pi_{\mathcal{S}} w\|_2^2.$$

Plugging that $(5) = \|\Pi_{\mathcal{S}} \Phi^\top b\|_2^2 = \|\Pi_{\mathcal{S}} \Phi^\top \Phi \Pi_{\mathcal{S}} w\|_2^2$, and taking the square root, we finally have

$$\alpha_k \|\Pi_{\mathcal{S}} w\|_2 \leq \|\Pi_{\mathcal{S}} \Phi^\top \Phi \Pi_{\mathcal{S}} w\|_2 \leq \beta_k \|\Pi_{\mathcal{S}} w\|_2.$$

□

C.2 Proof of Lemma 2

Lemma 2 (Restated). *Supposing Φ satisfies the RIP assumption, given two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [n]$ and $|\mathcal{S}_1 \cup \mathcal{S}_2| \leq k$, for $\forall w \in \mathbb{R}^n$ it holds that*

$$\|\Pi_{\mathcal{S}_1} \Phi^\top \Phi \Pi_{\mathcal{S}_1 \setminus \mathcal{S}_2} \Pi_{\mathcal{S}_2} w\|_2 \leq \tfrac{\beta_k - \alpha_k}{2} \cdot \|\Pi_{\mathcal{S}_2} w\|_2.$$

Proof. Similar to the proof of Lemma 1, we first write the norm in the form of an inner product. Given two sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [n]$ and $|\mathcal{S}_1 \cup \mathcal{S}_2| \leq k$, for $\forall w \in \mathbb{R}^n$, with regular linear algebra manipulation, it holds that

$$\begin{aligned}
&\|\Pi_{\mathcal{S}_1}\Phi^\top\Phi\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\|_2 \\
&\quad= \max_{b\in\mathbb{R}^n:\|b\|_2=1} |\langle b, \Pi_{\mathcal{S}_1}\Phi^\top\Phi\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\rangle| \\
&\quad= \max_{b\in\mathbb{R}^n:\|b\|_2=1} |\langle \Phi\Pi_{\mathcal{S}_1}b, \Phi\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\rangle|, \qquad (7)
\end{aligned}$$

where the second equality is due to the fact that $\Pi_{\mathcal{S}_1}$ is symmetric.

Define two unit-length vectors

$$X = \frac{\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w}{\|\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\|_2}, \qquad Y = \frac{\Pi_{\mathcal{S}_1}b}{\|\Pi_{\mathcal{S}_1}b\|},$$

and we can see that $\langle X, Y\rangle = 0$, as $\mathcal{S}_1^c$ and $\mathcal{S}_1$ are disjoint. As a result, $\|X+Y\|_2^2 = \|X\|_2^2 + \|Y\|_2^2 = 2$. Moreover, given that $|\mathcal{S}_1 \cup \mathcal{S}_2| \leq k$, we can see that $X+Y$ is k-sparse. Applying the RIP property, the following holds:

$$2\alpha_k = \alpha_k\|X+Y\|_2^2 \leq \|\Phi X + \Phi Y\|_2^2 \leq \beta_k\|X+Y\|_2^2 = 2\beta_k.$$

Similarly, $\|X-Y\|_2^2 = 2$ and $X-Y$ is also k-sparse:

$$2\alpha_k \leq \|\Phi X - \Phi Y\|_2^2 \leq 2\beta_k.$$

Noting that

$$\langle \Phi X, \Phi Y\rangle = \frac{\|\Phi X + \Phi Y\|_2^2 - \|\Phi X - \Phi Y\|_2^2}{4},$$

we can see the following,

$$-\frac{\beta_k - \alpha_k}{2} \leq \langle \Phi X, \Phi Y\rangle \leq \frac{\beta_k - \alpha_k}{2}. \qquad (8)$$

Recall that

$$(7) = \max_{\|b\|_2=1} |\langle \Phi X, \Phi Y\rangle| \cdot \|\Pi_{\mathcal{S}_1}b\|_2 \cdot \|\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\|_2,$$

and apply (8) to the above, we conclude that

$$\begin{aligned}
(7) &\leq \max_{\|b\|_2=1} \frac{\beta_k - \alpha_k}{2} \cdot \|\Pi_{\mathcal{S}_1}b\|_2 \cdot \|\Pi_{\mathcal{S}_1^c}\Pi_{\mathcal{S}_2}w\|_2 \\
&\leq \frac{\beta_k - \alpha_k}{2}\|\Pi_{\mathcal{S}_2}w\|_2.
\end{aligned}$$

□

C.3 Proof of Lemma 3

Lemma 3 (Restated)**.** *For procedure ①, the following iterative invariant holds.*

$$\|z_t - w^\star\|_2 \leq |1+\tau_t| \cdot \|w_t - w^\star\|_2 + |\tau_t| \cdot \|w_{t-1} - w^\star\|_2.$$

Proof. According to line 9 in Algorithm 2, with some regular linear algebra manipulation, we can derive

$$\begin{aligned}
\|z_t - w^\star\|_2 &= \|w_t + \tau_t(w_t - w_{t-1}) - w^\star\|_2 \\
&= \|(1+\tau_t)(w_t - w^\star) + \tau_t(w^\star - w_{t-1})\|_2 \\
&\leq |1+\tau_t|\|w_t - w^\star\|_2 + |\tau_t|\|w_{t-1} - w^\star\|_2,
\end{aligned}$$

where the last inequality is done by triangle inequality.

□

C.4 Proof of Lemma 4

Lemma 4 (Restated). *For procedure ②, the following iterative invariant holds.*

$$\|w_{t+1} - w^\star\|_2 \leq \rho\|z_t - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

where $\rho = \left(2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\} + \frac{\beta_{4k}-\alpha_{4k}}{\alpha_{3k}}\right)$, *and* $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2$ *is the optimal objective value.*

Proof. Denoting $v = z_t - \mu_t \nabla f(z_t)$, and set $\mathcal{S}_\star = \mathrm{supp}(w_{t+1}) \cup \mathrm{supp}(w^\star)$, we begin by the projection at line 7 in Algorithm 2. Applying the triangle inequality,

$$\|w_{t+1} - w^\star\|_2 \leq \|w_{t+1} - \Pi_{\mathcal{S}_\star} v\|_2 + \|\Pi_{\mathcal{S}_\star} v - w^\star\|_2. \tag{9}$$

As $\mathcal{S}_\star = \mathrm{supp}(w_{t+1}) \cup \mathrm{supp}(w^\star)$, we can observe that $\langle w_{t+1}, \Pi_{\mathcal{S}_\star{}^c} v\rangle = 0$ and $\langle w^\star, \Pi_{\mathcal{S}_\star{}^c} v\rangle = 0$. As a result,

$$\begin{aligned}
\|w_{t+1} - \Pi_{\mathcal{S}_\star} v\|_2^2 &= \|w_{t+1} - v + \Pi_{\mathcal{S}_\star{}^c} v\|_2^2 \\
&= \|w_{t+1} - v\|_2^2 + \|\Pi_{\mathcal{S}_\star{}^c} v\|_2^2 + 2\langle w_{t+1} - v, \Pi_{\mathcal{S}_\star{}^c} v\rangle \\
&= \|w_{t+1} - v\|_2^2 + \|\Pi_{\mathcal{S}_\star{}^c} v\|_2^2 + 2\langle -v, \Pi_{\mathcal{S}_\star{}^c} v\rangle \\
&\leq \|w^\star - v\|_2^2 + \|\Pi_{\mathcal{S}_\star{}^c} v\|_2^2 + 2\langle -v, \Pi_{\mathcal{S}_\star{}^c} v\rangle \\
&= \|w^\star - v\|_2^2 + \|\Pi_{\mathcal{S}_\star{}^c} v\|_2^2 + 2\langle w^\star - v, \Pi_{\mathcal{S}_\star{}^c} v\rangle \\
&= \|w^\star - v + \Pi_{\mathcal{S}_\star{}^c} v\|_2^2 \\
&= \|w^\star - \Pi_{\mathcal{S}_\star} v\|_2^2,
\end{aligned}$$

where the inequality is due to the projection step $w_{t+1} = \Pi_{\mathcal{C}_k \cap \mathbb{R}_+^n} v$ is done optimally, and $w^\star \in \mathcal{C}_k \cap \mathbb{R}_+^n$. Plugging the above inequality into (9), it holds that

$$\|w_{t+1} - w^\star\|_2 \leq 2\|\Pi_{\mathcal{S}_\star} v - w^\star\|_2. \tag{10}$$

Expanding v and denoting $\epsilon = \Phi w^\star - y$, we have

$$\begin{aligned}
v &= z_t - \mu_t\left(\nabla f(z_t)\right) \\
&= z_t - \mu_t\left(2\Phi^\top(\Phi z_t - y)\right) \\
&= z_t - \mu_t\left(2\Phi^\top\Phi(z_t - w^\star) + 2\Phi^\top(\Phi w^\star - y)\right) \\
&= z_t - 2\mu_t\Phi^\top\Phi(z_t - w^\star) - 2\mu_t\Phi^\top\epsilon.
\end{aligned}$$

Plugging the above into inequality (10), we can further expand

$$\begin{aligned}
\|w_{t+1} - w^\star\|_2 &\leq 2\|\Pi_{\mathcal{S}_\star}(z_t - 2\mu_t\Phi^\top\Phi(z_t - w^\star) - 2\mu_t\Phi^\top\epsilon) - w^\star\|_2 \\
&= 2\|\Pi_{\mathcal{S}_\star}(z_t - w^\star) - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\Phi(z_t - w^\star) - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon\|_2 \\
&\leq 2\|\Pi_{\mathcal{S}_\star}(z_t - w^\star) - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\Phi(z_t - w^\star)\|_2 + 4\mu_t\|\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon\|_2 \\
&= 2\|\Pi_{\mathcal{S}_\star}(z_t - w^\star) - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\Phi I(z_t - w^\star)\|_2 + 4\mu_t\|\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon\|_2.
\end{aligned} \tag{11}$$

Expanding the identity matrix by $I = \Pi_{\mathcal{S}_\star} + \Pi_{\mathcal{S}_\star{}^c}$, we have

$$\begin{aligned}
(11) \leq\ & \underbrace{2\|(I - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\Phi\Pi_{\mathcal{S}_\star})\Pi_{\mathcal{S}_\star}(z_t - w^\star)\|_2}_{A} \\
&+ \underbrace{4\mu_t\|\Pi_{\mathcal{S}_\star}\Phi^\top\Phi\Pi_{\mathcal{S}_\star{}^c}(z_t - w^\star)\|_2}_{B} \\
&+ \underbrace{4\mu_t\|\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon}_{C}\|_2.
\end{aligned}$$

Now we bound the three terms respectively.

Noting that $|\mathcal{S}_\star| \leq 2k$, according to Lemma 1, in the subspace with support $\mathcal{S}_\star$, *i.e.*, $\{w \mid \text{supp}(w) = \mathcal{S}_\star\}$, the eigenvalues $\alpha_{2k} \leq \lambda_{\mathcal{S}_\star}(\Pi_{\mathcal{S}_\star}\Phi^\top\Phi\Pi_{\mathcal{S}_\star}) \leq \beta_{2k}$. Therefore, eigenvalues

$$\lambda_{\mathcal{S}_\star}(I - 2\mu_t\Pi_{\mathcal{S}_\star}\Phi^\top\Phi\Pi_{\mathcal{S}_\star}) \in [1 - 2\mu_t\beta_{2k}, 1 - 2\mu_t\alpha_{2k}],$$

which means

$$\begin{aligned} A &\leq 2\max\{2\mu_t\beta_{2k} - 1, 1 - 2\mu_t\alpha_{2k}\}\|\Pi_{\mathcal{S}_\star}(z_t - w^\star)\|_2 \\ &\leq 2\max\{\beta_{2k}/\alpha_{3k} - 1, 1 - \alpha_{2k}/\beta_{3k}\}\|z_t - w^\star\|_2. \end{aligned}$$

For term B, demoting $\mathcal{S}' = \text{supp}(z_t) \cup \text{supp}(w^\star)$, it can be observed that

$$B = 4\mu_t\|\Pi_{\mathcal{S}_\star}\Phi^\top\Phi\Pi_{\mathcal{S}_\star{}^c}\Pi_{\mathcal{S}'}(z_t - w^\star)\|_2.$$

Noting that $|\mathcal{S}' \cup \mathcal{S}_\star| \leq 4k$, by directly applying Lemma 2 we have

$$\begin{aligned} B &\leq 4\mu_t\frac{\beta_{4k} - \alpha_{4k}}{2}\|\Pi_{\mathcal{S}'}(z_t - w^\star)\|_2 \\ &\leq \frac{\beta_{4k} - \alpha_{4k}}{\alpha_{3k}}\|z_t - w^\star\|_2. \end{aligned}$$

To complete the proof, let us deal with the last piece. Similar to the techniques used in the proof on Lemma 1,

$$\begin{aligned} \|\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon\|_2 &= \max_{x\in\mathbb{R}^n:\|x\|_2=1}\langle\Pi_{\mathcal{S}_\star}\Phi^\top\epsilon, x\rangle \\ &= \max_{x\in\mathbb{R}^n:\|x\|_2=1}\epsilon^\top\Phi\Pi_{\mathcal{S}_\star}x \\ &= \max_{x\in\mathbb{R}^n:\|x\|_2=1}\langle\epsilon, \Phi\Pi_{\mathcal{S}_\star}x\rangle \\ &\leq \max_{x\in\mathbb{R}^n:\|x\|_2=1}\|\epsilon\|_2\cdot\|\Phi\Pi_{\mathcal{S}_\star}x\|_2 \\ &\leq \sqrt{\beta_{2k}}\|\epsilon\|_2, \end{aligned}$$

where the last inequality is done by directly applying the definition of RIP. Therefore,

$$C \leq 4\mu_t\sqrt{\beta_{2k}}\|\epsilon\|_2 \leq 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2.$$

Combining the 3 pieces together, we finally derive

$$\begin{aligned} \|w_{t+1} - w^\star\|_2 \leq\ & 2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\}\|z_t - w^\star\|_2 \\ &+ \frac{\beta_{4k} - \alpha_{4k}}{\alpha_{3k}}\|z_t - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2. \end{aligned}$$

Rearranging the inequality completes the proof.

□

C.5 Proof of Theorem 1

Theorem 1 (Restated)**.** *In the worst case scenario, with Assumption 1, the solutions path find by Automated Accelerated IHT (Algorithm 2) satisfy the following iterative invariant.*

$$\|w_{t+1} - w^\star\|_2 \leq \rho|1 + \tau_t|\cdot\|w_t - w^\star\|_2 + \rho|\tau_t|\cdot\|w_{t-1} - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

where $\rho = \left(2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\} + \frac{\beta_{4k} - \alpha_{4k}}{\alpha_{3k}}\right)$*, and* $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2$ *is the optimal objective value.*

Proof. Lemma 3 suggests

$$\|z_t - w^\star\|_2 \leq |1 + \tau_t|\|w_t - w^\star\|_2 + |\tau_t|\|w_{t-1} - w^\star\|_2.$$

Combining with lemma 4, *i.e.*,

$$\|w_{t+1} - w^\star\|_2 \leq \rho\|z_t - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

where $\rho = \left(2\max\{\frac{\beta_{2k}}{\alpha_{3k}} - 1, 1 - \frac{\alpha_{2k}}{\beta_{3k}}\} + \frac{\beta_{4k}-\alpha_{4k}}{\alpha_{3k}}\right)$, and $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2$, we have

$$\|x_t - w^\star\|_2 \leq \rho|1+\tau_t|\|w_t - w^\star\|_2 + \rho|\tau_t|\|w_{t-1} - w^\star\|_2 + 2\beta_{3k}\sqrt{\beta_{2k}}\|\epsilon\|_2,$$

which completes the proof. □

C.6 Proof of Corollary 1

Corollary 1 (Restated)**.** *Given the iterative invariant as stated in Theorem 1, and assuming the optimal solution achieves $\|\epsilon\|_2 = 0$, the solution found by Algorithm 2 satisfies:*

$$f(w_{t+1}) - f(w^\star) \leq \phi^t \left(\frac{\beta_{2k}}{\alpha_{2k}} f(w_1) + \frac{\rho\tau\beta_{2k}}{\phi\alpha_k} f(w_0)\right),$$

where $\phi = (\rho(1+\tau) + \sqrt{\rho^2(1+\tau)^2 + 4\rho\tau})/2$ and $\tau = \max_{i\in[t]}|\tau_i|$. It is sufficient to show linear convergence to the global optimum, when $\phi < 1$, or equivalently $\rho < 1/(1+2\tau)$.

Proof. Theorem 1 provides an upper bound invariant among consecutive iterates of the algorithm. To have better sense of convergence rate, we assume the optimal solution achieves $\|\epsilon\|_2 = 0$. Theorem 1 then implies

$$\begin{aligned}\|w_{t+1} - w^\star\|_2 &\leq \rho(1+|\tau_t|)\|w_t - w^\star\|_2 + \rho|\tau_t| \cdot \|w_{t-1} - w^\star\|_2 \\ &\leq \rho(1+\tau)\|w_t - w^\star\|_2 + \rho\tau \cdot \|w_{t-1} - w^\star\|_2.\end{aligned}$$

Rearranging the inequality with some regular algebraic manipulations, we have

$$\begin{aligned}\|w_{t+1} - w^\star\|_2 + \frac{\rho\tau}{\phi}\|w_t - w^\star\|_2 &\leq \phi\left(\|w_t - w^\star\|_2 + \frac{\rho\tau}{\phi}\|w_{t-1} - w^\star\|_2\right) \\ &\leq \phi^t\left(\|w_1 - w^\star\|_2 + \frac{\rho\tau}{\phi}\|w_0 - w^\star\|_2\right),\end{aligned}$$

where $\phi = \frac{\sqrt{\rho^2(1+\tau)^2+4\rho\tau}+\rho(1+\tau)}{2}$.

Noting that all ρ, τ, ϕ are non-negative, we can relax the inequality a bit to be

$$\|w_{t+1} - w^\star\|_2 \leq \phi^t\left(\|w_1 - w^\star\|_2 + \frac{\rho\tau}{\phi}\|w_0 - w^\star\|_2\right). \tag{12}$$

It is sufficient for linear convergence when $\phi < 1$, *i.e.*,

$$\begin{aligned}&\frac{\sqrt{\rho^2(1+\tau)^2 + 4\rho\tau} + \rho(1+\tau)}{2} < 1 \\ \iff\quad &\sqrt{\rho^2(1+\tau)^2 + 4\rho\tau} < 2 - \rho(1+\tau) \\ \iff\quad &\begin{cases}\rho^2(1+\tau)^2 + 4\rho\tau & < (2-\rho(1+\tau))^2 \\ 0 & < 2 - \rho(1+\tau)\end{cases} \\ \iff\quad &\begin{cases}\rho(1+2\tau) & < 1 \\ \rho(1+\tau) & < 2\end{cases} \\ \iff\quad &\rho < 1/(1+2\tau)\end{aligned}$$

In our case, this also indicates the linear convergence of function values. Noting that $(w_{t+1} - w^\star)$ and $(w_1 - w^\star)$ are at most $2k$-sparse, and $(w_0 - w^\star) = -w^\star$ is k-sparse, we have the following statements according to RIP

property:

$$\|\Phi(w_{t+1} - w^\star)\|_2^2 \leq \beta_{2k}\|w_{t+1} - w^\star\|_2^2$$
$$\|\Phi(w_1 - w^\star)\|_2^2 \geq \alpha_{2k}\|w_1 - w^\star\|_2^2$$
$$\|\Phi(w_0 - w^\star)\|_2^2 \geq \alpha_k\|w_0 - w^\star\|_2^2$$

As we assume $\|\epsilon\|_2 = \|y - \Phi w^\star\|_2 = 0$, *i.e.*, $y = \Phi w^\star$ and $f(w^\star)$, we can see that

$$f(w_{t+1}) = \|\Phi w_{t+1} - y\|_2^2 \leq \beta_{2k}\|w_{t+1} - w^\star\|_2^2$$
$$f(w_1) = \|\Phi w_1 - y\|_2^2 \geq \alpha_{2k}\|w_1 - w^\star\|_2^2$$
$$f(w_0) = \|\Phi w_0 - y\|_2^2 \geq \alpha_k\|w_0 - w^\star\|_2^2$$

Plugging these into (12) completes the proof. □

D Additional Related Work

Thresholding-based optimization algorithms have been attractive alternatives to relaxing the constraint to a convex one or to greedy selection. Bahmani et al. (2013) provide a gradient thresholding algorithm that generalizes pursuit approaches for compressed sensing to more general losses. Yuan et al. (2018) study convergence of gradient thresholding algorithms for general losses. Jain et al. (2014) consider several variants of thresholding-based algorithms for high dimensional sparse estimation. Nguyen et al. (2014); Li et al. (2016) discuss convergence properties of thresholding algorithms for stochastic settings, while in (Jain et al., 2016) the algorithm is extended to structured sparsity. Greedy algorithms (Shalev-Shwartz et al., 2010) for cardinality constrained problems have similar convergence guarantees and smaller per iteration cost but tend to underperform when compared to thresholding-based algorithms (Khanna & Kyrillidis, 2018).

Acceleration using momentum term (Beck & Teboulle, 2009; Ghadimi et al., 2015) allows for faster convergence of first-order methods without increasing the per iteration cost. In the context of accelerating sparsity constrained first-order optimization, Khanna & Kyrillidis (2018); Blumensath (2012) use momentum terms in conjunction with thresholding and prove linear convergence of their method. We extend their work by also including additional constraints of non-negativity. More recently, there have also been works (Ma et al., 2019) that study acceleration in sampling methods such as MCMC that are relevant to Bayesian coresets.

E Additional Results for Synthetic Gaussian Posterior Inference

Additional results for experiments in section 5.1 are provided in this section.

From an optimization perspective, one may be curious about the convergence speed of the two proposed algorithms, *i.e.*, A-IHT and Accelerated A-IHT II (Algorithm 2 & 3). The convergence for the two algorithms compared to the solutions by baselines are presented in Figure 5. The x-axis is iteration number for A-IHT and A-IHT II, and the y-axis is the objective function to be minimized, *i.e.*,

$$f(w) = \|y - \Phi w\|_2^2,$$

where $y = \sum_{i=1}^n \hat{g}_i$ and $\Phi = [\hat{g}_1, \ldots, \hat{g}_n]$.

The two IHT algorithms' fast convergence speed reflects what our theory suggests. They surpass GIGA within about 30 iterations, and surpass SparseVI within 50 iterations (A-IHT II) and within 100 iterations (A-IHT), respectively. Although we should note that the objective function which SparseVI minimizes is reverse KL divergence instead of l_2 distance, the two IHT algorithms can achieve much better solutions when considering KL divergence as well, as shown in Figure 1. Moreover, the tendency of a further decrease in objective value is still observed for the two IHT algorithms at 300^{th} iteration.

Illustration of the coresets constructed by A-IHT II in the first trial after projecting to 2D is presented in Figure 6.

F Additional Results for Radial Basis Regression

In this section, we provide additional experimental results of posterior contours for the radial basis regression experiment (section 5.2).

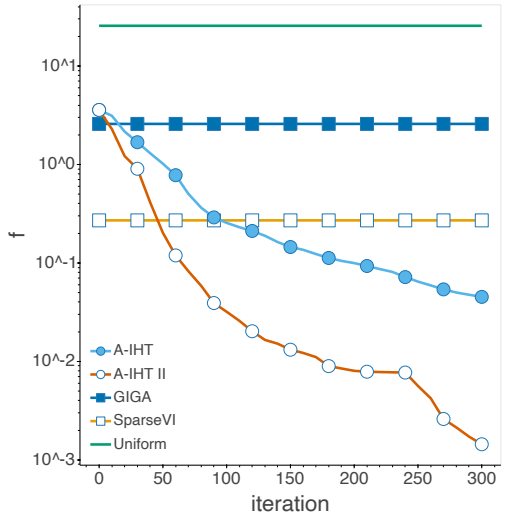


Figure 5: Convergence results for synthetic Gaussian posterior inference (subsection 5.1) when sparsity setting $k = 200$ in the first trial. For GIGA, SparseVI and Uniform, each of the objective function values f is calculated by the final output of each algorithms.

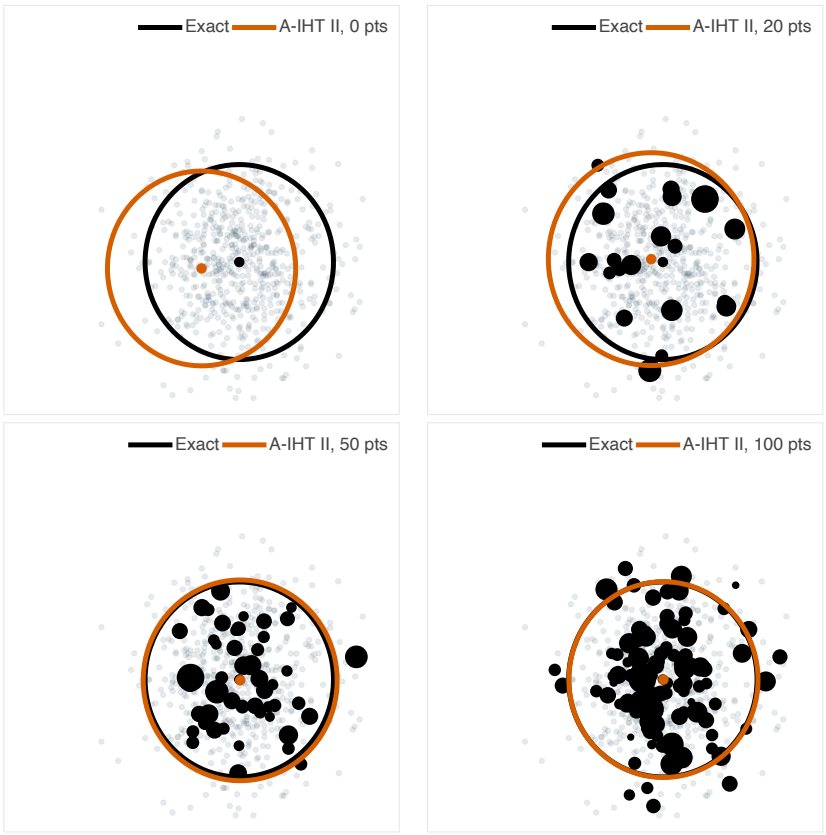


Figure 6: Illustration of true posterior and posterior constructed by A-IHT II after projecting to 2-dimensional plane for synthetic Gaussian posterior inference (Section 5.1). Results at different sparsity level are shown. The ellipses indicate 2σ-prediction of the posterior distribution, and the black dots represent coreset points selected with their radius denoting the respective weights.

We plot the posterior contours for both the true posterior and coreset posterior when sparsity level $k = 300$ in the first four random trials out of ten trials. The coreset posterior constructed by our Algorithm 3 recovers the

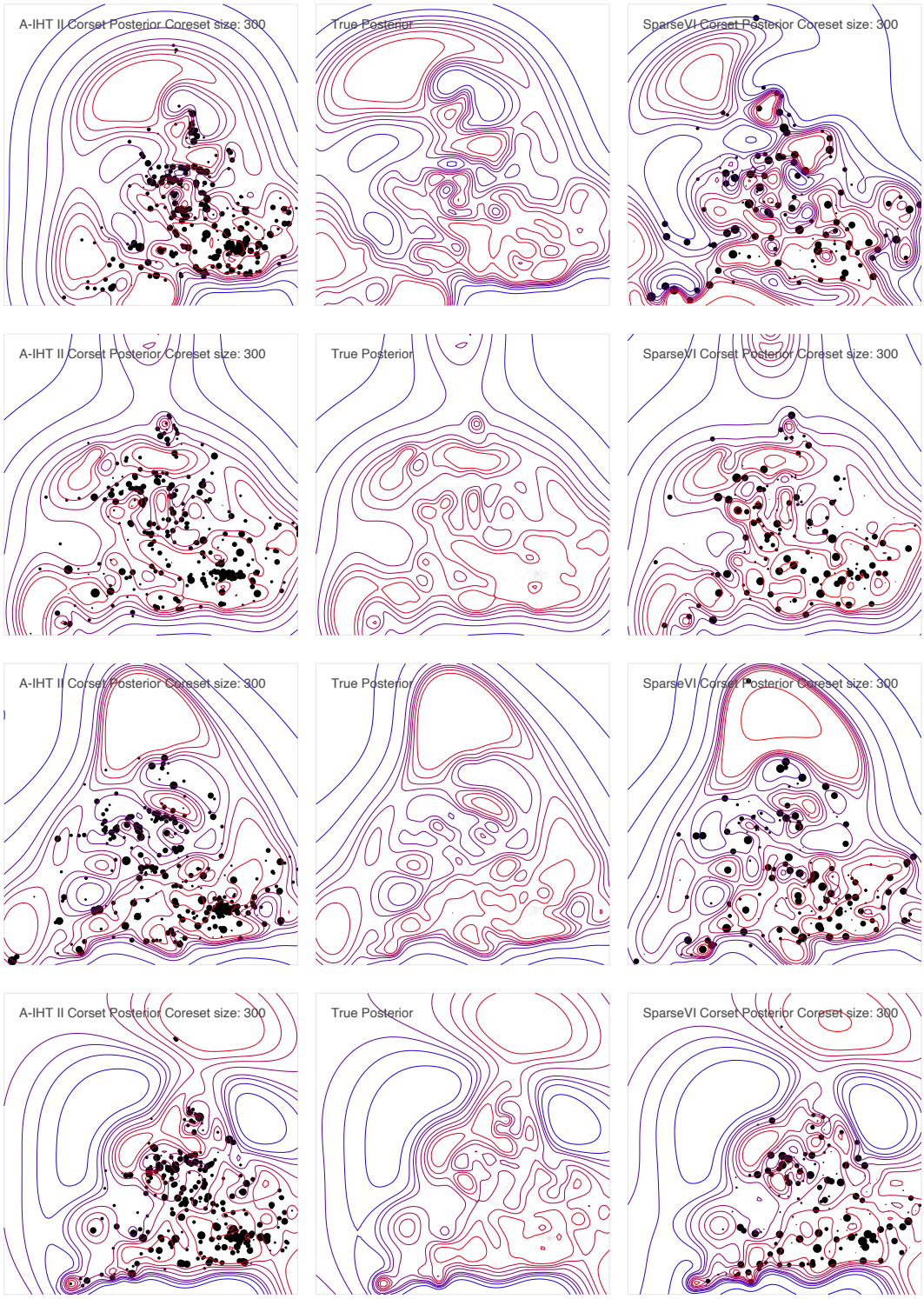
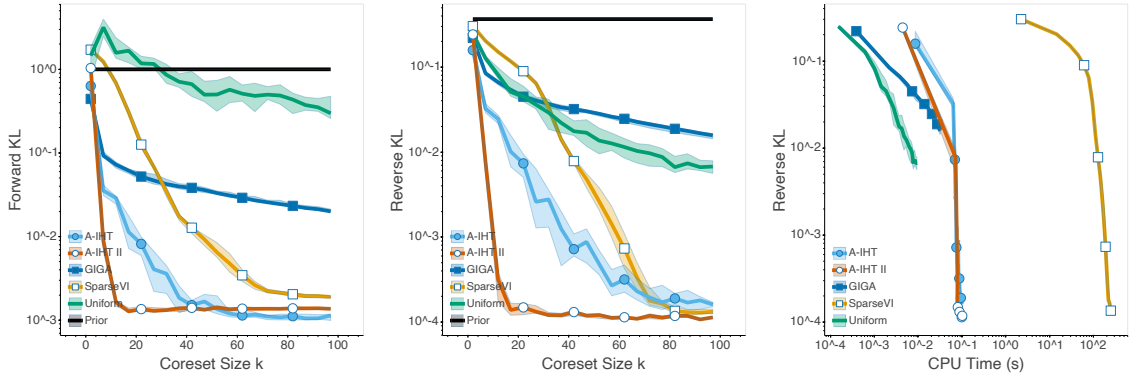
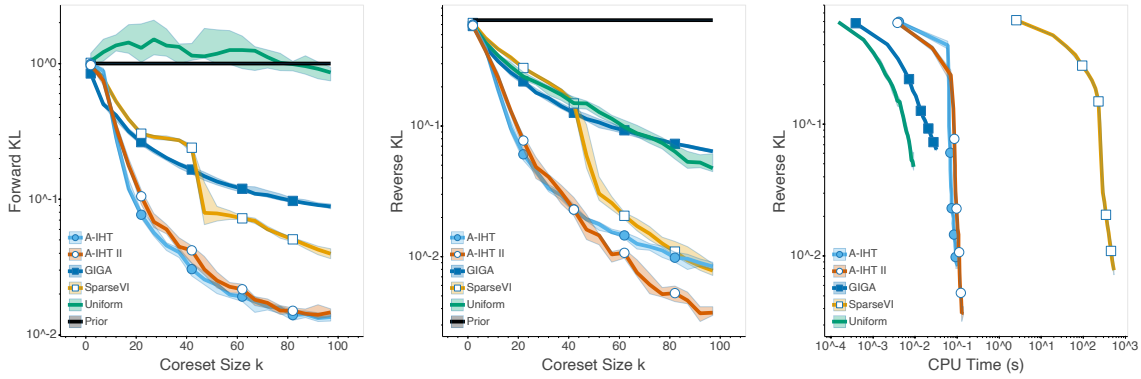


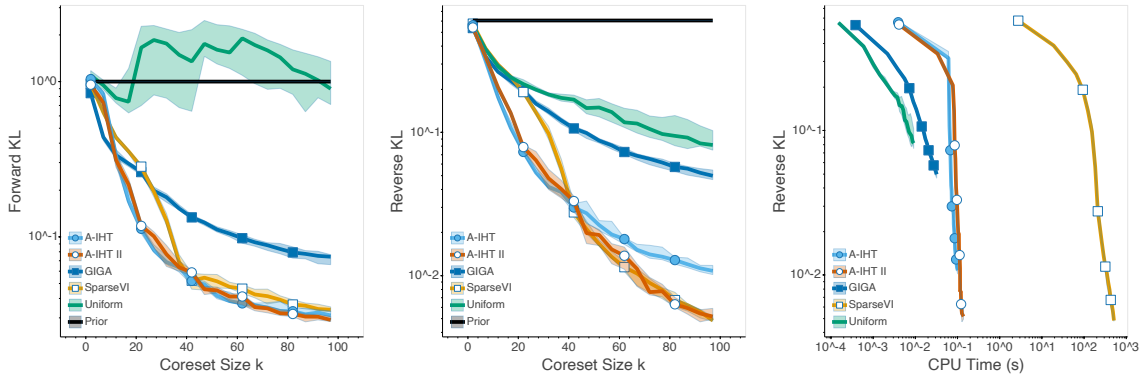
Figure 7: Experiments on Bayesian radial basis function regression in the first four random trials out of ten trails, where coreset sparsity setting $k = 300$. Coreset points are presented as black dots, with their radius indicating assigned weights. Posterior constructed by Accelerated IHT II (left) shows almost exact contours as the true posterior distribution (middle), while posterior constructed by SparseVI (right) shows deviated contours from the true posterior distribution.



(a) `synthetic dataset for logistic regression`



(b) `phishing dataset` for logistic regression



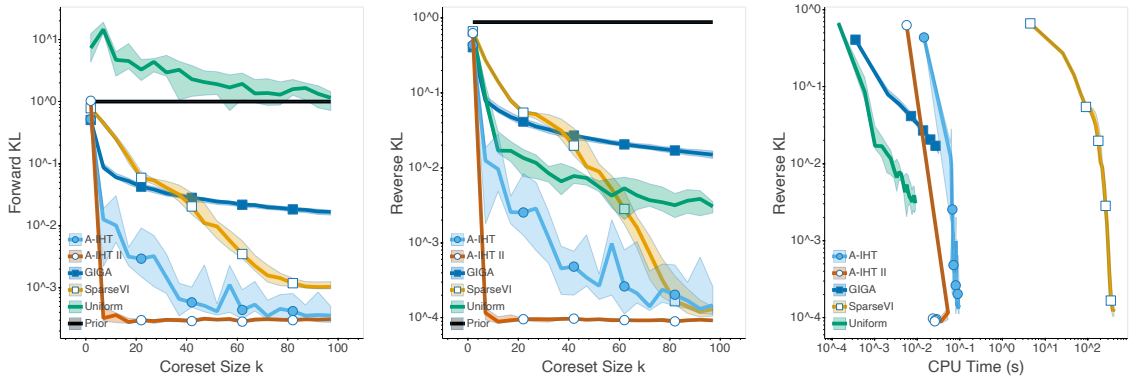
(c) `chemical reactivities dataset` for logistic regression

Figure 8: Bayesian coreset construction for logistic regression (LR) using the three different datasets. All the algorithms are run 20 times, and the median as well as the interval of 35^{th} and 65^{th} percentile, indicated as the shaded area, are reported. Different maximal coreset size k is tested from 1 to 100. Forward KL (left) and reverse KL (middle) divergence between estimated true posterior and coreset posterior indicate the quality of the constructed coreset. The smaller the KL divergence, the better the coreset is. The running time for each algorithm is also recorded (right).

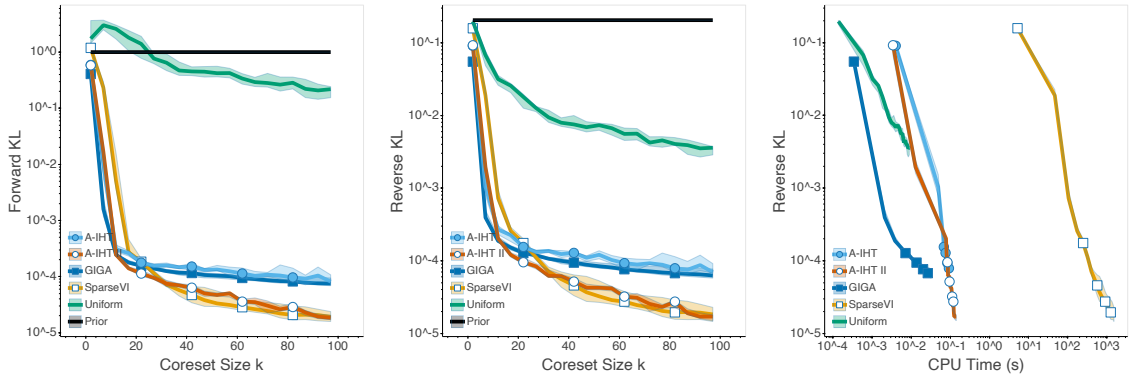
true posterior almost exactly, unlike SparseVI. Results are shown in Figure 7.

G Details and Extensive Results of the Bayesian Logistic and Poisson Regression Experiments

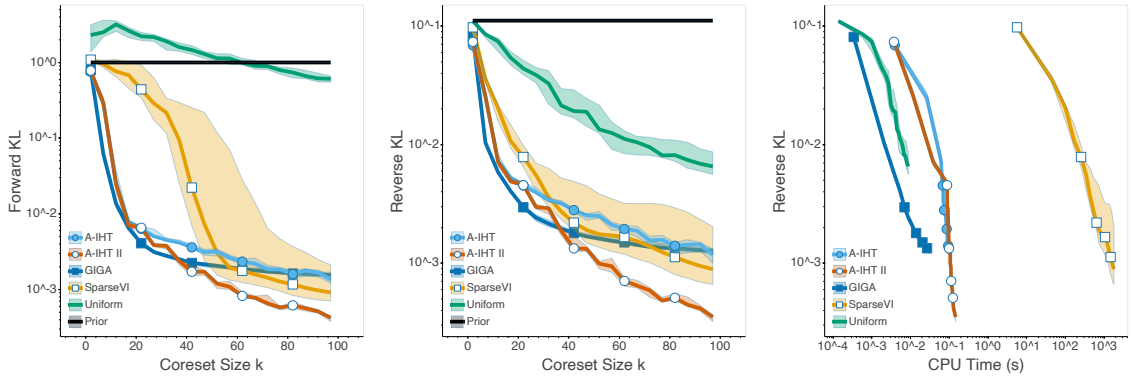
We consider how IHT performs when used in real applications where the closed-form expressions are unattainable. As the true posterior is unknown, a Laplace approximation is used for GIGA and IHT to derive the finite



(a) `synthetic dataset for Poisson regression`



(b) `biketrips dataset` for Poisson regression



(c) `airportdelays dataset` for Poisson regression

Figure 9: Bayesian coreset construction for Poisson regression (PR) using the three different datasets. All the algorithms are run 20 times, and the median as well as the interval of 35^{th} and 65^{th} percentile, indicated as the shaded area, are reported. Different maximal coreset size k is tested from 1 to 100. Forward KL (left) and reverse KL (middle) divergence between estimated true posterior and coreset posterior indicate the quality of the constructed coreset. The smaller the KL divergence, the better the coreset is. The running time for each algorithms is also recorded (right).

projection of the distribution, *i.e.*, $\hat{g}_i$. Further, Monte Carlo sampling is needed to derive gradients of D_{KL} for SparseVI. We compare different algorithms estimating the posterior distribution for logistic regression and Poisson regression. The reverse KL and forward KL between the coreset posterior and true posterior are estimated using another Laplace approximation. The experiment was proposed by Campbell & Broderick (2019), and is used in (Campbell & Broderick, 2018) (GIGA) and (Campbell & Beronov, 2019) (SparseVI). The experimental settings for each baseline algorithms are set following their original settings for this experiment. In addition, we conduct additional experiments using a stochastic gradient estimator or using an alternative evaluation for coreset quality.

For logistic regression, given a dataset $\{(x_n, y_n) \in \mathbb{R}^D \times \{1, -1\} \mid i \in [N]\}$, we aim to infer $\theta \in \mathbb{R}^{D+1}$ based on the model:

$$y_n \mid x_n, \theta \sim \text{Bern}\left(\frac{1}{1 + e^{-z_n^\top \theta}}\right),$$

where $z_n = [x_n^\top, 1]^\top$. Three datasets are used for logistic regression. The `synthetic dataset for logistic regression` consists of data x_n sampled i.i.d. from standard normal distribution $\mathcal{N}(0, I)$, and label y_n sampled from Bernoulli distribution conditioned on x_n and $\theta = [3, 3, 0]^\top$. The original `phishing` dataset[4] consists of $N = 11055$ data points with dimension $D = 68$. The `phishing` dataset used in this experiment is preprocessed (Campbell & Beronov, 2019) via principle component analysis to project each data points to dimension of $D = 10$ to mitigate high computation by SparseVI. The original `chemical reactivities` dataset[5] has $N = 26733$ data points with dimension $D = 10$. We uniformly sub-sample $N = 500$ data points from each datasets for this experiment, due to the high computation cost of SparseVI.

For Poisson regression, given $\{(x_n, y_n) \in \mathbb{R}^D \times \mathbb{N} \mid i \in [N]\}$, we aim to infer $\theta \in \mathbb{R}^{D+1}$ from model

$$y_n \mid x_n, \theta \sim \text{Poiss}\left(\log\left(1 + e^{-z_n^\top \theta}\right)\right),$$

where $z_n = [x_n^\top, 1]^\top$. Three other datasets are used for Poisson regression: the `synthetic dataset for Poisson regression` consists of data x_n sampled i.i.d. from a standard normal distribution $\mathcal{N}(0, 1)$, and target y_n sampled from Poisson distribution conditioned on x_n and $\theta = [1, 0]^\top$. The `biketrips` dataset[6] consists of $N = 17386$ data points with dimension $D = 8$. The `airportdelays` dataset[7] has $N = 7580$ data points with dimension $D = 15$. Same as logistic regression, we uniformly sub-sample $N = 500$ data points from each datasets for this experiment.

The comparison of the algorithms for Bayesian coreset construction for logistic regression are shown in Figure 8, and Bayesian coreset construction for Poisson regression are shown in Figure 9. The left column shows forward KL divergence given sparsity setting k, the middle column shows reverse KL divergence, and the right column presents the running time for corset construction for each algorithm.

It is observed that A-IHT and A-IHT II achieve state-of-the-art performance. The IHT algorithms often obtain coresets with smaller KL than GIGA and SparseVI, with computing time comparable to GIGA, significantly less than SparseVI. The experiments indicate that IHT outperforms the previous methods, improving the trade-off between accuracy and performance.

The results on large-scale datasets have been presented in the Figure 4 in the main paper. Next, we present two additional sets of experiments that are omitted in the main paper.

Stochastic Gradient Estimator. For large-scale datasets, it is often necessary to "batch" the algorithms. IHT can be easily batched by replacing the gradient with a stochastic gradient estimator that only a batch of data in each iteration.

Recall that for IHT the gradient of the objective function $f(w) = \|y - \Phi w\|^2$ is $\nabla f(w) = 2\Phi^\top(\Phi w - y)$, where $\Phi \in \mathbb{R}^{S \times n}$. As we introduced in section 2, S is the number of samples $\theta \sim \hat{\pi}$, and n is the number of data. Thus, we can form a unbiased gradient estimator as

$$\tilde{g}(w) = 2\Gamma_1^\top \Phi^\top(\Phi\Gamma_2 w - y),$$

where $\Gamma_1, \Gamma_2 \in \mathbb{R}^{n \times n}$ are *i.i.d* sampled from a distribution π_Γ with $\mathbb{E}_{\Gamma_1 \sim \pi_\Gamma}[\Gamma_1] = \mathbb{E}_{\Gamma_2 \sim \pi_\Gamma}[\Gamma_2] = I$, where $I \in \mathbb{R}^{n \times n}$ is the identity matrix. Therefore,

$$\mathbb{E}\tilde{g}(w) = 2(\mathbb{E}\Gamma_1)^\top \Phi^\top(\Phi \mathbb{E}\Gamma_2 w - y) = 2I\Phi^\top(\Phi I w - y) = \nabla f(w),$$

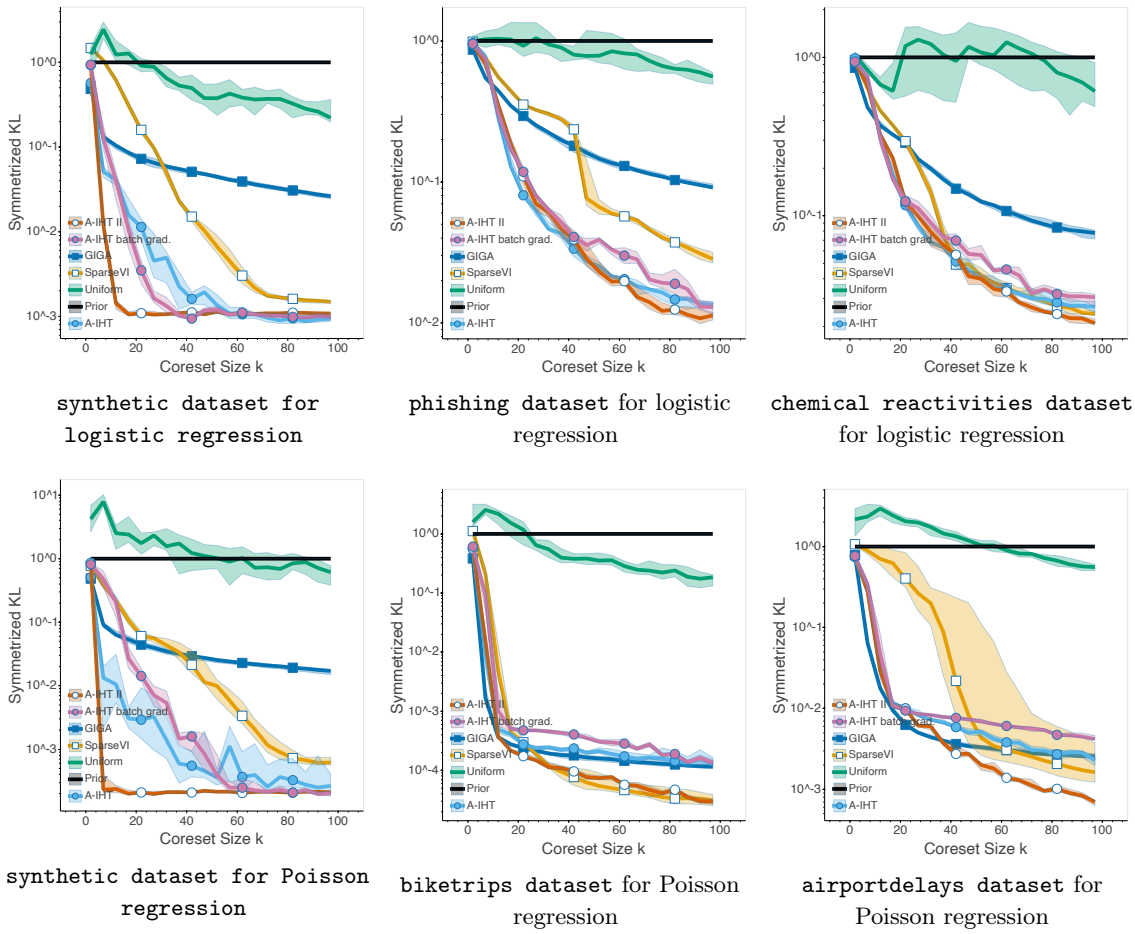
showing that $\tilde{g}(w)$ is an unbiased estimator of $\nabla f(w)$.

[4] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`

[5] `http://komarix.org/ac/ds`

[6] `http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`

[7] The `airportdelays` dataset was constructed (Campbell & Broderick, 2019) by combining flight delay data (`http://stat-computing.org/dataexpo/2009/the-data.html`) and weather data (`https://www.wunderground.com/history/.`).



`synthetic dataset for logistic regression` | `phishing dataset` for logistic regression | `chemical reactivities dataset` for logistic regression

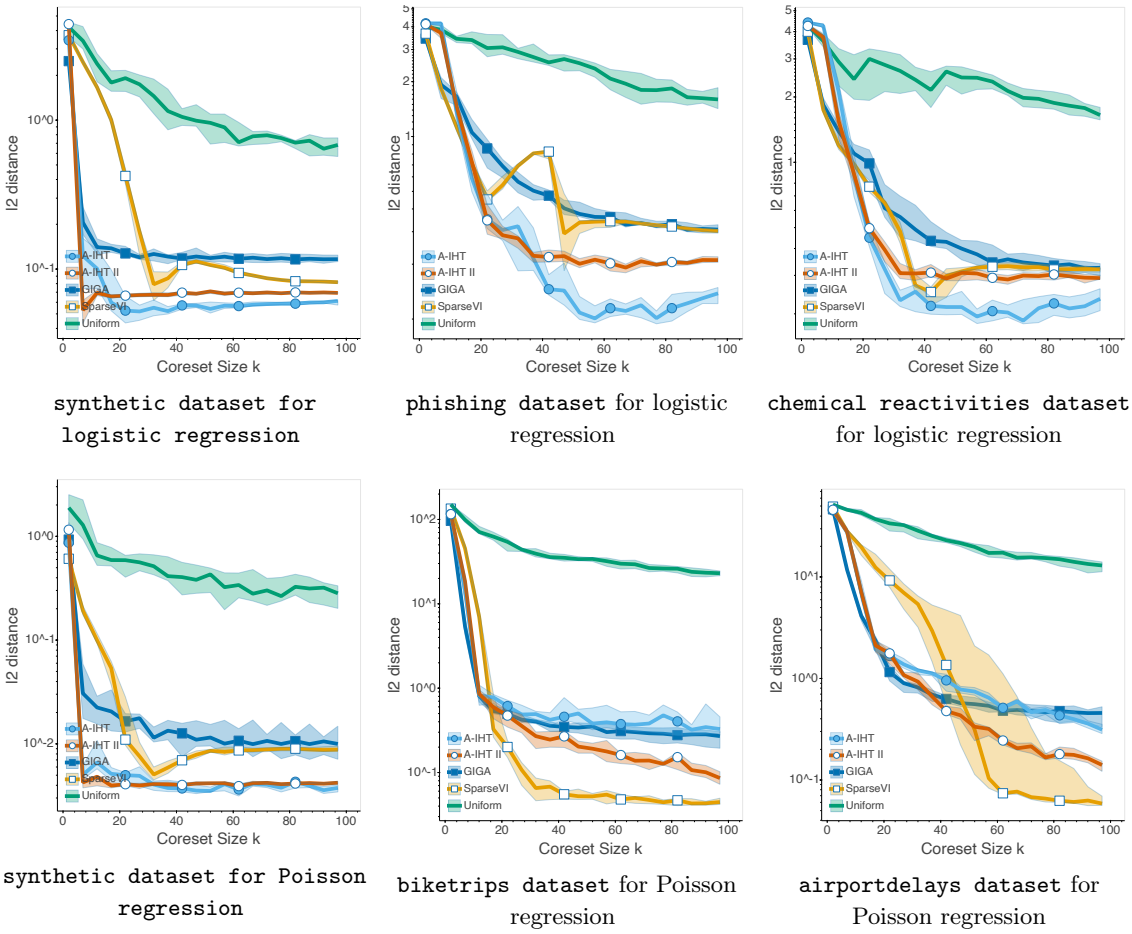
`synthetic dataset for Poisson regression` | `biketrips dataset` for Poisson regression | `airportdelays dataset` for Poisson regression

Figure 10: Bayesian coreset construction for logistic regression and Poisson regression using the six different datasets. All the algorithms are run 20 times, and the median as well as the interval of 35^{th} and 65^{th} percentile, indicated as the shaded area, are reported. Different maximal coreset size k is tested from 1 to 100. Symmetrized KL divergence between estimated true posterior and coreset posterior indicate the quality of the constructed coreset. The smaller the KL divergence, the better the coreset is.

For example, we can form the estimator using a batch of data with batch size B by letting Γ_1, Γ_2 be random matrices as randomly setting $n - B$ rows of $\frac{n}{B} I$ be zero. Equivalently, it is the same as randomly picking B columns of Φ, setting the rest columns be zero, and scale the matrix by n/B. Noting that each column of Φ corresponds to each of the n data points, this operation is essentially to approximate Φ using a batch of data with batch size B, and thus it approximates the gradient using a batch of a data.

We test how Algorithm 2 performs on the Bayesian logistic regression and Poisson regression using the stochastic estimator with batch size $B = n/5$. All of the experimental settings are the same as what we have introduced in this section. As a summary of both forward FL and reverse KL, we use the symmetrized KL (*i.e.*, the sum of forward KL and reverse KL) as the evaluation metric for coreset quality. The results are shown in Figure 10. It is observed that A-IHT with the stochastic gradient estimator (A-IHT batch grad.) performs comparably to the A-IHT. We note that the batched version of A-IHT can be improved by increasing its maximal number of iterations, *i.e.*, optimization with stochastic gradient needs more iterations to converge, or using a better batch gradient estimator. Theoretical study on accelerated IHT with approximated gradients is still an open question to the best of our knowledge. Further research on accelerated IHT with stochastic gradients is an interesting future work.

ℓ_2-distance Evaluation of Coreset Quality. In the previous experiments in the subsection, the coreset quality is evaluated by approximating the KL divergence between the full-dataset posterior and coreset posterior. As an alternative way to measure the coreset quality, we measure the ℓ_2-distance between the maximum-a-posteriori



`synthetic dataset for logistic regression`

`phishing dataset` for logistic regression

`chemical reactivities dataset` for logistic regression

`synthetic dataset for Poisson regression`

`biketrips dataset` for Poisson regression

`airportdelays dataset` for Poisson regression

Figure 11: Bayesian coreset construction for logistic regression and Poisson regression using the six different datasets. All the algorithms are run 20 times, and the median as well as the interval of 35^{th} and 65^{th} percentile, indicated as the shaded area, are reported. Different maximal coreset size k is tested from 1 to 100. ℓ_2-distance between the MAP estimators of the full-dataset posterior and coreset posterior indicate the quality of the constructed coreset. The smaller the ℓ_2-distance, the better the coreset is.

(MAP) estimation of the full-dataset posterior and coreset posterior. The results are shown in Figure 11. It is observed that the two IHT algorithms usually achieve the best results, except that SparseVI achieves the lowest ℓ_2-distance on two datasets. However, SparseVI costs $\times 10^4$ more time than IHT and GIGA.