A. Additional results

A.1. Orthonormal vectors

In this experiment, we generated a dataset of 5000 unit vectors in $\mathbb{R}^{5000}$, each aligned with one of the coordinate axes. This dataset is exactly that used in the proof of Proposition 2.1, except that the number of datapoints N is fixed to 5000. We constructed coresets for each of the datasets via uniformly random subsampling (RND), Frank–Wolfe (FW), and GIGA. We compared the algorithms on two metrics: reconstruction error, as measured by the 2-norm between $\mathcal{L}(w)$ and $\mathcal{L}$; and representation efficiency, as measured by the size of the coreset. Fig. 5 shows the results of the experiment, with reconstruction error in Fig. 5a and coreset size in Fig. 5b. As expected, for early iterations FW performs about as well as uniformly random subsampling, as these algorithms generate equivalent coresets (up to some reordering of the unit vectors) with high probability. FW only finds a good coreset after all 5000 points in the dataset have been added. These algorithms both do not correctly scale the coreset; in contrast, GIGA scales its coreset correctly, providing significant reduction in error.

A.2. Alternate inference algorithms

We reran the same experiment as described in Section 4.3, except we swapped the inference algorithm for random-walk Metropolis–Hastings (RWMH) and the No-U-Turn Sampler (NUTS) (Hoffman & Gelman, 2014). When using RWMH, we simulated a total of 50,000 steps: 25,000 warmup steps including covariance adaptation with a target acceptance rate of 0.234, and 25,000 sampling steps thinned by a factor of 5, yielding 5,000 posterior samples. If the acceptance rate for the latter 25,000 steps was not between 0.15 and 0.7, we reran the procedure. When using NUTS, we simulated a total of 6,000 steps: 1,000 warmup steps including leapfrog step size adaptation with a target acceptance rate of 0.8, and 5,000 sampling steps.

The results for these experiments are shown in Figs. 6 and 7, and generally corroborate the results from the experiments using Hamiltonian Monte Carlo in the main text. One difference when using NUTS is that the performance versus computation time appears to follow an "S"-shaped curve, which is caused by the dynamic path-length adaptation provided by NUTS. Consider the log-likelihood of logistic regression, which has a "nearly linear" region and a "nearly flat" region. When the coreset is small, there are directions in latent space that point along "nearly flat" regions; along these directions, u-turns happen only after long periods of travel. When the coreset reaches a certain size, these "nearly flat" directions are all removed, and u-turns happen more frequently. Thus we expect the computation time as a function of coreset size to initially increase smoothly, then drop quickly, followed by a final smooth increase, in agreement with Fig. 7b.

B. Technical Results and Proofs

Proof of Lemma 3.5. By setting $s_m = \frac{\|\mathcal{L}_m\|}{\|\mathcal{L}\|}$ for each $m \in [N]$ in Eq. (29), we have that $\tau \geq \frac{\|\mathcal{L}\|}{\sigma} > 0$. Now suppose $\epsilon \leq 0$; then there exists some conic combination d of $(d_{\infty m})_{m=1}^N$ for which $\|d\| = 1$, $\langle d, \ell \rangle = 0$, and $\forall m \in [N]$, $\langle -d, d_{\infty m} \rangle \leq 0$. There must exist at least one index $n \in [N]$ for which $\langle -d, d_{\infty n} \rangle < 0$, since otherwise d is not in the linear span of $(d_{\infty m})_{m=1}^N$. This also implies $\|d_{\infty n}\| > 0$ and hence $\|\ell_n - \langle \ell_n, \ell \rangle \ell\| > 0$. Then $\left\langle -d, \sum_{m=1}^N \xi_m d_{\infty m} \right\rangle < 0$ for any $\xi \in \Delta^{N-1}$ with $\xi_n > 0$. But setting $\xi_n \propto \sigma_n \|\ell_n - \langle \ell_n, \ell \rangle \ell\|$ results in $\sum_{n=1}^N \xi_n d_{\infty n} = 0$, and we have a contradiction. □

Proof of Lemma 3.6. We begin with the $\tau\sqrt{J_t}$ bound. For any $\xi \in \Delta^{N-1}$,

$$\langle d_t, d_{tn_t} \rangle = \max_{n \in [N]} \langle d_t, d_{tn} \rangle \geq \sum_{n=1}^N \xi_n \langle d_t, d_{tn} \rangle . \tag{40}$$

Suppose that $\ell = \sum_{n=1}^N s_n \ell_n$ for some $s \in \mathbb{R}_+^N$. Setting $\xi_n \propto s_n \|\ell_n - \langle \ell_n, \ell(w_t) \rangle \ell(w_t)\|$ yields

$$\langle d_t, d_{tn_t} \rangle \geq C^{-1} \|\ell - \langle \ell, \ell(w_t) \rangle \ell(w_t)\| \tag{41}$$

$$C := \left(\sum_{n=1}^N s_n \|\ell_n - \langle \ell_n, \ell(w_t) \rangle \ell(w_t)\| \right) . \tag{42}$$

Noting that the norms satisfy $\|\ell - \langle \ell, \ell(w_t) \rangle \ell(w_t)\| = \sqrt{J_t}$ and $\|\ell_n - \langle \ell_n, \ell(w_t) \rangle \ell(w_t)\| \leq 1$, we have

$$\langle d_t, d_{tn_t} \rangle \geq \|s\|_1^{-1} \sqrt{J_t} . \tag{43}$$

Maximizing over all valid choices of s yields

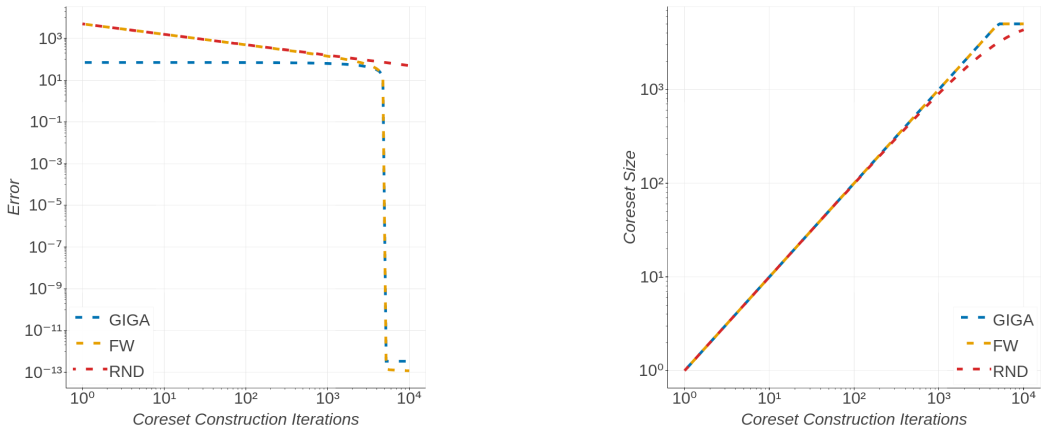
$$\langle d_t, d_{tn_t} \rangle \geq \tau \sqrt{J_t} . \tag{44}$$

Next, we develop the $f(J_t)$ bound. Note that

$$\sum_{n=1}^N w_{tn} \|\ell_n - \langle \ell_n, \ell \rangle \ell\| \, d_{\infty n} = \sum_{n=1}^N w_{tn} (\ell_n - \langle \ell_n, \ell \rangle \ell) \tag{45}$$

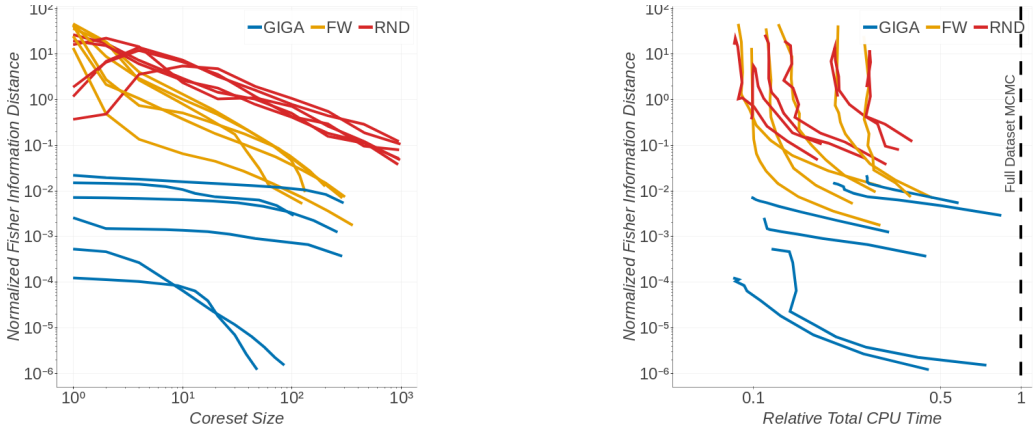
$$= \ell(w_t) - \langle \ell(w_t), \ell \rangle \ell, \tag{46}$$

so we can express $\ell(w_t) = \sqrt{J_t} d + \sqrt{1 - J_t} \ell$ and $d_t = \sqrt{J_t} \ell - \sqrt{1 - J_t} d$ for some vector d that is a conic combination of $(d_{\infty n})_{n=1}^N$ with $\|d\| = 1$ and $\langle d, \ell \rangle = 0$. Then by the definition of ϵ in Eq. (30) and Lemma 3.5, there exists



(a) (b)

Figure 5. Comparison of different coreset constructions on the synthetic axis-aligned vector dataset. Fig. 5a shows a comparison of 2-norm error between the coreset $\mathcal{L}(w)$ and the true sum $\mathcal{L}$ as a function of construction iterations. Fig. 5b shows a similar comparison of coreset size.



(a) (b)

Figure 6. Results for the experiment described in Section 4.3 with posterior inference via random-walk Metropolis–Hastings.

an $n \in [N]$ such that $\langle -d, d_{\infty n} \rangle \geq \epsilon > 0$. Therefore

$$\langle d_t, d_{tn_t} \rangle \tag{47}$$
$$\geq \langle d_t, d_{tn} \rangle \tag{48}$$
$$= \left\langle \sqrt{J_t}\,\ell - \sqrt{1-J_t}\,d, \frac{\ell_n - \langle \ell_n, \ell(w_t)\rangle\,\ell(w_t)}{\|\ell_n - \langle \ell_n, \ell(w_t)\rangle\,\ell(w_t)\|} \right\rangle \tag{49}$$
$$= \frac{\sqrt{1-J_t}\,\langle -d, \ell_n\rangle + \sqrt{J_t}\,\langle \ell, \ell_n\rangle}{\sqrt{1-\left(\sqrt{1-J_t}\,\langle \ell_n, \ell\rangle + \sqrt{J_t}\,\langle \ell_n, d\rangle\right)^2}} \tag{50}$$
$$= \frac{\sqrt{1-J_t}\sqrt{1-\langle \ell_n, \ell\rangle^2}\,\langle -d, d_{\infty n}\rangle + \sqrt{J_t}\,\langle \ell, \ell_n\rangle}{\sqrt{1-\left(\sqrt{1-J_t}\,\langle \ell_n, \ell\rangle + \sqrt{J_t}\sqrt{1-\langle \ell_n, \ell\rangle^2}\,\langle d, d_{\infty n}\rangle\right)^2}}. \tag{51}$$

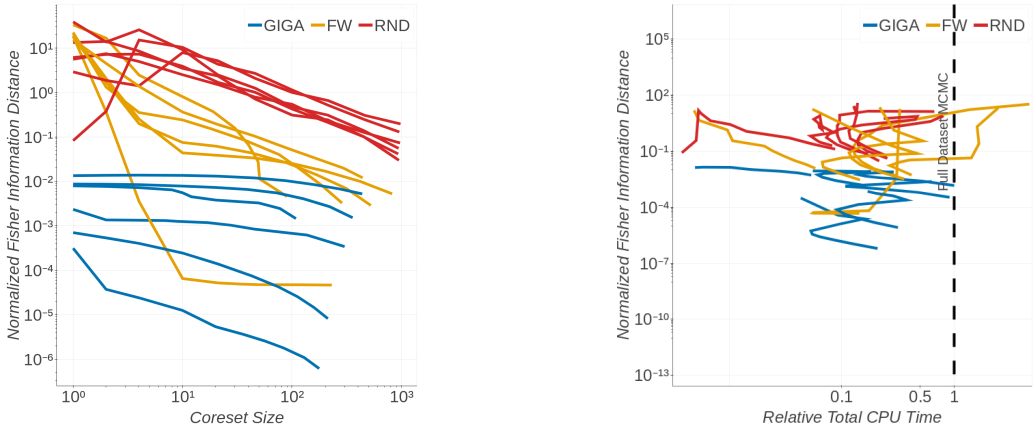
We view this bound as a function of two variables $\langle \ell, \ell_n \rangle$ and $\langle -d, d_{\infty n} \rangle$, and we view the worst-case bound as the minimization over these variables. We further lower-bound by removing the coupling between them. Fixing $\langle -d, d_{\infty n} \rangle$, the derivative in $\langle \ell, \ell_n \rangle$ is always nonnegative, and note that $\langle \ell_n, \ell \rangle > -1$ since otherwise $\langle -d, d_{\infty n} \rangle = 0$ by the remark after Eq. (18), so setting

$$\beta = 0 \wedge \left(\min_{n \in [N]} \langle \ell, \ell_n \rangle \;\text{ s.t. }\; \langle \ell, \ell_n \rangle > -1 \right), \tag{52}$$

we have

$$\langle d_t, d_{tn_t} \rangle \geq \tag{53}$$
$$\frac{\sqrt{1-J_t}\sqrt{1-\beta^2}\,\langle -d, d_{\infty n}\rangle + \sqrt{J_t}\,\beta}{\sqrt{1-\left(\sqrt{1-J_t}\,\beta + \sqrt{J_t}\sqrt{1-\beta^2}\,\langle d, d_{\infty n}\rangle\right)^2}}. \tag{54}$$

We add $\{0\}$ into the minimization since $\beta \leq 0$ guarantees that the derivative of the above with respect to J_t is nonpos-



(a) (b)

Figure 7. Results for the experiment described in Section 4.3 with posterior inference via NUTS.

itive (which we will require in proving the main theorem). For all J_t small enough such that $\sqrt{1-J_t}\sqrt{1-\beta^2}\epsilon + \sqrt{J_t}\beta \geq 0$, the derivative of the above with respect to $\langle -d, d_{\infty n}\rangle$ is nonnegative. Therefore, minimizing yields

$$\langle d_t, d_{tn_t}\rangle \geq \frac{\sqrt{1-J_t}\sqrt{1-\beta^2}\epsilon + \sqrt{J_t}\beta}{\sqrt{1-\left(\sqrt{1-J_t}\beta - \sqrt{J_t}\sqrt{1-\beta^2}\epsilon\right)^2}}. \tag{55}$$

which holds for any such small enough J_t. But note that we've already proven the $\langle d_t, d_{tn_t}\rangle \geq \tau\sqrt{J_t}$ bound, which is always nonnegative; so the only time the current bound is "active" is when it is itself nonnegative, i.e. when J_t is small enough. Therefore the bound

$$\langle d_t, d_{tn_t}\rangle \geq \tau\sqrt{J_t} \vee \frac{\sqrt{1-J_t}\sqrt{1-\beta^2}\epsilon + \sqrt{J_t}\beta}{\sqrt{1-\left(\sqrt{1-J_t}\beta - \sqrt{J_t}\sqrt{1-\beta^2}\epsilon\right)^2}} \tag{56}$$

holds for all $J_t \in [0,1]$. □

C. Cap-tree Search

When choosing the next point to add to the coreset, we need to solve the following maximization with $O(N)$ complexity:

$$n_t = \arg\max_{n\in[N]} \frac{\langle \ell_n, \ell - \langle \ell, \ell(w_t)\rangle \ell(w_t)\rangle}{\sqrt{1-\langle \ell_n, \ell(w_t)\rangle^2}}. \tag{57}$$

One option to potentially reduce this complexity is to first partition the data in a tree structure, and use the tree structure for faster search. However, we found that in practice (1) the cost of constructing the tree structure outlined below outweighs the benefit of faster search later on, and (2) the computational gains diminish significantly with high-dimensional vectors ℓ_n. We include the details of our proposed cap-tree below, and leave more efficient construction and search as an open problem for future work.

Each node in the tree is a spherical "cap" on the surface of the unit sphere, defined by a central direction ξ, $\|\xi\| = 1$ and a dot-product bound $r \in [-1, 1]$, with the property that all data in child leaves of that node satisfy $\langle \ell_n, \xi\rangle \geq r$. Then we can upper/lower bound the search objective for such data given ξ and r. If we progress down the tree, keeping track of the best lower bound, we may be able to prune large quantities of data if the upper bound of any node is less than the current best lower bound.

For the lower bound, we evaluate the objective at the vector ℓ_n closest to ξ. For the upper bound, define $u := \frac{\ell - \langle \ell, \ell(w_t)\rangle \ell(w_t)}{\|\ell - \langle \ell, \ell(w_t)\rangle \ell(w_t)\|}$, and $v := \ell(w_t)$. Then $\|u\| = \|v\| = 1$ and $\langle u, v\rangle = 0$. The upper bound is

$$\max_{\zeta} \frac{\langle \zeta, u\rangle}{\sqrt{1-\langle \zeta, v\rangle^2}} \quad \text{s.t.} \quad \langle \zeta, \xi\rangle \geq r \quad \|\zeta\| = 1. \tag{58}$$

If we write $\zeta = \alpha_u u + \alpha_v v + \sum_i \alpha_i z_i$ where z_i completes the basis of u, v etc, and $\xi = \beta_u u + \beta_v v + \sum_i \beta_i z_i$,

$$\begin{aligned} \max_{\alpha\in\mathbb{R}^d} \quad & \frac{\alpha_u}{\sqrt{1-\alpha_v^2}} \\ \text{s.t.} \quad & \alpha_u\beta_u + \alpha_v\beta_v + \sum_i \alpha_i\beta_i \geq r \\ & \alpha_u^2 + \alpha_v^2 + \sum_i \alpha_i^2 = 1. \end{aligned} \tag{59}$$

Noting that α_i doesn't appear in the objective, we maximize

$\sum_i \alpha_i \beta_i$ to find the equivalent optimization

$$\max_{\alpha_u, \alpha_v} \quad \frac{\alpha_u}{\sqrt{1-\alpha_v^2}} \tag{60}$$

$$\text{s.t.} \quad \alpha_u \beta_u + \alpha_v |\beta_v| + \|\beta\| \sqrt{1-\alpha_u^2-\alpha_v^2} \geq r \tag{61}$$

$$\alpha_u^2 + \alpha_v^2 \leq 1, \tag{62}$$

where the norm on $|\beta_v|$ comes from the fact that we can choose the sign of α_v arbitrarily, ensuring the optimum has $\alpha_v \geq 0$. Now define

$$\gamma := \frac{\alpha_u}{\sqrt{1-\alpha_v^2}} \quad \eta := \frac{1}{\sqrt{1-\alpha_v^2}}, \tag{63}$$

so that the optimization becomes

$$\max_{\gamma, \eta} \quad \gamma \tag{64}$$

$$\text{s.t.} \quad \gamma \beta_u + |\beta_v| \sqrt{\eta^2 - 1} + \|\beta\| \sqrt{1-\gamma^2} \geq r\eta \tag{65}$$

$$\gamma^2 \leq 1, \, \eta \geq 1. \tag{66}$$

Since η is now decoupled from the optimization, we can solve

$$\max_{\eta \geq 1} |\beta_v| \sqrt{\eta^2 - 1} - r\eta \tag{67}$$

to make the feasible region in γ as large as possible. If $|\beta_v| > r$, we maximize Eq. (67) by sending $\eta \to \infty$ yielding a maximum of 1 in the original optimization. Otherwise, note that at $\eta = 1$ the derivative of the objective is $+\infty$, so we know the constraint $\eta = 1$ is not active. Therefore, taking the derivative and setting it to 0 yields

$$0 = \frac{|\beta_v| \eta}{\sqrt{\eta^2 - 1}} - r \tag{68}$$

$$\eta = \sqrt{\frac{r^2}{r^2 - |\beta_v|^2}}. \tag{69}$$

Substituting back into the original optimization,

$$\max_{\gamma} \quad \gamma \tag{70}$$

$$\text{s.t.} \quad \gamma \beta_u + \|\beta\| \sqrt{1-\gamma^2} \geq \sqrt{r^2 - |\beta_v|^2} \tag{71}$$

$$\gamma^2 \leq 1. \tag{72}$$

If $\beta_u \geq \sqrt{r^2 - |\beta_v|^2}$, then $\gamma = 1$ is feasible and the optimum is 1. Otherwise, note that at $\gamma = -1$, the derivative of the constraint is $+\infty$ and the derivative of the objective is 1, so the constraint $\gamma = -1$ is not active. Therefore, we can solve the unconstrained optimization by taking the derivative and setting to 0, yielding

$$\gamma = \frac{\beta_u \sqrt{r^2 - \beta_v^2} + \|\beta\| \sqrt{1-r^2}}{\|\beta\|^2 + \beta_u^2}. \tag{73}$$

Therefore, the upper bound is as follows:

$$U = \begin{cases} 1 & |\beta_v| > r \\ 1 & \beta_u \geq \sqrt{r^2 - \beta_v^2} \\ \frac{\beta_u \sqrt{r^2 - \beta_v^2} + \|\beta\| \sqrt{1-r^2}}{\|\beta\|^2 + \beta_u^2} & \text{else.} \end{cases} \tag{74}$$

D. Datasets

The `Phishing` dataset is available online at https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html. The `DS1` dataset is available online at http://komarix.org/ac/ds/. The `BikeTrips` dataset is available online at http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset. The `AirportDelays` dataset was constructed using flight delay data from http://stat-computing.org/dataexpo/2009/the-data.html and historical weather information from https://www.wunderground.com/history/.