Coresets for Near-Convex Functions

Murad Tukan  
muradtuk@gmail.com

Alaa Maalouf  
alaamalouf12@gmail.com

Dan Feldman  
dannyf.post@gmail.com

The Robotics and Big Data Lab,  
Department of Computer Science,  
University of Haifa,  
Haifa, Israel

Abstract

Coreset is usually a small weighted subset of \( n \) input points in \( \mathbb{R}^d \), that provably approximates their loss function for a given set of queries (models, classifiers, etc.). Coresets become increasingly common in machine learning since existing heuristics or inefficient algorithms may be improved by running them possibly many times on the small coreset that can be maintained for streaming distributed data. Coresets can be obtained by sensitivity (importance) sampling, where its size is proportional to the total sum of sensitivities. Unfortunately, computing the sensitivity of each point is problem dependent and may be harder to compute than the original optimization problem at hand. We suggest a generic framework for computing sensitivities (and thus coresets) for wide family of loss functions which we call near-convex functions. This is by suggesting the \( f \)-SVD factorization that generalizes the SVD factorization of matrices to functions. Example applications include coresets that are either new or significantly improves previous results, such as SVM, Logistic regression, M-estimators, and \( \ell_1 \)-regression. Experimental results and open source are also provided.

1 Introduction

In common machine learning problems, we are given a set of input points \( P \subseteq \mathbb{R}^d \) (training data), and a loss function \( f : P \times \mathbb{R}^d \rightarrow [0, \infty) \), where the goal is to solve the optimization problem of finding a query (model, classifiers, centers) \( x^* \) that minimizes the sum of fitting errors \( \sum_{p \in P} f(p, x) \) over every query \( x \) in a given (usually infinite) set. For example, in \( k \)-median (or \( k \)-mean) clustering, each query is a set of \( k \) centers and the loss function is the distance (or squared distance) of a point to its nearest center. In linear regression or SVM, every input point includes a label, and the loss function is the fitting error between the classification of \( p \) via a given query to the actual label of \( p \). Empirical risk minimization (ERM) may be used to generalize the result from train to test data.

Modern machine learning. In practice, many of these optimization or learning problems are usually hard even to approximate. Instead, practical heuristics with no provable guarantees may be used to solve them. Even for well understood problems, which have close optimal solution, such as linear regression or classes of convex optimization, in the era of big data we may wish to maintain the solution in other computation models such as: streaming input data (“on-the-fly”) that provably uses small memory, parallel computations on distributed data (on the cloud, network or GPUs) as well as deletion of points, constrained optimization (e.g. sparse classifiers). Cross validation [34] or hyper-parameter tuning techniques such as AutoML [30, 32] need to evaluate many queries for different subsets of the data, and different constraints.

Coresets. One approach is to redesign existing machine learning algorithms for faster, approximate solutions and these new computation models. A different approach that is to use data summarization techniques. Coresets in particular were first used to solve problems in computational geometry [11] and 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.
We choose the following family of near-convex loss functions, with example supervised and unsupervised applications that include support vector machines, logistic regression, \(\ell_p\)-regression for any \(p \in [0, \infty)\), clustering \([2, 16, 24, 31, 37, 42, 53]\), logistic regression \([35, 47]\), LMS solvers and SVD \([28, 44, 45, 52]\), where all of these works have been dedicated to suggest a coreset for a specific problem.

A generic framework for constructing coresets was suggested in \([25, 40]\). It states that, with high probability, non-uniform sampling from the input set yields a coreset. Each point should be sampled i.i.d. with a probability that is proportional to its importance or sensitivity, and assigned a multiplicative weight which is inverse proportional to this probability, so that the expected original sum of losses over all the points will be preserved. Here, the sensitivity of an input point \(p \in P\) is defined to be the maximum of its relative fitting loss \(s(p) = f(p, x)/\sum_{q \in P} f(q, x)\) over every possible query \(x\). The size of the coreset is near-linear in the total (sum) \(t\) of these sensitivities; see Theorem 3 for details. It turns out in the recent years that many classical and hard machine learning problems \([7, 33, 55]\) have total sensitivity that is near-logarithmic or independent of the input size \(|P|\) which implies small coresets via sensitivity sampling.

Paper per problem. The main disadvantage of this framework is that the sensitivity \(s(p)\), as defined above, is problem dependent: namely on the loss function \(f\) and the feasible set of queries. Moreover, maximizing \(s(p) = f(p, x)/\sum_{q \in P} f(q, x)\) is equivalent to minimizing the inverse \(\sum_{q \in P} f(q, x)/f(p, x)\). Unfortunately, minimizing the enumerator is usually the original optimization problem which motivated the coreset in the first place. The denominator may make the problem harder, in addition to the fact that now we need to solve this optimization problem for each and every input point in \(P\). While approximations of the sensitivities usually suffice, sophisticated and different approximation techniques are frequently tailored in papers of recent machine learning conferences for each and every problem.

1.1 Problem Statement

To this end, the goal of this paper is to suggest a framework for sensitivity bounding of a family of functions, and not for a specific optimization problem. This approach is inspired by convex optimization: while we do not have a single algorithm to solve any convex optimization, we do have generic solutions for family of convex functions. E.g., linear programming, Semi-definite programming, and so on.

We choose the following family of near-convex loss functions, with example supervised and unsupervised applications that include support vector machines, logistic regression, \(\ell_p\)-regression for any \(z \in (0, \infty)\), and functions that are robust to outliers. In the Supplementary Material we suggest a more generalized version that handles a bigger family of functions; see Definition 13 and hope that this paper will inspire the research of more and larger families.

**Definition 1 (Near-convex functions).** Let \(P \subseteq \mathbb{R}^d\) be a set of \(n\) points, and let \(f : P \times \mathbb{R}^d \to [0, \infty)\) be a loss function. We call \(f\) a near-convex function if there are a convex loss function \(g : P \times \mathbb{R}^d \to [0, \infty)\) (see Definition 12 at Supplementary Material), a function \(h : P \times \mathbb{R}^d \to [0, \infty)\), and a scalar \(z > 0\) satisfying:

(i) There exist \(c_1, c_2 > 0\) such that for every \(p \in P\), and \(x \in \mathbb{R}^d\),
\[
c_1 \left( g(p, x)^z + h(p, x)^z \right) \leq f(p, x) \leq c_2 \left( g(p, x)^z + h(p, x)^z \right).
\]

(ii) For every \(p \in P\), \(x \in \mathbb{R}^d\) and \(b > 0\), we have \(g(p, bx) = b \cdot g(p, x)\).

(iii) For every \(p \in P\) and \(x \in \mathbb{R}^d\), we have \(\sum_{q \in P} h(q, x)^z \leq \frac{2}{n}\).
We denote by $F$, which is given at Corollary 10.

with respect to the loss function of $\ell$ with respect to the hinge loss, which is most used form of SVM in practice (see Sklearn library at [49]).

with a "simpler" pair of functions where the first is a convex function "g" that is linear in its argument (and thus size of coreset) would be small, depending on the "hardness" of the loss function that is used tools similar to the well-conditioned basis which was first suggested at [21] to compute such factorization in order to compute coresets for the near-convex functions. To our knowledge, we suggest the first coreset for the problem formulation $f$, the union of all functions $f$ with the above properties.

The intuition behind Definition $f$ Properties [1](ii)(iii) are used to reduce the problem to dealing with a "simpler" pair of functions where the first is a convex function "g" that is linear in its argument $x$ and the second function "h" being independent of the input points. Property [1](iv) ensures that the ellipsoid which encloses the level set of $g$ (the convex function) exists and is centered at the origin to avoid dealing with the center. By combining the properties associated with the level set of $g$ (the convex function) and Properties [1][iv] we manage to bound the loss function from above and below by the mahalanobis distance with respect to the enclosing ellipsoid. This is due to the fact that the level set encloses a contracted version of the ellipsoid which encloses the level set of $g$.

We are interested in a generic algorithm that would get a set of input points, and a loss function as above, and compute a sensitivity for each point, based on the parameters of the given loss function. In addition, we wish to use worst-case analysis and prove that for every input the total sensitivity (and thus size of coreset) would be small, depending on the “hardness” of the loss function that is encapsulated in the above parameters $z$, $R$, etc.

2 Related Work

Logistic Regression. A coreset construction algorithms for the problem of logistic regression were suggested by [33, 50, and 47]. All of these works handled variations of the problem, e.g., they all lack the incorporation of the bias term (intercept) in their loss function. Specifically speaking, both [33] and [47] didn’t account for the regularization term and its parameter. Furthermore, the coreset’s size established by [47], was dependant on the structure of the input data. As for [50], the coreset only succeed for a small subset of queries (a ball in $\mathbb{R}^d$ of radius $r$, where the coreset’s size is near linear in $r$). Contrary to previous works, our coreset approximates the logistic regression loss function including the bias parameter (intercept) and the regularization term for every possible query. This is the loss function that is usually used in practice, e.g., see Sklearn library in [49]. Finally, our coreset’s size is independent of the structure of the data.

SVM. [11, 57, 58] addressed the problem of coreset construction for SVM, yet they used squared hinge loss to enforce the SVM cost function to be strongly convex. At [60], the coreset is constructed with respect to the hinge loss which most used form of SVM in practice (see Sklearn library at [49]). However for the coreset to be constructed, a (sub-)optimal solution was required for the problem itself. In addition, the coreset size depended heavily on the ratio between the variance of each class of points. In this paper, we also address a coreset with respect to the hinge loss, yet we don’t require any (sub-)optimal solution to construct the coreset, and our coreset’s size depends on the ratio between the number of points of each class (see Corollary [9]).

$\ell_2$-Regression. A notable line of work [10, 18, 21, 54, 65] addressed the construction of coresets and sketches in this area, however, all such papers addressed the case of $z \geq 1$. Most of these works used tools similar to the well-conditioned basis which was first suggested at [21] to compute such coresets. Intuitively it can be thought of as a generalization of the SVD factorization of an input set with respect to the loss function of $\ell_2$-regression for any $z \geq 1$. In our framework we generalize this factorization in order to compute coresets for the near-convex functions. To our knowledge, we suggest the first coreset for the problem of $\ell_2$-regression for any $z \in (0, 1)$.

Outlier resistant functions (similar to M-estimators). Among such functions, is the $\ell_2$-regression for any $z \in (0, 1)$ that is mentioned above, Huber loss function [13], Tukey loss functions [12], and many more [14]. However, to our knowledge, we present the first coreset for the problem formulation which is given at Corollary [10].

3 Our contribution

In this paper, we suggest an $\varepsilon$-coreset construction algorithm with respect to any near-convex function. Specifically speaking, we provide:
(i) A generalization of the well conditioned bases of [21] to a broader family of functions, i.e., not just for $\ell_2$-Regression problems where $z \geq 1$. This informally describes a factorization of the input data with respect to a given near-convex loss function. We call such factorization the $f$-SVD of $P$ (see Definition 4).

(ii) A framework for bounding the sensitivity of each point in an input set with respect to any near-convex function. The heart of the framework relies on computing the $f$-SVD factorization described in (i); see Lemma 5 and Algorithm 1.

(iii) By (ii), we provide the first $\varepsilon$-coreset for the problem of $\ell_2$-regression where $z \in (0, 1)$, and the first $\varepsilon$-coreset for certain outlier resistant functions. We also unify existing works of coreset construction for the problems of logistic regression and SVM; see Section 6.

(iv) Experimental results on real-world and synthetic datasets for common machine learning solvers (supported by our framework) of Scikit-learn library [49], assessing the practicability and efficacy of our algorithm.

(v) An open source code implementation of our algorithm, for reproducing our results and future research [61].

3.1 Novelty

$f$-SVD factorization. In this work, we suggest a novel factorization technique of an input dataset with respect to a specific loss function $f$, we call it the $f$-SVD factorization. Roughly speaking, the heart of the $f$-SVD factorization lies in finding a diagonal matrix $D \in [0, \infty)^{d \times d}$ and an orthogonal matrix $V \in \mathbb{R}^{d \times d}$ such that the total loss $\sum_{p \in P} f(p, x)$ for any query $x \in \mathbb{R}^d$ can be bounded from above by $\sqrt{d} \|DV^T x\|_2$ and from below by $\|DV^T x\|_2$. In some sense, this can be thought of as a $(1 - 1/\sqrt{d})$-coreset (or a sketch) since it approximates the total loss for any query in $\mathbb{R}^d$ up to a multiplicative factor of $(1 - 1/\sqrt{d})$. In order to obtain such factorization, we forge a link between the Löwner ellipsoid [36] and the properties of near-convex functions; see Fig. 1 for a detailed illustrative explanation, Definition 4 and Lemma 16 for the formal details.

Note that SVD factorization is a special case of $f$-SVD due to that fact that SVD handles functions of the form $\sqrt{\sum_{p \in P} |p^T x|^2}$ and attempts to achieve the same purpose. The $f$-SVD factorization is a generalization of the well-conditioned bases of [21].

From $f$-SVD to sensitivity bounds. With the lower bound on the total loss that is guaranteed by the $f$-SVD, we show how to bound the sensitivity of each point in the dataset. On the other hand, the upper bound on the total loss provided by the $f$-SVD factorization, helps us in bounding the total sensitivity. Having this being said, we use the $f$-SVD factorization to suggest a sensitivity bounding framework for a set of points with respect to any near-convex function $f \in \mathcal{F}$; see Lemma 5.

4 Preliminaries

Notations. For integers $n, d \geq 2$, we denote by $0_d$ the origin of $\mathbb{R}^d$, and by $[n]$ the set $\{1, \ldots, n\}$. The set $\mathbb{R}^{n \times d}$ denotes the union over every $n \times d$ real matrix, and $I_d \in \mathbb{R}^{d \times d}$ denotes the identity matrix. We say that a matrix $A \in \mathbb{R}^{d \times d}$ is orthogonal if and only if $A^T A = A A^T = I_d$. Finally, throughout the paper, vectors are addressed as column vectors, and $\| \cdot \| : \mathbb{R}^d \to \mathbb{R}$ is a weight function.

In what follows, we formally prove the notion of $\varepsilon$-coreset in our context.

Definition 2 ($\varepsilon$-coreset). Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, $f : P \times \mathbb{R}^d \to [0, \infty)$ be a near-convex function, and let $\varepsilon \in (0, 1)$. An $\varepsilon$-coreset for $P$ with respect to $f$, is a pair $(S, v)$ where $S \subseteq P$, $v : S \to (0, \infty)$ is a weight function, such that for every $x \in \mathbb{R}^d$, $1 - \frac{\sum_{p \in P} v(q) f(q, x)}{\sum_{p \in P} f(p, x)} \leq \varepsilon$.

The following theorem formally describes how to construct an $\varepsilon$-coreset via the sensitivity framework.

Theorem 3 (Restatement of Theorem 5.5 in [17]). Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, and let $f : P \times \mathbb{R}^d \to [0, \infty)$ be a loss function. For every $p \in P$ define the sensitivity of $p$ as $\sup_{x \in \mathbb{R}^d} \frac{f(p, x)}{\sum_{q \in P} f(q, x)}$, where the sup is over every $x \in \mathbb{R}^d$ such that the denominator is non-zero. Let $s : P \to [0, 1]$ be
Let $S$ be a set of $n$ points, $f \in F$ be a near-convex loss function (see Definition 7), and let $g, h, c_1, z$ be defined as in the context of Definition 7 with respect to $f$. Let $D \in [0,\infty)^{d \times d}$ be a diagonal matrix, and let $V \in [0,\infty)^{d \times d}$ be an orthogonal matrix, such that for every $x \in \mathbb{R}^d$, $c_1 \left( \|DV^T x\|_2^2 + \sum_{p \in P} h(p, x) z \right) \leq \sum_{p \in P} f(p, x)$, and let $\alpha \in \Theta \left( \sqrt{d} \right)$ such that for every $x \in \mathbb{R}^d$, $\sum_{p \in P} g(p, x)\max_{1 \leq z} \leq \alpha \|DV^T x\|_2 \max_{1 \leq z}$. Define $U : P \rightarrow \mathbb{R}^d$ such that $U(p) = (VD)^{-1} p$ for every $p \in P$. The tuple $(U, D, V)$ is the $f$-SVD of $P$. 

Note that (i) such factorization exists for any set of points $P$ and any near-convex loss function $f : P \times \mathbb{R}^d \rightarrow [0,\infty)$ satisfying Definition 1, and (ii) the matrix $V^T D V$ is invertible due to the fact that $D$ is of full rank which is a result of Property (iv) of Definition 4. Both (i)-(ii) hold by using Löwner ellipsoid; see Fig. 1 for intuitive explanation, and Lemma 5 at the Supplementary Material for formal proof.

In what follows, we proceed to bound the sensitivity of each point and the total sensitivity, with respect to a loss function $f \in F$. This is by using the $f$-SVD of $P$.

**Lemma 5.** Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, and let $f \in F$ be a near-convex loss function as in Definition 7. Let $g, h, c_1, c_2, z$ be defined as in the context of Definition 7 with respect to $f$. $(U, D, V)$
be the $f$-SVD of $P$, and let $\alpha \in \Theta \left( \sqrt{d} \right)$ which satisfies the conditions in Definition 4. Suppose that there exists a set \( \{v_j\}_{j=1}^{O(d)} \subseteq \mathbb{R}^d \) of $O(d)$ unit vectors and $c > 0$, such that for every unit vector \( y \in \mathbb{R}^d \) and $p \in P$, \( g \left( p, (DV^T)^{-1} y \right) \leq c \sum_{j=1}^{O(d)} g \left( p, (DV^T)^{-1} v_j \right) \). Then, for every $p \in P$, the sensitivity of $p$ is bounded by \( s(p) \leq \frac{2c}{c_1} + \frac{2c}{c_1} \sum_{j=1}^{O(d)} \left( g \left( p, (DV^T)^{-1} v_j \right) \right)^{-1} \), and the total sensitivity is bounded by \( \sum_{p \in P} s(p) \leq \frac{2c}{c_1} + \frac{2c}{c_1} \max \{ n^{1-\varepsilon}, 1 \} \alpha^2 O(d) \).

The existence of the set \( \{v_j\}_{j=1}^{O(d)} \) is discussed in details at the supplementary material at Section D.

### 5.2 The coreset construction

Algorithm 1 receives as input, a set $P$ of $n$ points in $\mathbb{R}^d$, a loss function $f \in \mathcal{F}$ (see Definition 1), and a sample size $m > 0$. As Theorem 6 states, if the sample size $m$ is sufficiently large, then Algorithm 1 outputs a pair $(S, v)$ that is with high probability, an $\varepsilon$-coreset for $P$ with respect to $f$.

First, we set $d'$ to be VC dimension of the quadruple $(P, \mathbb{I}, \mathbb{R}^d, f)$; See Definition 15. The crux of our algorithm lies in generating the importance sampling distribution via efficiently computing upper bound on the sensitivity of each point (Lines 5-7). To do so, we compute the $f$-SVD of $P$ at Lines [3-4] and we use it to bound the sensitivity of each $p \in P$ as stated in Lemma 5, see Line 6.

Now we have all the needed ingredients to use Theorem 3 in order to obtain an $\varepsilon$-coreset, i.e., we sample i.i.d $m$ points from $P$ based on their sensitivity bounds (see Line 9), and assign a new weight for every sampled point at Line 10.

**Algorithm 1: CORESET($P, f, m$)**

Input: A set $P \subseteq \mathbb{R}^d$ of $n$ points, a near-convex loss function $f : P \times \mathbb{R}^d \rightarrow [0, \infty)$, and a sample size $m \geq 1$.

Output: A pair $(S, v)$ that satisfies Theorem 6.

1. Set $d' :=$ the VC dimension of quadruple $(P, \mathbb{I}, \mathbb{R}^d, f)$ // See Definition 15
2. Set $g$ and $(c_1, c_2)$ to be a function and a set of real positive numbers respectively, satisfying Property (1) and (II) of Definition 4 with respect to $f$.
3. Set $c > 0$ and $\{v_1, \ldots, v_d\}$ to be a positive scalar and a set of $d$ unit vectors in $\mathbb{R}^d$ respectively satisfying Lemma 5.
4. Set $(U, D, V)$ to be the $f$-SVD of $(P, w)$ // See Definition 1.
5. For every $p \in P$ do:
   6. Set $s(p) := \frac{2c}{c_1} \sum_{j=1}^{d} g \left( p, (DV^T)^{-1} v_j \right) + \frac{2c}{c_1} \max \{ n^{1-\varepsilon}, 1 \} \alpha^2 n$ // the bound of the sensitivity of $p$ as in Lemma 5.
   7. Set $\tilde{c} := \frac{1}{c_1}$ to be a sufficiently large constant // Can be determined from Theorem 6.
   8. Pick an i.i.d sample $S$ of $m$ points from $P$, where each $p \in P$ is sampled with probability $\frac{s(p)}{\tilde{c}}$.
   9. Set $v : \mathbb{R}^d \rightarrow [0, \infty]$ to be a weight function such that for every $q \in S$, $v(q) = \frac{\tilde{c}}{s(q) m}$.
10. return $(S, v)$.

**Theorem 6.** Let $P \subseteq \mathbb{R}^d$ be set of $n$ points, and $f \in \mathcal{F}$ be a near-convex function. Let $R, r > 0$ be a pair of positive scalars as in Definition 7 with respect to $f$, and let $c, c_1, c_2, \alpha$ be defined as in the context of Lemma 5 with respect to $f$. Let $\varepsilon, \delta \in (0, 1)$ be an error parameter and a probability of failure respectively, and let $d'$ be the VC dimension of the triplet $(P, f, \mathbb{R}^d)$. Let $t = \frac{2c}{c_1} + \frac{2c}{c_1} \max \{ n^{1-\varepsilon}, 1 \} \alpha^2 n$, $m \in O \left( \frac{1}{\delta} \left( d' \log \left( t \right) + \log \left( \frac{1}{\delta} \right) \right) \right)$, and let $(S, v)$ be the output of a call to CORESET($P, f, m$). Then, (i) with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset of size $m$ for $P$ with respect to $f$; see Definition 2. (ii) The overall time for constructing $(S, v)$ is bounded by $O \left( T(n, d^2) \log \left( \frac{1}{\delta} \right) \right)$, where $T(n, d)$ is a bound on the time it takes to compute a gradient of $\sum_{p \in P} f(p, x)$ with respect to any query $x \in \mathbb{R}^d$.

**Poly-logarithmic coreset size.** We provide an analysis that shows how to obtain a coreset of size poly-logarithmic in the input size $n$; see Algorithm 2 and Lemma 17 at the Supplementary Material.
6 Applications

In what follows, we provide various applications for our framework, e.g. SVM, Logistic Regression, \( \ell_z \) for \( z \in (0, 1) \), outlier resistant functions (similar to Tukey in behavior). For additional problems supported by our framework, we refer the reader to Section G at the Supplementary Material.

Table 1: Results: The table below presents the coreset size and the time needed for constructing it with respect to a specific set of problems, where the input is a set of \( n \) points in \( \mathbb{R}^d \) denoted by \( P \). In the table, \( \text{nnz} (P) \) denotes the total number of nonzero entries in the set \( P \), \( \tilde{C} \) denotes the ratio between the number of positive and negative labeled points (in practice, it’s a constant number), \( \lambda = \sqrt{n} \) is the given regularization parameter for the problems, \( \gamma \geq 1 \) is defined as in Corollary 10 \( \epsilon \) is the error parameter, and \( \delta \) is the probability of failure.

<table>
<thead>
<tr>
<th>Problem type</th>
<th>Coreset’s size</th>
<th>Construction time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>( O \left( \frac{dn}{\gamma} \left( d \log (d \sqrt{n}) + \log \left( \frac{1}{\delta} \right) \right) \right) )</td>
<td>( O (nd^2) )</td>
</tr>
<tr>
<td>( \ell_z )-Regression for ( z \in (0, 1) )</td>
<td>( O \left( \frac{n^{1-z} \lambda^2 + 1}{\epsilon ^2} \left( d \log \left( d \sqrt{n} + \frac{C_z^2 + 1}{C} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) )</td>
<td>( O (\text{nnz} (P) \log n + dO(1)) )</td>
</tr>
<tr>
<td>SVM</td>
<td>( O \left( \frac{n^{1-z} \lambda^2 + 1}{\epsilon ^2} \left( d \log \left( \frac{dn}{\gamma} + \frac{C_z^2 + 1}{C} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) )</td>
<td>( O (nd^2) )</td>
</tr>
<tr>
<td>Restricted ( \ell_z )-regression</td>
<td>( O \left( \frac{n^{1-z} \lambda^2 + 1}{\epsilon ^2} \left( d \log \left( \gamma d^2 + \frac{1}{\delta} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) )</td>
<td>( O (\text{nnz} (P) \log n + dO(1)) )</td>
</tr>
</tbody>
</table>

**Corollary 7 (Logistic Regression).** Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points such that for every \( p \in P \), \( \| p \|_2 \leq 1 \), \( y : P \to \{-1, 1\} \) be a labeling function, \( \lambda \geq 1 \) be a regularization parameter such that for every \( p \in P \), \( x \in \mathbb{R}^d \) and \( b \in \mathbb{R} \), \( f_{\text{Log}} \left( p, \begin{bmatrix} x \\ b \end{bmatrix} \right) = \frac{1}{2} \ln \left( 1 + e^{p^T x + y(p) \cdot b} \right) + \frac{1}{2n} \| x \|_2^2 \).

Let \( \epsilon, \delta \in (0, 1) \) be an error parameter and a probability of failure respectively, \( m \in O \left( \frac{dn}{\gamma \lambda} \left( d \log \left( \frac{dn}{\gamma} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \), and let \( (S, v) \) be the output of a call to CORESET \( (P, f_{\text{Log}}, m) \). Then, with probability at least \( 1 - \delta \), \( (S, v) \) is an \( \epsilon \)-coreset (of size \( m \)) for \( P \) with respect to \( f_{\text{Log}} \).

**Corollary 8 (\( \ell_z \)-Regression where \( z \in (0, 1) \)).** Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points, \( z \in (0, 1) \) and let \( f_{\text{SVM}} \in \mathcal{P} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \) be a loss function such that for every \( x \in \mathbb{R}^d \), and \( p \in P \), \( f_{\text{SVM}}(p, x) = \| p^T x \|^2 \).

Let \( \epsilon, \delta \in (0, 1) \), \( m \in O \left( \frac{n^{1-z} \lambda^2 + 1}{\epsilon ^2} \left( d \log \left( n \lambda^{-z} \lambda^2 + 1 \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \), and let \( (S, v) \) be the output of a call to CORESET \( (P, f_{\text{SVM}}, m) \). Then, with probability at least \( 1 - \delta \), \( (S, v) \) is an \( \epsilon \)-coreset (of size \( m \)) for \( P \) with respect to \( f_{\text{SVM}} \).

We now show how our framework can be used to compute an \( \epsilon \)-coreset for some query spaces where the involved loss functions are not from the family \( \mathcal{F} \). The coreset construction algorithms are hidden in the constructive proofs of the following corollaries.

**Corollary 9 (Support Vector Machines).** Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points such that for every \( p \in P \), \( \| p \| \leq 1 \). Let \( y : P \to \{-1, 1\} \) be a labeling function, \( \lambda \geq 1 \) be a regularization parameter such that for every \( p \in P \), \( x \in \mathbb{R}^d \), and \( b \in \mathbb{R} \), \( f_{\text{SVM}} \left( p, \begin{bmatrix} x \\ b \end{bmatrix} \right) = \lambda \max \left\{ 0, 1 - (p^T x + y(p) \cdot b) \right\} + \frac{1}{2n} \| x \|_2^2 \). Let \( P_+ = \left\{ \| p \|_2 \leq 1, y(p) = 1 \right\}, P_- = P \setminus P_+, \tilde{C} = \frac{|P_+|}{|P_-|} \).

Then, there exists an algorithm that gets the set \( P \) as an input, and returns a pair \( (S, v) \), such that (i) with probability at least \( 1 - \delta \), \( (S, v) \) is an \( \epsilon \)-coreset for \( P \) with respect to \( f_{\text{SVM}} \), and (ii) the size of the coreset is \( |S| \leq O \left( \frac{1}{\delta} \left( \frac{dn}{\gamma \lambda} + \frac{C_z^2 + 1}{C} \right) \left( d \log \left( \frac{dn}{\gamma \lambda} \left( \frac{dn}{\gamma} + \frac{C_z^2 + 1}{C} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \right) \).

**Corollary 10 (Outlier resistant functions).** Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points, and let \( f_{\text{Rest}} : \mathbb{R} \times \mathbb{R}^d \to [0, \infty) \) be loss function such that for every \( x \in \mathbb{R}^d \), and \( p \in P \), \( f_{\text{Rest}}(p, x) = \min \left\{ \| p^T x \|_2, \| x \|_2 \right\} \).

Then, there exists an algorithm that gets the set \( P \) as an input, and returns a pair \( (S, v) \), such that (i) with probability at least \( 1 - \delta \), \( (S, v) \) is an \( \epsilon \)-coreset for \( P \) with respect to \( f_{\text{Rest}} \), and (ii) the size of the coreset is \( O \left( \frac{n^{1-z} \lambda^2 + 1}{\epsilon ^2} \left( d \log \left( \gamma d^2 + \frac{1}{\delta} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \), where \( \gamma \) is defined in the proof.

1 Problems which are reduced to \( \ell_z \)-regression problems for any \( z \geq 1 \), are easier to deal with in term of coreset construction time due to the existence of randomized algorithm of computing the Löwner ellipsoid by [15]; see Section H for detailed description.
7 Experimental Results

In what follows we evaluate our coreset against uniform sampling on real-world datasets, with respect to the SVM problem, Logistic regression problem and $\ell_z$-regression problem for $z \in (0, 1)$. Additional details of our setup can be found at Section F of the Supplementary Material.

**Software/Hardware.** Our algorithms were implemented in Python 3.6 [63] using “Numpy” [48], “Scipy” [64] and “Scikit-learn” [49]. Tests were performed on 2.59GHz i7-6500U (2 cores total) machine with 16GB RAM.

**Datasets.** The following datasets were used for our experiments mostly from UCI machine learning repository [22]: (i) HTRU [22] — 17,898 radio emissions of the Pulsar star each consisting of 9 features. (ii) Skin [22] — 245,057 random samples of R,G,B from face images consisting of 4 dimensions. (iii) Cod-rna [62] — consists of 59,535 samples, 8 features, which has two classes (i.e. labels), describing RNAs. (iv) Web dataset [9] — 49,749 web pages where each record consists of 300 features. (v) 3D spatial networks [22] — 3D road network with highly accurate elevation information (+-20cm) from Denmark used in eco-routing and fuel/Co2-estimation routing algorithms consisting of 434,874 records where each record has 4 features.

**Evaluation against uniform sampling.** At Fig. 2a-2i we have chosen 20 sample sizes, starting from 50 till 500, at Figures 22-25 we have chosen 20 sample sizes starting from 4000 till 16,000. At each sample size, we generate two coresets, where the first is using uniform sampling and the latter is using Algorithm 1. For each coreset $(S, v)$, we find $x^* \in \arg\min_{x \in \mathbb{R}^d} \sum_{p \in S} v(p) f(p,x)$, and the approximation error $\varepsilon$ is set to be $(\sum_{p \in S} v(p) f(p,x^*))/(\min_{x \in \mathbb{R}^d} \sum_{p \in S} f(p,x)) - 1$. The results were averaged across 40 trials, while the shaded regions correspond to the standard deviation.

**Evaluation against prior work.** We can not have a fair comparison between our coreset to prior coresets for Logistic regression [47,56] due to the fact that our formulation of the problem is different. As for support vector machines, we compared our efficacy against [60], the same way that we have compared against uniform sampling. Although not in all cases our approach outperforms [60] in terms of relative error (i.e., $\varepsilon$), our approach is much faster than that of [60]: see Figure 4.

Figure 2: Experimental results
8 Conclusions and open problems

In this paper, we have provided what we call the $f$-SVD of $P$ with respect a given near-convex loss function $f \in \mathcal{F}$, as well as sensitivity bounding framework using the $f$-SVD. What interests us is to draw back forcing $f$ to have a centrally symmetric level set as well as embedding the center of the Lüwner ellipsoid into the sensitivity bound. This is crucial step for generalizing the framework towards a much broader family of functions, e.g., log-Lipschitz functions. We are aware that for $\ell_p$-regression problems where $p \geq 1$, Lewis weights have been used by [17] and are considered to be the state of the art coreset for these problems. We aim to generalize the applicability of Lewis weights and other sketching techniques towards different functions, and as far as we know, we consider the above issues to be open problems.
Broader Impact

Our work provides a strong theoretical result, where we have suggested a generic framework for bounding the sensitivity with respect to broad family of functions. Practically, this family imposes widely used applications such as SVM, Logistic regression, ℓ_p-Regression and more.

Although, Broader Impact discussion is not directly applicable, our work can be used to accelerate many known machine learning solvers under various settings such as distributed, streaming, etc.

References


A Generalization of our tools

We first define the term query space which will aid us in simplifying the proofs as well as the corresponding theorems.

**Definition 11 (Query space).** Let \( P \) be a set of \( n \geq 1 \) points in \( \mathbb{R}^d \), \( w: P \to [0, \infty) \) be a non-negative weight function, and let \( f: P \times \mathbb{R}^d \to [0, \infty) \) denote a loss function. The tuple \((P, w, \mathbb{R}^d, f)\) is called a query space.

Our paper relies on using known theorems associated with convex loss functions to prove our technical results. Thus, for completeness we give a formal definition of a convex loss functions as follows.

**Definition 12 (Convex loss function).** Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points, and let \( f: P \times \mathbb{R}^d \to [0, \infty) \) be a loss function. We say that \( f \) is a convex loss function if for every \( p \in P \), \( f(p, \cdot): \mathbb{R}^d \to [0, \infty) \) is a convex function i.e., for every \( \theta \in [0, 1] \) and every \( x, y \in \mathbb{R}^d \)

\[
 f(p, \theta x + (1 - \theta)y) \leq \theta f(p, x) + (1 - \theta)f(p, y).
\]

Below, we present a straightforward generalization of the properties in Definition 1, is applied to grasp much more variety of functions, by taking the weights into account and not setting them to 1 for every point in the input set of points as well as other properties.

**Definition 13 (Generalization of Definition 1).** Let \((P, w, \mathbb{R}^d, f)\) be a query space, where \( f: P \times \mathbb{R}^d \to [0, \infty) \) is a loss function. We call \( f \) a near-convex loss function if there exists a convex loss function \( g: P \times \mathbb{R}^d \to [0, \infty) \), a function \( h: P \times \mathbb{R}^d \to [0, \infty) \) and a scalar \( z > 0 \) that satisfies:

(i) There exist \( c_1, c_2 \in (0, \infty) \) such that for every \( p \in P \), and \( x \in \mathbb{R}^d \),

\[
 c_1 (g(p, x)^z + h(p, x)^z) \leq f(p, x) \leq c_2 (g(p, x)^z + h(p, x)^z).
\]

(ii) There exist \( c_3, c_4 \in (0, \infty) \) such that for every \( p \in P \), \( x \in \mathbb{R}^d \) and \( b \in (0, \infty) \),

\[
 c_3 bg(p, x) \leq g(p, bx) \leq c_4 bg(p, x).
\]

(iii) There exists \( c_5 \in (0, \infty) \) such that for every \( p \in P \) and \( x \in \mathbb{R}^d \),

\[
 \frac{w(p)h(p, x)^z}{\sum_{q \in P} w(q)h(q, x)^z} \leq \frac{c_5 w(p)}{\sum_{q \in P} w(q)}.
\]

(iv) The set \( \mathcal{X}_g = \left\{ x \in \mathbb{R}^d \left| \sum_{p \in P} w(p)^{\max\{1, z\}} g(p, x)^{\max\{1, z\}} \leq 1 \right. \right\} \) is centrally symmetric, i.e., for every \( x \in \mathcal{X}_g \), we have \(-x \in \mathcal{X}_g\), and there exist \( R, r \in (0, \infty) \) such that \( B(0_d, r) \subset \mathcal{X}_g \subset B(0_d, R) \), where \( B(x, y) \) denotes a ball of radius \( y > 0 \), centered at \( x \in \mathbb{R}^d \).

We denote by \( \mathcal{F} \) the union of all functions \( f \) with the above properties.

Due to such changes, we also give a generalization towards the definition of \( f \)-SVD, as in what follows.

**Definition 14 (Generalization of Definition 4).** Let \((P, w, \mathbb{R}^d, f)\) be a query space, such that \( f \in \mathcal{F} \), and let \( g, h, c_1, c_2, c_3, c_4, z \) be defined as in the context of Definition 3 with respect to \( f \). Let \( D, V \in \mathbb{R}^{d \times d} \) be a diagonal matrix and an orthogonal matrix respectively, and let \( \alpha \in \Theta \left( \sqrt{d} \right) \) such that for every \( x \in \mathbb{R}^d \),

\[
 c_1 \left( (c_3 \|DV^Tx\|_2^z + \sum_{p \in P} w(p)h(p, x)^z) \right) \leq \sum_{p \in P} w(p)f(p, x),
\]

and

\[
 \sum_{p \in P} w(p)^{\max\{1, z\}} g(p, x)^{\max\{1, z\}} \leq (c_4 \alpha \|DV^Tx\|_2^{\max\{1, z\}}).
\]

Let \( U: P \to \mathbb{R}^d \) such that \( U(p) = (VD)^{-1}p \) for every \( p \in P \). The tuple \((U, D, V)\) is the \( f \)-SVD of \((P, w)\).
B VC dimension

Definition 15 (VC-dimension \[7\]). For a query space \((P, w, \mathbb{R}^d, f)\) and \(r \in [0, \infty)\), we define
\[
\text{ranges}(x, r) = \{ p \in P \mid w(p)f(p, x) \leq r \},
\]
for every \(x \in \mathbb{R}^d\) and \(r \geq 0\). The dimension of \((P, w, \mathbb{R}^d, f)\) is the size \(|S|\) of the largest subset \(S \subset P\) such that
\[
|\{ S \cap \text{ranges}(x, r) \mid x \in \mathbb{R}^d, r \geq 0 \}| = 2^{|S|},
\]
where \(|A|\) denotes the number of points in \(A\) for every \(A \subseteq \mathbb{R}^d\).

C Existence of \(f\text{-SVD factorization}

Lemma 16. Let \((P, w, \mathbb{R}^d, f)\) be a query space, such that \(f \in \mathcal{F}\). Let \(g, h, c_1, c_2, c_3, c_4, z\) be defined as in the context of Definition 13 with respect to \(f\), \(\alpha \in \Theta\left(\sqrt{d}\right)\) and let \(\beta = \max\{1, z\}\). Then there exists a diagonal matrix \(D \in \mathbb{R}^{d \times d}\) and an orthogonal matrix \(V \in \mathbb{R}^{d \times d}\) such that for every \(x \in \mathbb{R}^d\),
\[
\sum_{p \in P} w(p) \max\{1, \frac{1}{\sqrt{d}}\} g(p, x)^\beta \leq \left(c_4 \alpha \|DV^T x\|_2\right)^\beta,
\]
and
\[
c_1 \left((c_3 \|DV^T x\|_2) + \sum_{p \in P} w(p) h(p, x)^z\right) \leq \sum_{p \in P} w(p)f(p, x).
\]

Proof. We prove Lemma 16 using Löwner ellipsoid; See [36].

Using Löwner ellipsoid. Let \(X_g = \left\{ x \in \mathbb{R}^d \mid \left(\sum_{p \in P} w(p) \max\{1, \frac{1}{\sqrt{d}}\} g(p, x)^\beta\right)^\frac{1}{\beta} \leq 1 \right\}\). Since \(f \in \mathcal{F}\) (see Definition 13), and \(g\) is convex, we have that (i) \(X_g\) is a convex set, and (ii) \(X_g\) is centrally symmetric. Then by Theorem III of [36], there exists an ellipsoid \(E\), known as the Löwner ellipsoid that is centered at the origin \(0_d\), such that
\[
\frac{1}{\sqrt{d}} E \subseteq X_g \subseteq E,
\]
where \(\frac{1}{\sqrt{d}} E\) denotes the set \(\left\{ \frac{1}{\sqrt{d}} x \mid x \in E \right\}\).

By combining Property [iv] of Definition 13 with (3), there exists \(r \in (0, \infty)\) such that \(B(0_d, r) \subseteq X_g \subseteq E\). Since \(B(0_d, r) \subseteq E\), then \(E\) is an ellipsoid where each of its axes has positive length. By that, there exists a diagonal matrix \(D \in \mathbb{R}^{d \times d}\) of positive entries and an orthogonal matrix \(V \in \mathbb{R}^{d \times d}\) such that, (i) \(E = \{ y \in \mathbb{R}^d \mid \|DV^T y\|_2 \leq 1 \}\), and (ii) \(V^T D V\) is a positive semi-definite matrix.

Put \(x \in \mathbb{R}^d\) and now we proceed to derive the bounds.

Proving (1). Let \(y = \frac{1}{\|DV^T x\|_2} x\). By the definition of \(E\) in (3) and the definition of \(y\), we have that \(y \in E\), and by combining (3) with the assumption that \(\alpha \in \Theta\left(\sqrt{d}\right)\) we obtain that \(y \in E \subseteq \alpha X_g\). Then
\[
\frac{1}{\alpha} y \in \frac{1}{\alpha} E \subseteq X_g,
\]
which consequently leads to
\[
\sum_{p \in P} w(p) \max\{1, \frac{1}{\sqrt{d}}\} g\left(p, \frac{1}{\alpha} y\right)^\beta \leq 1,
\]
where the inequality holds by Property (iv) of Definition 13 with respect to \( g \). Hence,
\[
\sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g(p, x)^{\beta} \leq \left( c_4 \alpha \left\| DV^T x \right\|_2 \right)^\beta \sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g\left(p, \frac{x}{\alpha \left\| DV^T x \right\|_2}\right)^\beta
\]
where the first inequality is by substituting \( b := \alpha \left\| DV^T x \right\|_2 \) and \( x := \frac{1}{\alpha} \) in Property (ii) of \( g \) (see Definition 13), and the second inequality is by combining the fact that \( y = \frac{1}{\alpha \left\| DV^T x \right\|_2} x \) with (5).

**Proving (2).** Let \( b' = \alpha \left\| DV^T x \right\|_2 \). By (5), we get that \( \sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g\left(p, \frac{1}{\alpha} x \right) \geq 1 \). In addition, Property (iv) of Definition 13 states that every vector in \( \mathbb{R}^d \) of norm \( r \) is inside \( \mathcal{X}_g \). Thus, there exists \( b \geq b' \) such that \( \sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g\left(p, \frac{1}{\alpha} x \right) \geq \beta \). By (3), we have that \( \left\| DV^T x \right\|_2 = b \left\| DV^T x \right\|_2 \leq b \) where \( z = \frac{1}{\beta} x \). Hence, by plugging \( x := z \) and \( b := b \) in Property (ii) of Definition 13 we obtain that
\[
\sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g\left(p, x \right) \geq \beta \sum_{p \in P} w(p)^{\max\{\frac{1}{2}, 1\}} g\left(p, z \right) = \beta (c_3 b)^\beta = (c_3 )^\beta \alpha \left\| DV^T x \right\|_2 \left\| DV^T x \right\|_2 \beta . \quad (7)
\]

By combining Property (ii) of Definition 13 with (6) and (7), Lemma 16 holds.

**D The intuition behind the existence of the set \( \{v_j\}_{j=1}^{O(d)} \)**

Let \((P, w, \mathbb{R}^d, f)\) be a query space (see Definition 11) such that \( f \in \mathcal{F} \) as in Definition 13. Let \( g, c_4, z \) be defined as in the context of Definition 13 with respect to \( f \). Since by Definition 13 the level set of \( g \) is bounded and is contained in a ball of radius \( R \), then it holds that Löwner ellipsoid of the level set of \( g \) is contained in the ball \( B(0_d, \sqrt{dR}) \), using similar arguments to those at (3).

Then there exist a point \( \hat{c} \) on the ball \( B(0_d, R) \) such that on the ball \( B(\hat{c}, 2d \max \{ R, 1/R \}) \) there exist a set of \( d+1 \) vectors \( \{ \hat{v}_i \}_{i=1}^{d+1} \) each of norm 2\(dR\) such that any unit vector \( x \in \mathbb{R}^d \) is a convex combination of them. For instance, the set \( \{ \hat{v}_i \}_{i=1}^{d+1} \) can be computed by finding a \( d \)-simplex which inscribes the unit ball \( B(0_d, 1) \).

Let \( A \in \mathbb{R}^{d \times d} \) and observe that by using Jensen’s inequality, for any unit vector \( x \in \mathbb{R}^d \) and \( p \in P \), we have
\[
g(p, Ax)^2 \leq \left( \sum_{i=1}^{d+1} g\left(p, A \hat{v}_i \right) \right)^2 \leq d^2 \sum_{i=1}^{d+1} g\left(p, A \hat{v}_i \right)^2 \leq (2c_4 d^2 R)^2 \sum_{i=1}^{d+1} g\left(p, A \hat{v}_i \right)^2 ,
\]
where \( v_i = \frac{1}{\sqrt{dR}} \hat{v}_i \) for each \( i \in [d+1] \).

Thus the assumption in Lemma 20 holds.

**Note that** this holds in the general case of \( g \) however in the applications that we handled, since \( g \) is also homogeneous function, then only \( d \) vectors suffices for satisfying this assumption in Lemma 20; see Section 3.

**E Extension towards Streaming and distributed settings**

Algorithm 1 can be easily extended towards streaming and distributed settings as presented at Algorithm 2. At the beginning, the data arrives in a streaming fashion, e.g. in batches, where our coreset scheme (see Algorithm 1) is applied on each of these batches. When we have two \( \varepsilon \)-coresets in memory, we merge them and an \( \varepsilon \)-coreset is constructed upon their merge. This procedure is done until (i) there is no points left in the stream and (ii) there is exactly one coreset left in memory.

Algorithm 2 begins with initializing the batches to an empty sets as well as setting the height of the tree to 1; see lines 1-2. In what follows, for each 2\(t \) of streamed points, we generate an \( \varepsilon \)-coreset on
Algorithm 2: Streaming-Coreset\((P, w, f, \ell, \varepsilon, \delta)\)

**Input:** A set \(P \subseteq \mathbb{R}^d\) of \(n\) points, a weight function \(w : \mathbb{R}^d \rightarrow [0, \infty)\), a leaf size \(\ell > 0\) a convex loss function function \(f : \mathbb{R}^d \rightarrow [0, \infty)\), an error parameter \(\varepsilon \in (0, 1)\), and probability \(\delta \in (0, 1)\).

**Output:** A pair \((S, v)\) which is an \((h)\)-coreset for \((P, \mathbb{R}^d, X, f)\), with probability of at least \(1 - \delta h\).

1. \(B_i \leftarrow \emptyset\) for every \(1 \leq i \leq \infty\)
2. \(h \leftarrow 1\)
3. **for each** set \(Q\) of consecutive \(2\ell\) points from \(P\) **do**
   4. \((T, v) := \text{CORESET}(Q, w, f, \varepsilon, \frac{2}{\log n}, \frac{2}{\log n})\)
   5. \(j \leftarrow 1\)
   6. \(B_j := B_j \cup (T, v)\)
   7. **for each** \(j \leq h\) **do**
      8. **while** \(|B_j| \geq 2\) **do**
         9. \((T_1, v_1), (T_2, v_2) := \text{pop first pair of consecutive items in } B_j\)
         10. **for every** \(p \in T_1 \cup T_2\) **set** \(v'(p) := \begin{cases} v_1(p) & p \in T_1, \\ v_2(p) & \text{Otherwise} \end{cases}\)
         11. \((T, v) := \text{CORESET}(T_1 \cup T_2, v', f, \varepsilon, \frac{2}{\log n}, \frac{2}{\log n})\)
         12. \(B_{j+1} := B_{j+1} \cup (T, v)\)
         13. \(h := \max\{h, j + 1\}\)
     14. \((S, v) := B_h\)
     15. **return** \((S, v)\)

this set as presented at lines [6][7]. Lines [8][9] depict the core of the merge-and-reduce tree, which is the binary tree building fashion from the leaves (the incoming batches) towards the root of the tree. Finally, we return the root of the tree as shown at lines [14][15]. For much broader and detailed explanation regarding the merge-and-reduce tree, we refer the reader towards [7].

E.1 From sublinear to poly-logarithmic coreset size

**Lemma 17** (Variant of Lemma 4, [60]). Let \(P \subseteq \mathbb{R}^d\) be a set of \(n\) points, and let \(f \in \mathcal{F}\) be a near-convex loss function. Let \(\varepsilon \in \left[\frac{1}{\log n}, \frac{1}{2}\right]\), \(\delta \in \left[\frac{1}{\log n}, 1\right]\) and let \(t\) denote the total sensitivity from Lemma 3. Suppose that there exists some \(\beta \in (0, 1, 0.8)\) such that \(t \in \Theta(n^\beta)\) and let \(\ell \geq 2^{\frac{\delta}{\varepsilon}}\). Let \((S, v)\) be the output of a call to Streaming-Coreset\((P, w, f, \ell, \varepsilon, \delta)\). Then \((S, v)\) is an \(\varepsilon\)-coreset of size

\[
|S| \in (\log n)^{O(1)}.
\]

**Proof.** First we note that using Theorem 6 on each node in the merge-and-reduce tree, would attain that the root of the tree, i.e., \((S, v)\) attains that for every \(w\)

\[
(1 - \varepsilon)^{\log n} \sum_{p \in P} w(p)f(p, x) \leq \sum_{p \in S} w(p)f(p, x) \leq (1 + \varepsilon)^{\log n} \sum_{p \in P} w(p)f(p, x),
\]

with probability at least \((1 - \delta)^{\log n}\).

We observe by the properties of the natural number \(e\),

\[
(1 + \varepsilon)^{\log n} = \left(1 + \frac{\varepsilon \log n}{\log n}\right)^{\log n} \leq e^{\varepsilon \log n},
\]

which when replacing \(\varepsilon\) with \(\varepsilon' = \frac{\varepsilon}{\frac{\log n}{\log n}}\) in the above inequality as done at Lines [4] and [9] of Algorithm 2, we obtain that

\[
(1 + \varepsilon')^{\log n} \leq e^{\frac{\varepsilon'}{2}} \leq 1 + \varepsilon,
\]

where the second inequality holds since \(\varepsilon \in \left[\frac{1}{\log n}, \frac{1}{2}\right]\).
As for the lower bound, observe that
\[(1 - \varepsilon)^{\log n} \geq 1 - \varepsilon \log n,\]
where the inequality holds since \(\varepsilon \in \left[\frac{1}{\log n}, \frac{1}{2}\right].\)

Hence,
\[(1 - \varepsilon')^{\log n} \geq 1 - {\varepsilon'} \log n = 1 - \frac{\varepsilon}{2} \geq 1 - \varepsilon.\]

Similar arguments holds also for the failure probability \(\delta\). What is left for us to do is setting the leaf size which will attain us an \(\varepsilon\)-coreset of size poly-logarithmic in \(n\) (the number of points in \(P\)).

Let \(\ell \in (0, \infty)\) be the size of a leaf in the merge-and-reduce tree. We observe that a coreset of size poly-logarithmic in \(n\), can be achieved by solving the inequality
\[
\frac{2\ell}{2} \geq (2^{\ell})^{\beta},
\]
which is invoked when ascending from any two leafs and their parent node at the merge-and-reduce tree.

Rearranging the inequality, we yield that
\[
\ell^{1-\beta} \geq 2^\beta.
\]

Since \(\ell \in (0, \infty)\), any \(\ell \geq \sqrt[\beta]{2^\beta} \) would be sufficient for the inequality to hold. What is left for us to do, is to show that when ascending through the merge-and-reduce tree from the leaves towards the root, each parent node can’t be more than half of the merge of it’s children (recall that the merge-and-reduce tree is built in a binary tree fashion, as depicted at Algorithm 2).

Thus, we need to show that,
\[
\sum_{i=1}^{i} \beta_i \cdot \ell^i \leq \frac{2 \sum_{i=0}^{\beta_k} \cdot \ell^{\beta_k-1}}{2} = 2 \sum_{i=1}^{\beta_k} \cdot \ell^{\beta_k-1},
\]
holds, for any \(i \in [\lceil \log n \rceil]\) where \(\log n\) is the height of the tree. Note that the left most term is the parent node’s size and the right most term represents half the size of both parent’s children nodes.

In addition, for \(i = 1\), the inequality above represents each node which is a parent of leaves. Thus, we observe that for every \(i \geq 1\), the inequality represents ascending from node which is a root of a sub-tree of height \(i - 1\) to it’s parent in the merge-and-reduce tree.

By simplifying the inequality, we obtain the same inequality which only addressed the leaves. Hence, by using any \(\ell \geq 2^{\frac{1}{1+\beta}}\) as a leaf size in the merge and reduce tree, we obtain an \(\varepsilon\)-coreset of size poly-logarithmic in \(n\).

\[\square\]

\section{Proofs for the Main Theorems}

Throughout this section, we will present generalized versions of the lemmata and theorems that are presented at Section 5 and Section 6.

\subsection{Generalization of Lemma 5}

\textbf{Lemma 18 (Equivalence of norms, \cite{51}).} Let \(a, b > 0\) such that \(a \leq b\). Then for every \(x \in \mathbb{R}^d\),
\[
\|x\|_b \leq \|x\|_a \leq d^{\frac{1}{1+z}} \|x\|_b.
\]

\textbf{Claim 19.} \([\text{Result of Hölder’s Inequality}]\) Let \(\{a_i\}_{i=1}^{n}\) be a set of \(n\) non-negative numbers, \(z \in (0, 1)\) be a real number. Then
\[
\sum_{i=1}^{n} |a_i|^z \leq n^{1-z} \left( \sum_{i=1}^{n} |a_i| \right)^z.
\]
Proof. Let $z' = \frac{1}{z}$ and for every $i \in [n]$, let $\hat{a}_i = |a_i|^z$. Let $e \in [1]^n$. We have

$$
\sum_{i=1}^{n} |a_i|^z = \sum_{i=1}^{n} \hat{a}_i \leq \|e\|_1 \left( \sum_{i=1}^{n} \hat{a}_i^{z'} \right)^{1/z'} = n^{1-z} \left( \sum_{i=1}^{n} |a_i| \right)^z,
$$

where the first and last equalities are by definition of $\hat{a}_i$, and the inequality is by Hölder’s inequality.

Lemma 20. Let $(P, w, \mathbb{R}^d, f)$ be a query space (see Definition 17) such that $f \in \mathcal{F}$ as in Definition 17. Let $g, h, c_1, c_2, c_3, c_4, c_5, z$ be defined as in the context of Definition 17 with respect to $f$. $(U, D, V)$ be the $f$-SVD of $(P, w)$, and let $\alpha \in \Theta \left( \sqrt{n} \right)$ which satisfies the conditions in Definition 4.

Suppose that there exists a set of $O(d)$ unit vectors $\{v_j\}_{j=1}^{O(d)}$ and $c \in (0, \infty)$, such that for every unit vector $y \in \mathbb{R}^d$ and $p \in P$,

$$
g(p, (DV^T)^{-1}y)^z \leq c \sum_{j=1}^{O(d)} g(p, (DV^T)^{-1}v_j)^z.
$$

Then, for every $p \in P$, the sensitivity of $p$ with respect to the query space $(P, w, \mathbb{R}^d, f)$ is bounded by

$$
s(p) \leq \left( \frac{2c_2c_5 w(p)}{c_1 \sum_{q \in P} w(q)} \right)^z + \frac{cc_2}{c_1c_3^2} \sum_{j=1}^{O(d)} w(p) \left( g \left( p, (DV^T)^{-1}v_j \right) \right)^z,
$$

and the total sensitivity is bounded by

$$
\sum_{p \in P} s(p) \leq \frac{c_2c_5}{c_1} + \frac{cc_2c_4}{c_1c_3^2} \max \left\{ n^{1-z}, 1 \right\} \alpha^z O(d).
$$

Proof. Let $n$ denote the number of points in $P$. Put $p \in P$, $x \in \mathbb{R}^d$ such that $\sum_{q \in P} w(q)f(q, x) > 0$, and let $y = \frac{1}{\|DV^T x\|_2}DV^T x$. We observe that,

$$
\frac{f(p, x)}{\sum_{q \in P} w(q)f(q, x)} \leq \frac{f(p, x)}{c_1 \left( (c_3 \|DV^T x\|_2)^z + \sum_{q \in P} w(q)h(q, x, z) \right)} \leq \frac{c_2g(p, x)^z + c_2h(p, x)^z}{c_1 \left( c_3 \|DV^T x\|_2^z \right) + \sum_{q \in P} w(q)h(q, x, z)} \leq \frac{c_2g(p, x)^z + c_2h(p, x)^z}{c_1 \left( c_3 \|DV^T x\|_2^z \right)} + \frac{c_2h(p, x)^z}{c_1 \sum_{q \in P} h(q, x, z)},
$$

where (9) holds by Lemma 16, (10) holds by Property (10) of Definition T with respect to $f$, and the last inequality follows from plugging $a_1 := c_2g(p, x)$, $r_1 := c_2h(p, x)$, $a_2 := c_1c_3 \|DV^T x\|_2$ and $r_2 := c_3 \sum_{q \in P} h(q, x, z)$ into Claim 21.

Note that when $h(q, z) = 0$ for every $q \in P$ and $z \in \mathbb{R}^d$, then we obtain from (10), that the rightmost term of (11) is zero.

We also have,

$$
\frac{1}{\|DV^T x\|_2^z} g(p, x)^z \leq \frac{1}{c_3^z} g \left( p, \frac{x}{\|DV^T x\|_2} \right)^z \leq \frac{c}{c_3^z} \sum_{j=1}^{d} g \left( p, (DV^T)^{-1}v_j \right)^z,
$$

(12)

(13)

(14)
As for the total sensitivity, we first observe that if $z \in (0, 1)$,  
\[
\sum_{q \in P} \sum_{j=1}^{O(d)} w(q)g \left( q, (DV^T)^{-1} v_j \right)^z = \sum_{j=1}^{O(d)} w(q)g \left( q, (DV^T)^{-1} v_j \right)^z 
\]
\[
= \sum_{j=1}^{O(d)} \left( \sum_{q \in P} w(q)g \left( q, (DV^T)^{-1} v_j \right) \right)^z 
\]
\[
\leq n^{1-z} \sum_{j=1}^{O(d)} \left( \sum_{q \in P} w(q)g \left( q, (DV^T)^{-1} v_j \right) \right)^z 
\]
\[
\leq n^{1-z} (c_4 \alpha)^z \sum_{j=1}^{O(d)} \left\| DV^T (DV^T)^{-1} v_j \right\|_2^z 
\]
\[
= n^{1-z} (c_4 \alpha)^z \sum_{j=1}^{O(d)} \| v_j \|_2 
\]
\[
= n^{1-z} (c_4 \alpha)^z O(d), 
\]
where (16) holds by the independency between the summation over $q \in P$ and summation over $j \in [d]$, (17) holds since the weights are non-negative by definition, (18) holds by plugging $z := z$, $n := n$, $a_i := w(q)^{\frac{1}{2}} g \left( q, (DV^T)^{-1} v_j \right)$ for every $i \in [n]$ into Claim 19 where $q$ denotes the $i$th point in $P$. (19) holds by Lemma 16 (20) follows since $DV^T (DV^T)^{-1} = I_d$, and finally (21) follows from the assumption of Lemma 5.

Similarly for the case of $z \geq 1$,  
\[
\sum_{q \in P} \sum_{j=1}^{O(d)} w(q)g \left( q, (DV^T)^{-1} v_j \right)^z \in (c_4 \alpha)^z O(d). 
\]

Hence, Lemma 5 holds as  
\[
\sum_{p \in P} s(p) \leq \sum_{p \in P} \frac{c_2c_5 w(p)}{c_1} + \sum_{p \in P} \frac{cc_2}{c_1 c_3^2} \sum_{j=1}^{O(d)} w(p)g \left( p, (DV^T)^{-1} v_j \right)^z
\]
\[
\leq \frac{c_2c_5}{c_1} + \frac{cc_2}{c_1 c_3^2} \max \left\{ n^{1-z}, 1 \right\} \alpha^2 O(d),
\]
where the first inequality holds by (15), and the second inequality holds by combining (16)–(22).

F.2 Proof of Theorem 6

Theorem 6. Let $P \subseteq \mathbb{R}^d$ be set of $n$ points, and $f \in \mathcal{F}$ be a near-convex function. Let $R, r > 0$ be a pair of positive scalars as in Definition 1 with respect to $f$, and let $c, c_1, c_2, \alpha$ be defined as in the context of Lemma 5 with respect to $f$. Let $\varepsilon, \delta \in (0, 1)$ be an error parameter and a probability of failure respectively, and let $d'$ be the VC dimension of the triplet $(P, f, \mathbb{R}^d)$. Let $t = \frac{c_2}{c_1} + \frac{cc_2}{c_1 c_3^2} \max \left\{ n^{1-z}, 1 \right\} \alpha^2 d' \log \left( \frac{1}{\varepsilon^2} \right) + \log \left( \frac{1}{\delta} \right)$, and let $(S, v)$ be the output of a call to CORESET$(P, f, m)$. Then, (i) with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset of size
where the first inequality holds by the definition of \( b \).

Algorithm 1. This is done by computing the Löwner ellipsoid, as explained in the proof of Lemma 16. The computation of the Löwner ellipsoid requires a separation oracle, where we use the gradient of \( f \). The overall time is dominated by computing the \(-\)SVD of \((P, w)\), i.e., \((U, D, V)\) at Line 4 of Algorithm 1. This is done by computing the Löwner ellipsoid, as explained in the proof of Lemma 16. The computation of the Löwner ellipsoid, requires a separation oracle, where we use the gradient of \( g \) as a candidate, similarly to [10]. We refer to [41] for more details on the computation of Löwner ellipsoid.

F.3 Proof of Corollary 7

Claim 21. Let \( a_1, r_1, a_2, r_2 \in [0, \infty) \) such that \( a_2, r_2 > 0 \). Then,

\[
\frac{a_1 + r_1}{a_2 + r_2} \leq \frac{a_1 + r_1}{r_2}.
\]

Proof. Observe that,

\[
\frac{a_1 + r_1}{a_2 + r_2} = \frac{a_1}{a_2 + r_2} + \frac{r_1}{a_2 + r_2} \leq \frac{a_1}{a_2} + \frac{r_1}{r_2},
\]

where the inequality holds since \( a_2, r_2 > 0 \) and \( a_1, r_1 \geq 0 \).

Claim 22. Let \( N \geq 2 \). For every \( i \in [N] \), let \( a_i \geq 0 \) and \( b_i > 0 \). Then,

\[
\frac{\max \{ a_1, a_2, \ldots, a_N \}}{\max \{ b_1, b_2, \ldots, b_N \}} \leq \max \left\{ \frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_N}{b_N} \right\}.
\]

Proof. Let \( \hat{i} \in \arg \max_{i \in [N]} a_i \) and let \( \hat{j} \in \arg \max_{i \in [N]} b_i \). Then,

\[
\frac{\max \{ a_1, a_2, \ldots, a_N \}}{\max \{ b_1, b_2, \ldots, b_N \}} = \frac{a_{\hat{i}}}{b_{\hat{j}}} \leq \frac{a_{\hat{i}}}{b_{\hat{i}}} \leq \max \left\{ \frac{a_1}{b_1}, \frac{a_2}{b_2}, \ldots, \frac{a_N}{b_N} \right\},
\]

where the first inequality holds by the definition of \( b_{\hat{j}} \).

Claim 23. For every \( z, b \in \mathbb{R} \),

\[
\ln (1 + e^{z+b}) \leq 2 \ln \left(1 + e^{z^2 + b} \right).
\]

Proof. Put \( b \in \mathbb{R} \), and note that for every \( z \in \mathbb{R} \), we have

\[
\ln (2) + z^2 - z \geq 0,
\]

by rearranging the above, we get that

\[
\ln (2) + z^2 \geq z.
\]

Applying the exponentiation operation on both sides with respect to the natural number \( e \) as the base, yields

\[
2e^{z^2} \geq e^z,
\]

and since \( e^{z^2+b} > 0 \),

\[
e^z \leq e^{z^2} \left(2 + e^{z^2+b} \right).
\]
By multiplying each side by $e^b$ and adding 1, we obtain that
\[1 + e^{x+b} \leq 1 + 2e^{x^2+b} + e^{x^2+2b}.\]
Applying the logarithm function on both sides of the inequality above proves Claim 23 as
\[\ln \left(1 + e^{x+b}\right) \leq 2 \ln \left(1 + e^{x^2+b}\right).\]

\[\square\]

**Lemma 24** (Bernoulli’s inequality, [39]). Let $x \geq -1$ be a real number and let $r \in [0, 1]$ be a positive real number. Then,
\[(1 + x)^r \leq 1 + rx\]

**Lemma 25.** Let $N > 1$, $c \in [1, N]$ and let $p \in \mathbb{R}^d$ such that $\|p\|_2 \leq 1$. Then for every $(x, b) \in \mathbb{R}^d \times \mathbb{R}$,
\[
(i) \quad \frac{1}{c} \ln \left(1 + e^{p^T x + b}\right) + \frac{1}{2N} \|x\|^2 \leq \frac{1}{c} (p^T x)^2 + 4 \max \left\{ \frac{1}{c} \ln (1 + e^b), \frac{1}{2N} \|x\|^2 \right\},
\]
\[
(ii) \quad \text{and} \quad \frac{1}{c} \ln \left(1 + e^{p^T x + b}\right) + \frac{1}{2N} \|x\|^2 \geq \frac{c}{8N} \left( (p^T x)^2 + 4 \max \left\{ \frac{1}{c} \ln (1 + e^b), \frac{1}{2N} \|x\|^2 \right\} \right).
\]

**Proof.** Put $(x, b) \in \mathbb{R}^d \times \mathbb{R}$. We now proceed to prove Lemma 25.

**Proof of Claim (i).** By plugging $z := p^T x$ and $b := b$ into Claim 23, we obtain that
\[
\ln \left(1 + e^{p^T x + b}\right) \leq 2 \ln \left(1 + e^{(p^T x)^2 + b}\right) \leq 2 \ln \left(e^{(p^T x)^2} (1 + e^b)\right) = 2 (p^T x)^2 + 2 \ln (1 + e^b),
\]
where the second inequality holds since $e^{(p^T x)^2} \geq 1$, and the equality follows from properties of the logarithm function.

Thus, Claim (i) holds since
\[
\frac{1}{c} \ln \left(1 + e^{p^T x + b}\right) + \frac{1}{2N} \|x\|^2 \leq \frac{2}{c} (p^T x)^2 + \frac{1}{c} \ln (1 + e^b) + \frac{1}{2N} \|x\|^2
\]
\[
\leq \frac{4}{c} (p^T x)^2 + 4 \max \left\{ \frac{1}{c} \ln (1 + e^b), \frac{1}{2N} \|x\|^2 \right\},
\]
where the first inequality is by (23), the second inequality holds by properties of the $\max$ operator.

**Proof of Claim (ii).** We start by noting that since $\|p\| \leq 1$, we have that
\[
\|x\|^2 \geq |p^T x|,
\]
which consequently leads to
\[
\frac{1}{c} \ln \left(1 + e^{-\|x\|^2 + b}\right) \geq \frac{1}{c} \ln \left(1 + e^{-\|x\|^2 + b}\right) \geq \frac{1}{2N} \ln \left(1 + e^{-\|x\|^2 + b}\right),
\]
where the second inequality holds since $c \leq N$.

We show that
\[
\ln \left(1 + e^{-\|x\|^2 + b}\right) + \|x\|^2 \geq \frac{1}{2} \ln \left(1 + e^b\right),
\]
holds for every $x \in \mathbb{R}^d$ and $b \in \mathbb{R}$. In order to to that, we first define the function $q : \mathbb{R} \to (0, \infty)$ such that for every $r \in \mathbb{R}$, $q(r) = R \ln \left(1 + e^{-|r|+b}\right) + r^2$.

Let $W$ denotes the Lambert W function (see [19]). Minimizing $q(r)$ over $r \in \mathbb{R}$, requires computing the derivative of $q(r)$ with respect to $r$, and setting it to zero. We observe that when setting the derivative to zero we obtain that $r^* \in [-W(1), W(1)]$, i.e., the left term of (26) attains its minimal value at some $x^* \in \mathbb{R}^d$ such that $\|x^*\|^2 \in [0, W(1)]$.

Observe that for every $x \in \mathbb{R}^d$
\[
\ln \left(1 + e^{-\|x\|^2 + b}\right) + \|x\|^2 \geq \ln \left(1 + e^{-\|x^*\|^2 + b}\right) + \|x^*\|^2 \geq \ln \left(1 + e^{-\|x^*\|^2 + b}\right) \geq \ln \left(1 + e^{-W(1)+b}\right),
\]

\[22\]
where the first inequality holds by the definition of \( x^* \), the second inequality holds since \( \| x^* \|^2 \geq 0 \), and the last inequality follows from the observation that \( \| x^* \| \in [0, W(1)] \).

Since \( e^{-W(1)} \in (0, 1) \), we have that

\[
\ln \left( 1 + e^{-W(1)+b} \right) \geq \ln \left( (1 + e^b) e^{-W(1)} \right) = e^{-W(1)} \ln \left( 1 + e^b \right) \geq \frac{1}{2} \ln \left( 1 + e^b \right),
\]

where the first inequality holds by plugging \( r := e^{-W(1)} \) and \( x := e^b \) into Lemma [24], the equality holds by properties of the logarithm function, and the last inequality holds since \( e^{-W(1)} \geq \frac{1}{2} \).

We also observe that

\[
\frac{1}{2N} \| x \|^2 \geq \frac{1}{2N} \| p^T x \|^2 = \frac{c}{2N} \left( \frac{1}{c} \| p^T x \|^2 \right),
\]

where the first inequality holds by properties of the logarithm function, and the last inequality holds since \( \frac{c}{2N} \cdot \frac{1}{c} = \frac{1}{2N} \).

Thus by combining [25], [26], and [27], Claim (ii) holds as

\[
\frac{1}{c} \ln \left( 1 + e^{p^T x + b} \right) + \frac{1}{2N} \| x \|^2 \geq \frac{c}{4N} \left( \frac{1}{2c} \| p^T x \|^2 + \max \left\{ \frac{1}{2c} \ln \left( 1 + e^b \right), \frac{1}{2N} \| x \|^2 \right\} \right).
\]

\[\Box\]

**Lemma 26.** Let \( (P, w, \mathbb{R}^{d+1}, f_{\text{LOG}}) \) be a query space, \( y : P \to \{1, -1\} \) be a labelling function, \( \lambda \geq 1 \) be a regularization parameter, such that for every \( p \in P \), \( b \in \mathbb{R} \) and \( x \in \mathbb{R}^d \),

\[
f_{\text{LOG}}(p, (x \mid b)) = \frac{1}{2} \sum_{q \in P} w(q) \| x \|_2^2 + \frac{1}{\lambda} \ln \left( 1 + e^{p^T x + y(p) b} \right).
\]

For every \( p \in P \), let \( P_{y(p)} = \{ q \mid q \in P, y(q) = y(p) \} \) denote the set of points with the same label as the label assigned to \( p \). Let \( (U, D, V) \) be the f-SVD of \( (P, w) \) with respect to \( f_{\text{LOG}} \). Then, claims (i) – (ii) hold as follows:

(i) for every \( p \in P \), the sensitivity of \( p \) with respect to the query space \( (P, w, \mathbb{R}^{d+1}, f_{\text{LOG}}) \) is bounded by

\[
s(p) = \frac{32}{\lambda} \left( \frac{2w(p)}{\sum_{q \in P_{y(p)}} w(q)} + w(p) \| U(p) \|^2 \right) \sum_{q \in P_{y(p)}} w(q),
\]

(ii) and the total sensitivity is bounded by

\[
\sum_{p \in P} s(p) \leq \frac{32}{\lambda} (2 + d) \sum_{p \in P} w(p).
\]

**Proof.** Put \( p \in P \) and let \( P_{y(p)} \) denote the subset of points from \( P \) with same label as \( p \), \( P_{y(p)} = \{ q \mid q \in P, y(q) = y(p) \} \). Observe that for every \( q \in P_{y(p)} \)

\[
\sup_{(x, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{w(p) f_{\text{LOG}}(p, (x \mid b))}{w(q) f_{\text{LOG}}(q, (x \mid b))} \leq \sup_{(x, b) \in \mathbb{R}^d \times \mathbb{R}} \frac{w(p) f_{\text{LOG}}(p, (x \mid b))}{w(q) f_{\text{LOG}}(q, (x \mid b))},
\]

where the inequality holds since \( P_{y(p)} \subseteq P \), and \( f_{\text{LOG}}(q, (x \mid b)) \geq 0 \) for every \( q \in P \), and \( (x \mid b) \in \mathbb{R}^d \times \mathbb{R} \).

Note the following:

(a) For every \( q \in P \), \( x \in \mathbb{R}^d \) and \( \gamma \geq 0 \) we have \( |q^T \gamma x| = \gamma |q^T x| \).
(b) Since \(|q^T x|\) is convex function, it also holds that \(\sum_{q \in P} |q^T x|^2\) is convex due to the fact that sum of convex functions is also convex.

(c) The level set \(\left\{ x \in \mathbb{R}^d, \sum_{q \in P} w(q) |q^T x|^2 \leq 1 \right\}\) is convex and is centrally symmetric.

(d) For every \(x \in \mathbb{R}^d, b \in \mathbb{R}\) we have that
\[
\begin{align*}
  w(q) \max \left\{ \frac{1}{\lambda} \ln (1 + e^b), \frac{1}{2} \sum_{q \in P_{y(p)}} \frac{1}{w(q)} \|x\|_2^2 \right\} \\
  \leq 2 \sum_{q \in P_{y(p)}} w(q) \max \left\{ \frac{1}{\lambda} \ln (1 + e^b), \frac{1}{2} \sum_{q \in P_{y(p)}} \frac{1}{w(q)} \|x\|_2^2 \right\}
\end{align*}
\]
where the inequality holds by plugging \(a_1 := w(q) \frac{1}{\lambda} \ln (1 + e^b), a_2 := \frac{1}{2} \sum_{q \in P_{y(p)}} \frac{1}{w(q)} \|x\|_2^2\), \(b_1 := \frac{\sum_{q \in P_{y(p)}} w(q)}{\lambda} \ln (1 + e^b), b_2 := \frac{1}{2} \|x\|_2^2\) into Claim 22.

Thus, combining (a), (b), (c), (d) and Lemma 25 allows us to plug

- \(f(p(x | b)) := f_{\text{log}}(p(x | b)), g(p(x | b)) := \frac{1}{\lambda} |p^T x|\) and \(h(p(x | b)) := \max \left\{ \sqrt{\frac{1}{\lambda} \ln (1 + e^b)}, \sqrt{\frac{1}{2} \sum_{q \in P_{y(p)}} \frac{1}{w(q)} \|x\|_2^2} \right\}\), for every \(p \in P, x \in \mathbb{R}^d, b \in \mathbb{R}\),

- \(\alpha = 1\),
- \(c_1 := \frac{1}{8N}\) and \(c_2 = 4\),
- \(c_i := 1\) for every \(i \in [3, 4]\)
- \(c_5 := 2\),
- \(z := 2\),
- \(v_j := e_j\) for every \(j \in [d]\) where \(e_j\) denotes the vector with a 1 in the \(j\)th coordinate and 0’s elsewhere,
- and \(c := 1\),

into Lemma 5 which yields that \(f_{\text{log}} \in \mathcal{F}\) and the sensitivity of each point \(q \in P_{y(p)}\) is bounded by
\[
s(q) = \frac{32}{\lambda} \left( 2 \sum_{q \in P_{y(p)}} \frac{w(q)}{\|q\|_2^2} + w(q) \sum_{j=1}^{d} |U(q)^T e_j|^2 \right) \sum_{q \in P_{y(p)}} w(q).
\]

Claim (i) now holds since for every \(q \in P\),
\[
\sum_{j=1}^{d} |U(q)^T e_j|^2 = \|U(q)\|_2^2,
\]
where the equality follows from definition of \(e_j\) for every \(j \in [d]\).

As for the total sensitivity, we have by Lemma 5
\[
\sum_{q \in P_{y(p)}} s(q) \leq \frac{32}{\lambda} \left( 2 \sum_{q \in P_{y(p)}} \frac{w(q)}{\|q\|_2^2} + d \right) \sum_{q \in P_{y(p)}} w(p),
\]

24
and
\[ \sum_{q \in P \setminus P_y(p)} s(q) \leq \frac{32}{\lambda} \left( 2 \frac{w(q)}{\sum_{\tilde{q} \in P \setminus P_y(p)} w(\tilde{q})} + d \right) \sum_{q \in P \setminus P_y(p)} w(p). \]

Hence, Claim (ii) holds as
\[ \sum_{q \in P} s(q) \leq \frac{32}{\lambda} (2 + d) \sum_{q \in P} w(q). \]

\[ \square \]

**Corollary 7 (Logistic Regression).** Let \( P \subset \mathbb{R}^d \) be a set of \( n \) points such that for every \( p \in P \), \( \| p \|_2 \leq 1 \), \( y : P \to \{-1, 1\} \) be a labeling function, \( \lambda \geq 1 \) be a regularization parameter such that for every \( p \in P \), \( x \in \mathbb{R}^d \) and \( b \in \mathbb{R} \), \( f_{\text{LOG}}(p, \left[ \begin{array}{c} x \\ b \end{array} \right]) = \frac{1}{2} \ln \left( 1 + e^{p^T x + y(p) b} \right) + \frac{1}{2\lambda} \| x \|_2^2 \).

Let \( \varepsilon, \delta \in (0, 1) \) be an error parameter and a probability of failure respectively, \( m \in O \left( \frac{4n}{\varepsilon^2} \left( d \log \left( \frac{4n}{\varepsilon^2} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \), and let \((S, v)\) be the output of a call to \( \text{Coreset} (P, f_{\text{LOG}}, m) \). Then, with probability at least \( 1 - \delta \), \((S, v)\) is an \( \varepsilon \)-coreset (of size \( m \)) for \( P \) with respect to \( f_{\text{LOG}} \).

**Proof.** First, observe that by Lemma 26, the total sensitivity is bounded by \( t := \frac{32}{\lambda} (2 + d) \sum_{q \in P} w(p) \).

Hence, plugging \( s(p) \) for every \( p \in P \) from Lemma 26 yields that \((S, v)\) is an \( \varepsilon \)-coreset of size \( O \left( \frac{4n}{\varepsilon^2} \left( d \log \left( \frac{4n}{\varepsilon^2} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \).

\[ \square \]

**F.4 Proof of Corollary 8**

**Lemma 27.** Let \((P, w, \mathbb{R}^{d+1}, f_{\text{SVD}})\) be a query space, such that for every \( p \in P \) and \( x \in \mathbb{R}^d \),
\[ f_{\text{SVD}}(p, x) = |p^T x|^z. \]

Let \((U, D, V)\) be the \( f \)-SVD of \((P, w)\) with respect to \( f_{\text{SVD}} \). Then, claims (i) – (ii) hold as follows:

(i) For every \( p \in P \), the sensitivity of \( p \) with respect to the query space \((P, w, \mathbb{R}^d, f_{\text{SVD}})\) is bounded by
\[ s(p) = w(p) \| U(p) \|_z^z, \]

(ii) and the total sensitivity is bounded by
\[ \sum_{p \in P} s(p) \leq n^{1-z} d^{z+1}. \]

**Proof.** Let \( g : P \to [0, \infty) \) such that for every \( p \in P \) and \( x \in \mathbb{R}^d \), \( g(p, x) = |p^T x| \), and for every \( i \in [d] \) let \( e_i \) denote the vector with 1 in the \( i \)th coordinate and 0’s elsewhere. Observe that:

(a) For every \( q \in P \), \( x \in \mathbb{R}^d \) and \( b \geq 0 \) we have \( g(p, b \cdot x) = b \cdot g(p, x) \).

(b) Since \( g(q, x) \) is a convex function for every \( q \in P \), it also holds that \( \sum_{q \in P} w(q)^{\frac{1}{z}} g(q, x) \) is convex due to the fact that sum of convex functions is also convex.

(c) The level set \( \left\{ x \mid x \in \mathbb{R}^d \setminus \sum_{q \in P} w(q)^{\frac{1}{z}} g(q, x) \leq 1 \right\} \) is convex and is centrally symmetric.

(d) In addition, for every unit vector \( y \in \mathbb{R}^d \)
\[ \| p^T y \|^z \leq \| p \|_y^z \leq \| p \|_z^z = \sum_{i=1}^d |p^T e_i|^z, \]
where the first inequality holds by Cauchy’s inequality, the second inequality is by Lemma 18 and the equality is by properties of norm.
Hence combining (a), (b), (c) and (d), allows us to plug

• \( f(q, x) := f_{\text{NC}, t}(q, x), g(q, x) := c(q^T x) \) and \( h(q, x) := 0 \) for every \( q \in P \) and \( x \in \mathbb{R}^d \).
• \( c_i := 1 \) for every \( i \in [4] \),
• \( c_5 := 0 \),
• \( \alpha := \sqrt{d} \),
• \( v_j := e_j \) for every \( j \in [d] \) where \( e_j \) denotes the vector with a 1 in the \( j \)th coordinate and 0’s elsewhere,
• \( e := 1 \), and
• \( z := z \)

into Lemma\(^5\) which yields that \( f_{\text{NC}, t} \in \mathcal{F} \) and the sensitivity of each point \( p \in P \) is bounded by

\[
s(p) = \sum_{i=1}^{d} \left| U(p)^T e_i \right|^2,
\]

and the total sensitivity is bounded by

\[
\sum_{q \in P} s(q) \leq n^{1-z}d^{z+1}.
\]

\[\square\]

**Corollary 8** (\( \ell_z \)-Regression where \( z \in (0, 1) \)). Let \( P \subseteq \mathbb{R}^d \) be a set of \( n \) points, \( z \in (0, 1) \) and let \( f_{\text{NC}, t} : P \times \mathbb{R}^d \) be a loss function such that for every \( x \in \mathbb{R}^d \), and \( p \in P \), \( f_{\text{NC}, t}(p, x) = |p^T x|^2 \).

Let \( \varepsilon, \delta \in (0, 1) \), \( m \in O \left( n^{1-z}d^{z+1} \left( d \log \left( n^{1-z}d^{z+1} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \), and let \( (S, v) \) be the output of a call to \text{CORESET}(P, f_{\text{NC}, t}, m) \). Then, with probability at least \( 1 - \delta \), \((S, v)\) is an \( \varepsilon \)-coreset (of size \( m \)) for \( P \) with respect to \( f_{\text{NC}, t} \).

**Proof.** First, observe that by Lemma\(^{27}\) the total sensitivity is bounded by \( t := n^{1-z}d^{z+1} \). Plugging \( s(p) \) for every \( p \in P \) from Lemma\(^{27}\), \( t := t, \varepsilon := \varepsilon \) and \( \delta := \delta \) into Theorem\(^{6}\) yields an \( \varepsilon \)-coreset of size \( O \left( n^{1-z}d^{z+1} \left( d \log \left( n^{1-z}d^{z+1} \right) + \log \left( \frac{1}{\delta} \right) \right) \right) \).

\[\square\]

**F.5 Proof of Corollary 9**

**Lemma 28.** Let \( \gamma \in (0, 1) \), \( N \geq 1 \), and let \( c \in [1, N] \). Let \( p \in \mathbb{R}^d \) such that \( \|p\|_2 \leq 1 \) and let \( X = \{(x, b) \in \mathbb{R}^d \times \mathbb{R} \mid \|x\|_2 \geq \gamma, |b| \leq 9 \|x\|_2 \} \). Then, for every \( (x, b) \in X \), claims (i) – (ii) hold as follows:

(a) \[
\frac{\|x\|_2^2}{N} + \frac{1}{\varepsilon} \max \left\{ 0, 1 + p^T x + b \right\} \leq \frac{2}{\varepsilon} \max \left\{ \frac{1}{\varepsilon}, \frac{b}{\varepsilon}, \frac{\|x\|_2^2}{N} \right\}.
\]

(b) \[
\frac{\|x\|_2^2}{N} + \frac{1}{\varepsilon} \max \left\{ 0, 1 + p^T x + b \right\} \geq \frac{c_5^2}{(1+10\gamma)N} \left( \frac{1}{\varepsilon} \right) \max \left\{ \frac{1}{\varepsilon}, \frac{b}{\varepsilon}, \frac{\|x\|_2^2}{N} \right\}.
\]

**Proof.** Put \((x, b) \in X \).

**Proof of Claim (i).** The proof is by the following case analysis:

1. If \( p^T x + b \geq 0 \), we have

\[
\max \left\{ 0, 1 + p^T x + b \right\} = 1 + p^T x + b \leq 2 + 2 |p^T x|^2 + b \leq 2 |p^T x|^2 + 2 \max \{1, b\},
\]

where the equality holds by the assumption of the case, and the first inequality holds since for every \( z \in \mathbb{R} \), we have \( 1 + z \leq 2z^2 + 2 \).
2. Otherwise,
\[
\max \left\{ 0, 1 + p^T x + b \right\} \leq 1 \leq 2 \max \left\{ 1, b \right\} \leq 2 \|p^T x\|^2 + 2 \max \left\{ 1, b \right\}
\]
de where the first inequality holds by the assumption of the case.

By taking both the above cases in mind, Claim (i) holds.

**Proof of Claim (ii).** Similar to the proof of Claim (i), we use the same case analysis:

1. If \( p^T x + b \geq 0 \), we observe that
\[
\frac{1}{c} \max \left\{ 0, 1 + p^T x + b \right\} + \frac{\|x\|^2}{N} = \frac{1}{c} \left( 1 + p^T x + b \right) + \frac{\|x\|^2}{N} \geq \frac{c}{2N} \left( \frac{1}{c} |p^T x|^2 + \max \left\{ 1, b, \frac{1}{2N} \|x\|^2 \right\} \right),
\]
de where the equality holds by the assumption of this case, and the inequality holds since \( |p^T x|^2 \leq \|x\|^2 \) due to the assumption that \( \|p\|_2 \leq 1 \).

2. Otherwise,
\[
\frac{|p^T x|^2}{c} + \max \left\{ 1, b \right\} \|x\|^2 \leq \frac{1}{c} + \frac{10 \|x\|^2}{\gamma c},
\]
de where the inequality follows since by the definition of the set \( X \), we have that \( b \leq 9 \|x\|_2 \leq \frac{9\|x\|_2^2}{\gamma} \).

We also note that,
\[
\frac{1}{c} \max \left\{ 0, 1 + p^T x + b \right\} + \frac{\|x\|^2}{N} \geq \frac{\|x\|^2}{N},
\]
de holds since the max term is non-negative.

Let \( l = \frac{c\gamma^2}{N(1 + 10\gamma)} \). Observe that
\[
\frac{l}{c} \left( 1 + \frac{10 \|x\|^2}{\gamma} \right) \leq \frac{\|x\|^2}{N},
\]
de since \( \|x\|^2 \geq \gamma \).

Hence, we obtain that
\[
\frac{1}{c} \max \left\{ 0, 1 + p^T x + b \right\} + \frac{\|x\|^2}{N} \geq \frac{c\gamma^2}{N(1 + 10\gamma)} \left( \frac{|p^T x|^2}{c} + \max \left\{ \frac{1}{c}, \frac{b}{c}, \frac{1}{2N} \|x\|^2 \right\} \right),
\]
de where the inequality holds by combining (29), (30) and (31).

Combining both cases proves Claim (ii).

\( \square \)

**Lemma 29.** Let \( (P, w, \mathbb{R}^{d+1}, f_{\text{SVM}}) \) be a query space, \( y : P \to \{1, -1\} \) be a labelling function, \( \lambda \geq 1 \) be a regularization parameter such that for every \( p \in P, x \in \mathbb{R}^d, \) and \( b \in \mathbb{R}, \)
\[
f_{\text{SVM}}(p, (x \mid b)) = \frac{1}{2 \sum_{q \in P} w(q)} \|x\|^2 + \lambda \max \left\{ 0, 1 - (p^T x + y(p)b) \right\}.
\]

For every \( p \in P \), let \( P_{y(p)} = \{ q \mid q \in P, y(q) = y(p) \} \) denote the set of points with the same label as the label assigned to \( p \).

Let \( (U, D, V) \) be the \( f \)-SVD of \( (P, w) \) with respect to \( f_{\text{SVM}} \). Then, claims (i) – (ii) hold as follows:
(i) for every \( p \in P \), the sensitivity of \( p \) with respect to the query space \( (P, w, \mathbb{R}^{d+1}, f_{\text{SVM}}) \) is bounded by

\[
\begin{align*}
  s(p) &= \max \left\{ \frac{9w(p)}{\sum_{q \in P_{x}(p)} w(q)}, \frac{2w(p)}{\sum_{q \in P \setminus P_{y}(p)} w(q)} \right\} + \frac{13w(p)}{4 \sum_{q \in P_{y}(p)} w(q)} \\
  &+ \frac{125 \sum_{q \in \bar{P}} w(q)}{4\lambda} \cdot \left( w(q) \|U(p)\|_{2}^{2} + \frac{w(p)}{\sum_{q \in \bar{P}} w(q)} \right),
\end{align*}
\]

(ii) and the total sensitivity is bounded by

\[
\sum_{p \in P} s(p) \leq 25 + \frac{\sum_{p \in P_{x}(p)} w(p)}{\sum_{q \in P \setminus P_{y}(p)} w(q)} + \frac{\sum_{q \in \bar{P}} w(q)}{4\lambda} \cdot (d + 2).
\]

**Proof.** Put \( p \in P \), let \( P_{y}(p) \) denote the subset of points from \( P \) with same label as \( p \), i.e., \( P_{y}(p) = \{ q \mid q \in P, y(q) = y(p) \} \), let \( \gamma = 0.4 \), and let \( X = \{ (x, b) \mid x \in \mathbb{R}, b \in \mathbb{R}, \|x\|_{2} \leq \gamma \} \). We have that

\[
\sup_{(x, b) \in \mathbb{R}^{d+1}} \frac{w(p)f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q)f_{\text{SVM}}(q, (x | b))} \leq \sup_{(x, b) \in X} \frac{w(p)f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q)f_{\text{SVM}}(q, (x | b))} + \sup_{(x, b) \in \mathbb{R}^{d+1} \setminus X} \frac{w(p)f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q)f_{\text{SVM}}(q, (x | b))}, \tag{32}
\]

**Proof of Claim (i)**. By the above inequality, in order to bound the sensitivity of a point \( p \in P \), we can bound the term in \( \text{(32)} \). For that, we first proceed to bound the left hand side of \( \text{(32)} \).

**Handling queries from \( X \).** We observe that for every \( q \in P \)

\[
-\gamma \leq -\gamma \|q\|_{2} - \|x\|_{2} \|q\|_{2} \leq q^{T}x \leq \|x\|_{2} \|q\|_{2} \leq \gamma \|q\|_{2} \leq \gamma,
\]

where the first and last inequalities hold since \( \|q\|_{2} \leq 1 \) for every \( q \in P \), the second and fifth inequalities hold since \( \|x\|_{2} \leq \gamma \), and the third and forth inequalities hold by Cauchy-Schwartz’s inequality.

In addition,

\[
\sup_{(x, b) \in X} \frac{w(p)f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q)f_{\text{SVM}}(q, (x | b))} \leq \sup_{(x, b) \in X} \frac{w(p)\|x\|_{2}^{2}}{\sum_{q \in P} w(q)\|x\|_{2}^{2}} + \sup_{(x, b) \in X} \frac{w(p) \max \{0, 1 + p^{T}x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^{T}x + y(q)b\}},
\]

where the inequality holds by plugging \( a_{1} := \frac{w(p)}{\sum_{q \in P} w(q)}\|x\|_{2}^{2}, r_{1} := w(p) \max \{0, 1 + p^{T}x + y(p)b\}, a_{2} := \frac{1}{2}\|x\|_{2}^{2} \) and \( r_{2} := \sum_{q \in P} w(q)f_{\text{SVM}}(q, (x, b)) - \frac{1}{2}\|x\|_{2} \)

into Claim [21].

Bounding the rightmost term of \( \text{(34)} \) requires carefully checking three cases:
(a) If $y(p)b > 0$, then we have
\[
\frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\}} \leq \frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0 + q^T x + y(q)b\}}
\]
\[
= \frac{w(p) (1 + p^T x + y(p)b)}{\sum_{q \in P} w(q) (1 + q^T x + y(q)b)}
\]
\[
\leq \frac{w(p) (1 + p^T x) + y(p)w(p)b}{\sum_{q \in P} w(q) (1 + q^T x) + y(q)w(q)b},
\]
where the first inequality holds since $P_{y(p)} \subseteq P$, the equality follows from combining the assumption that $y \in (0, 1)$ and (33), and the last inequality holds by combining the fact that $1 + q^T x \geq 0$ for every $q \in P_{y(q)}$, the assumption of the case, and the result of plugging $a_1 := w(p) (1 + p^T x)$, $r_1 := w(p)y(p)b$, $a_2 := \sum_{q \in P} w(q) (1 + q^T x)$ and $r_2 := \sum_{q \in P} w(q)y(q)b$ into Claim 21.

We also have
\[
\frac{w(p) \max \{0, 1 + p^T x\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x\}} \leq \frac{w(p) \max \{0, 1 + \|p\|\}}{\sum_{q \in P} w(q) \max \{0, 1 - \|q\|\}}
\]
\[
\leq \frac{w(p) (1 + \gamma)}{\sum_{q \in P} w(q) (1 - \gamma)},
\]
where the first inequality holds by (33), and the second inequality follows from the assumption that for every $q \in P$, $\|q\| \leq 1$.

(b) If $y(p)b \in [-\gamma, 0]$, then
\[
\frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\}} \leq \frac{w(p) \max \{0, 1 + \|p\| + \gamma\}}{\sum_{q \in P} w(q) \max \{0, 1 - \|q\| - \gamma\}} \leq \frac{w(p) \max \{0, 1 + 2\gamma\}}{\sum_{q \in P} w(q) (1 - 2\gamma)},
\]
where the first inequality holds since $|y(p)b| \leq \gamma$ and $P_{y(p)} \subseteq P$, and the second inequality holds since $\gamma \in (0, \frac{1}{2})$.

(c) Otherwise, we have $-\gamma > y(p)b$, which means that for every $q \in P$ such that $y(q) \neq y(p)$, we have $\gamma < y(q)b$.

Thus,
\[
\frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x - \gamma\}} \leq \frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P \setminus P_{y(p)}} w(q) \max \{0, 1 + q^T x - \gamma\}}
\]
\[
\leq \frac{w(p) \max \{0, 1 + \|p\| + \gamma\}}{\sum_{q \in P \setminus P_{y(p)}} w(q) \max \{0, 1 - \|q\| + \gamma\}}
\]
\[
\leq \frac{w(p)}{\sum_{q \in P \setminus P_{y(p)}} w(q)},
\]
where the first inequality holds since $P \setminus P_{y(p)} \subseteq P$, the second inequality follows from (33), and the last inequality holds by the assumption that $\|q\| \leq 2$ for every $q \in P$. 

29
Since $\gamma \in (0, \frac{1}{2})$, we have $1 \leq \frac{1+\gamma}{1-\gamma} \leq \frac{1+2\gamma}{1-2\gamma}$, and by that we get

$$\frac{w(p) \max \{0, 1 + \gamma\}}{\sum_{q \in P} w(q) (1 - \gamma)} \leq \frac{w(p) \max \{0, 1 + 2\gamma\}}{\sum_{q \in P} w(q) (1 - 2\gamma)}.$$  \hspace{1cm} (36)

Combining the cases above with (36), yields that

$$\sup_{(x,b) \in X} \frac{w(p) f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q) f_{\text{SVM}}(q, (x | b))} \leq 2 \max \left\{ \frac{w(p) (1 + \gamma)}{\sum_{q \in P} w(q) (1 - \gamma)}, \frac{w(p) (1 + 2\gamma)}{\sum_{q \in P \setminus P_{y(p)} \setminus P_{x(p)}} w(q)} \right\}.  \hspace{1cm} (37)$$

**Handling queries from $\mathbb{R}^d \times \mathbb{R} \setminus X$.** Put $(x, b) \in \mathbb{R}^d \times \mathbb{R} \setminus X$, and consider the following case analysis:

(a) If $|b| \leq 9 \|x\|_2$, then we note the following:

(A) For every $q \in P$, $x \in \mathbb{R}^d$ and $\beta \geq 0$ we have $|q^T \beta x| = \beta \cdot |q^T x|$.

(B) Since $|q^T x|$ is a convex function for every $q \in P$, it also holds that $\sum_{q \in P} w(q) |q^T x|^2$ is convex due to the fact that sum of convex functions is also convex.

(C) The level set $\left\{ x \in \mathbb{R}^d \mid \sum_{q \in P} w(q) |q^T x|^2 \leq 1 \right\}$ is convex and is centrally symmetric.

(D) In addition, for every unit vector $y \in \mathbb{R}^d$

$$|q^T y|^2 \leq \|q\|^2 = \sum_{i=1}^d |q^T e_i|^2,$$

where the inequality holds by Cauchy’s inequality and the equality holds by properties of norm.

By combining (A), (B), (C), (D) and the result of substituting $c := \lambda$, $N := 2 \sum_{q \in P} w(q)$ and $\gamma := 0.4$ into Lemma 28, we get that we can plug

- $f(p, (x | b)) := f_{\text{SVM}}(p, (x | b))$, $g(p, (x | b)) := \frac{1}{\lambda} \|p^T x\|$, and $h(p, (x | b)) := \max \left\{ \frac{1}{\lambda} \frac{b}{\|x\|_N} \right\}$,

- $\alpha := 1$,

- $c_1 := \frac{\lambda^2}{(1+10\gamma) \sum_{q \in P} w(q)}$ and $c_2 := 2$,

- $c_i := 1$ for every $i \in [3, 4]$,

- $c_5 := 2$,

- $v_j := e_j$ for every $j \in [d]$ where $e_j$ denotes the vector with a 1 in the $j$th coordinate and 0’s elsewhere,

and $c := 1$,

into Lemma 5, to obtain that $f_{\text{SVM}} \in \mathcal{F}$ with respect to any $x \in \mathbb{R}^d \times \mathbb{R} \setminus X$ and the sensitivity $p$ is bounded by

$$\frac{w(p) f_{\text{SVM}}(p, (x | b))}{\sum_{q \in P} w(q) f_{\text{SVM}}(q, (x | b))} \leq \frac{(1 + 10\gamma) \sum_{q \in P} w(q)}{\gamma^2 \lambda \sum_{q \in P} w(q)} \left( \frac{2w(p)}{\sum_{q \in P} w(q)} + \sum_{j=1}^d |U(p)^T e_j|^2 \right).$$ \hspace{1cm} (38)

with respect to any query in $\mathbb{R}^d \times \mathbb{R} \setminus X$. 30
(b) If \( y(p)b \geq 9 \|x\|_2 \) then we have that

\[
\sum_{q \in P} w(q) f_{\text{SVM}}(q, (x \mid b)) \leq \frac{w(p) \|x\|_2^2}{\sum_{q \in P} w(q) \|x\|_2^2} + \frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\}}
\]

\[
= \frac{w(p)}{\sum_{q \in P} w(q)} + \frac{w(p) \max \{0, 1 + p^T x + y(p)b\}}{\sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\}}
\]

where the inequality holds by plugging \( a_1 := \frac{w(p) \|x\|_2^2}{\sum_{q \in P} w(q)}, \)

\( r_1 := w(p) \max \{0, 1 + p^T x + y(p)b\}, \ a_2 := \frac{1}{2} \|x\|_2^2 \) and \( r_2 = \sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\} \) into Claim \( 21 \) and the equality holds since \( \|x\|_2 \geq \gamma. \)

In addition, we observe that

\[
\sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\} \leq \sum_{q \in P} w(q) \max \{0, 1 + q^T x + y(q)b\}
\]

\[
\leq \sum_{q \in P} w(q) \max \{0, 1 - \|x\|_2 + y(q)b\}
\]

\[
= \sum_{q \in P} w(q) \max \{0, 1 + \frac{10w(q)}{9}b\}
\]

\[
\leq \sum_{q \in P} w(q) \left(1 + \frac{10w(q)}{9}b\right)
\]

\[
= \sum_{q \in P} w(q) \left(1 + \frac{8w(q)}{9}b\right)
\]

\[
\leq \sum_{q \in P} w(q) + 4 \sum_{q \in P} w(q) = \frac{9w(p)}{4} \sum_{q \in P} w(q)
\]

where the first inequality holds since \( P \subseteq P_{w(p)}, \) the second inequality holds since \( \|q\|_2 \leq 1 \) for every \( q \in P, \) both the third inequality and the equality is by the assumption of the case, and the last inequality follows from plugging \( a_1 := w(p), r_1 := w(p) \frac{11w(p)}{10} b, a_2 := \sum_{q \in P} w(q), \) and \( r_2 := \frac{8}{9} \sum_{q \in P} w(q) \) into Claim \( 21. \)

Combining (39) and (40), yields that

\[
\sum_{q \in P} w(q) f_{\text{SVM}}(q, (x \mid b)) \leq \frac{13w(p)}{4} \sum_{q \in P} w(q)
\]

(c) Otherwise, i.e., \( y(p)b \leq -9 \|x\|_2, \) we have that for every \( q \in P_{w(p)} \)

\[
\max \{0, 1 + q^T x + y(q)b\} \leq \max \{0, 1 + \|x\|_2 + y(q)b\} = 0,
\]

where the first inequality holds since \( \|q\|_2 \leq 1 \) for every \( q \in P, \) and the fact that \( 1 - 8\gamma < 0. \)

Thus,

\[
\sum_{q \in P} f_{\text{SVM}}(p, (x \mid b)) \leq \sum_{q \in P} \frac{w(p) f_{\text{SVM}}(p, (x \mid b))}{w(q) f_{\text{SVM}}(q, (x \mid b))} = \frac{w(p)}{\sum_{q \in P} w(q)}.
\]

(41)
By combining the three cases above, we obtain that

$$s(p) \leq 2 \max \left\{ \frac{w(p) (1 + 2\gamma)}{\sum_{q \in P_{y(p)}} w(q)/\left(1 - 2\gamma\right)}, \frac{w(p)}{\sum_{q \in P_{y(p)}} w(q)} \right\} + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)} \right\} + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)} + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)}$$

(42)

Claim (ii) now holds as

$$\sum_{j=1}^{d} |U(p)e_j|^2 = ||U(p)||_2^2,$$

where the equality follows from the definition of $e_j$ for every $j \in [d]$.

**Proof of Claim (ii)** As for the total sensitivity, we first note that that

$$\sum_{q \in P_{y(p)}} \frac{w(q)}{q' \in P_{y(p)}} w(q') = 1.$$

(43)

In addition, by Lemma 5

$$\sum_{q \in P} w(q) \|U(q)\|_2^2 \leq d.$$

(44)

Hence,

$$\sum_{q \in P} s(q) \leq \sum_{q \in P} 2 \left( \frac{w(p) (1 + 2\gamma)}{\sum_{q \in P_{y(p)}} w(q)/\left(1 - 2\gamma\right)} + \frac{w(p)}{\sum_{q \in P_{y(p)}} w(q)} \right) + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)} + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)} + \frac{13w(p)}{4 \sum_{q \in P_{y(p)}} w(q)}$$

(45)

$$\leq 4 \frac{(1 + 2\gamma)}{1 - 2\gamma} + \frac{\sum_{q \in P_{y(p)}} w(q)}{\sum_{q' \in P_{y(p)}} w(q')} + \frac{\sum_{q \in P_{y(p)}} w(q)}{\sum_{q \in P_{y(p)}} w(q')} + \frac{13}{2}$$

(46)

where (45) holds since both arguments of the max operator at (42) are non-negative and their sum exceeds the max among them, and (46) holds by combining (43) with (44).

Claim (ii) now holds by substituting $\gamma = 0.4$. 

**Corollary 9** (Support Vector Machines). Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points such that for every $p \in P$, $\|p\| \leq 1$. Let $y : P \to \{1, -1\}$ be a labelling function, $\lambda \geq 1$ be a regularization parameter such that for every $p \in P$, $x \in \mathbb{R}^d$, and $b \in \mathbb{R}$, $f_{\text{SVM}} \left( p, \begin{bmatrix} x \\ b \end{bmatrix} \right) = \lambda \max \{0, 1 - (p^T x + y(p) \cdot b)\}$ + $\frac{1}{2\lambda} \|x\|^2$. Let $P_+ = \{p | p \in P, y(p) = 1\}$, $P_- = P \setminus P_+$, $C = \frac{|P_+|}{|P_-|}.$

Then, there exists an algorithm that gets the set $P$ as an input, and returns a pair $(S, v)$, such that (i) with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for $P$ with respect to $f_{\text{SVM}}$, and (ii) the size of the coreset is $|S| \leq O \left( \frac{1}{\varepsilon} \left( \frac{dn}{\lambda} + \frac{\tilde{C}^2 + 1}{\lambda} \right) \left( d \log \left( \frac{dn}{\lambda} + \frac{\tilde{C}^2 + 1}{\lambda} \right) + \log \frac{1}{\delta} \right) \right)$. 

32
Proof. First, observe that by Lemma, the total sensitivity of the query space $(P, w, \mathbb{R}^d, f_{\text{SVM}})$ is bounded by $O\left(\left(\frac{d C + \tilde{C}^2 + 1}{C}\right)\right)$. Let $s(p)$ be the upper bound on the sensitivity of each point $p \in P$ as in Lemma and let $t = \sum_{q \in P} s(q)$. Let $S$ be an i.i.d random sample of size $O\left(\left(\frac{d C + \tilde{C}^2 + 1}{C}\right) \left(d \log \left(\frac{d C + \tilde{C}^2 + 1}{C}\right) + \log \frac{1}{\delta}\right)\right)$, where each point $p \in P$ is sampled with probability $\frac{s(p)}{t}$, and let $v(p) = \frac{w(p) d}{s(p)}$. Hence by Theorem, we get that with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for the query space $(P, w, \mathbb{R}^d, f_{\text{SVM}})$ of size $|S| \in O\left(\left(\frac{d C + \tilde{C}^2 + 1}{C}\right) \left(d \log \left(\frac{d C + \tilde{C}^2 + 1}{C}\right) + \log \frac{1}{\delta}\right)\right)$. \hfill \Box

## F.6 Proof of Corollary

First, we provide the following definitions.

**Definition 30 (Induced matrix norm).** Let $z \in [1, \infty]$. Then the $\ell_z$ induced norm for any matrix $A \in \mathbb{R}^{d \times d}$, is defined by,

$$
\|A\|_z = \max_{\|x\|_z = 1} \|Ax\|.
$$

**Definition 31 (SVD factorization of a square matrix).** Let $A \in \mathbb{R}^{d \times d}$ be matrix. The SVD factorization of $A$ is defined to be

$$
A = U \Sigma V^T,
$$

where $U \in \mathbb{R}^{d \times d}$ is an orthogonal matrix, $\Sigma \in \mathbb{R}^{d \times d}$ is a diagonal matrix of non-negative entries in a descending order, i.e. for every $i, j \in [d]$ such that $i \leq j$, $\Sigma_{i,i} \geq \Sigma_{j,j}$, and finally $V \in \mathbb{R}^{d \times d}$ is an orthogonal matrix.

**Lemma 32.** For every vector $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$ there is $j \in [d]$ such that

$$
\|x\|_1 \leq \left(\frac{3d}{2} - 1\right) |x_1 + x_j|.
$$

Equality holds for $x = (1, -3, \cdots, -3)$ and every $j \in [d]$, i.e., the bound is tight.

**Proof.** Without loss of generality, assume that $x_1 \in \{0, 1\}$. Otherwise, we divide $x$ by $x_1$.

Let $j \in \arg \max_{i \in [d]} |x_1 + x_i|$, $m \in \arg \max_{i \in [d]} |x_i|$, $a = \max_{i \in [d]} x_i$, and $b = \max_{i \in [d]} -x_i$. The proof is by case analysis of three cases: (i) $x_1 = 0$, (ii) $x_1 = 1$ and $|1 + x_j| = 1 + a$, and (iii) $x_1 = 1$ and $|1 + x_j| = b - 1$.

There are no other cases, since if $x_1 = 1$,

$$
|x_1 + x_j| = |1 + x_j| = \max \{1 + x_j, -x_j - 1\} = \max \{1 + a, b - 1\}.
$$

We observe that:

(i) If $x_1 = 0$,

$$
\|x\|_1 = \sum_{i=1}^d |x_i| \leq d |x_m| = d |x_1 + x_m| \leq \left((3d/2) - 1\right) |x_1 + x_m|,
$$

where the last inequality is by assumption $d \geq 2$, otherwise the lemma is trivial.

(ii) If $|1 + x_j| = 1 + a$ and $x_1 = 1$, then for every $i \in [d]$,

$$
|x_i| = |1 + x_i - 1| \leq |1 + x_i - 1| + 1 \leq |1 + x_j + 1 = 2 + a, \quad (47)
$$

where the first inequality is by the triangle inequality, and the last equality holds by the assumption of the case. Hence

$$
\frac{\|x\|_1}{|x_1 + x_j|} \leq \frac{1 + (d - 1)(a + 2)}{1 + a}. \quad (48)
$$

33
The right hand side is decreasing with $a$ since the numerator of its derivative is
$$(d-1)(1+a) - (1+(d-1)(a+2)) = -d < 0.$$ Its maximum is achieved at $a \geq x_1 = 1$ by the assumption $x_1 = 1$ of this case. By this and (48),
$$\frac{\|x\|_1}{|x_1 + x_j|} \leq \frac{1 + (d-1)(a + 2)}{1 + a} \leq \frac{1 + 3(d-1)}{2} = \frac{3d}{2} - 1.$$ (iii) $|1 + x_j| = b - 1$ and $x_1 = 1$. For every $i \in [d]$ we thus have $|x_i| \leq |1 + x_j| + 1 = b$, similarly to (47). Hence
$$\frac{\|x\|_1}{|x_1 + x_j|} \leq \frac{1 + b(d-1)}{b - 1}.$$ The right hand side is decreasing with $b$ since the numerator of its derivative is
$$(d-1)(b - 1) - (1 + b(d-1)) = -d < 0.$$ Its maximum is achieved at $b = |1 + x_j| + 1 \geq |1 + x_1| + 1 = 3$, where the first equality is by the assumptions of Case (iii). By this and (49),
$$\frac{\|x\|_1}{|x_1 + x_j|} \leq \frac{1 + b(d-1)}{b - 1} \leq \frac{1 + 3(d-1)}{2} = \frac{3d}{2} - 1.$$}

\[\square\]

**Claim 33.** Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix, and let $A = U\Sigma V$ be the SVD factorization of $A$ (see Definition 31). Then for every $i \in [2, d]$, $\|A\|_2 \leq \|A(V_1 + V_i)\|_2$, where $V_j$ denotes the $j$th column of $V$ for every $j \in [d]$.

**Proof.** First, put $i \in [2, d]$, and note that by (46), we have that $\|A\|_2 = \|AV_1\|_2$.

For every $j \in [d]$, let $e_j$ denotes the vector with a 1 in the $j$th coordinate and 0’s elsewhere. We observe that
$$\|AV_1 + AV_i\|_2^2 = \|AV_1\|_2^2 + 2V_i^T A^T AV_1 + \|AV_i\|_2^2.$$ By orthogonality of $V$ and $U$,
$$V_i^T A^T AV_i = V_i^T \Sigma \Sigma^T U^T U \Sigma \Sigma^T V_i = V_i^T \Sigma \Sigma^T V_i = e_i^T \Sigma \Sigma^T e_i = \Sigma_{1,1} \Sigma_{i,i} e_1 e_i^T = 0,$$ where the first equality holds by Definition 31, the second equality is by orthogonality of $U$, the third equality is by orthogonality of $V$, the forth equality holds since $\Sigma$ is a diagonal matrix and the last equality holds by definition of $e_j$ for every $j \in [d]$.

Combining (50) and (51), yields that
$$\|AV_1 + AV_i\|_2 = \sqrt{\|AV_1\|_2^2 + \|AV_i\|_2^2} \geq \|AV_1\|_2 = \|A\|_2.$$ \[\square\]

**Lemma 34.** Let $(P, w, \mathbb{R}^d, f_{\text{Rest}_x})$ be a query space as in Definition 11 such that for every $x \in \mathbb{R}^d$, and $p \in P$, the loss function $f_{\text{Rest}_x}$ is defined to be
$$f_{\text{Rest}_x}(p, x) = \min \left\{ \|p^T x\|, \|x\|_2 \right\}.$$ Let $g_{\text{Rest}_x} \in F$ such that for every $x \in \mathbb{R}^d$ and $p \in P$, $g_{\text{Rest}_x}(p, x) = |p^T x|$. Let $(U, D, V)$ be the $F$-SVD of $P$ with respect to $g_{\text{Rest}_x}$. Let $\gamma = \max \left\{ 1, 2d \frac{\|x\|_2}{\|Dv\|_2} \right\}$. Then claims (i) – (ii) hold as follows:
(i) For every \( p \in P \), its sensitivity with respect to the query space \((P, w, \mathbb{R}^d, f_{\text{Res}^z})\) is bounded by
\[
    s(p) = w(p) \min \left\{ \|U(p)\|_2, d^{\frac{1}{2} - \frac{1}{4}} \left\| \left(DV^T\right)^{-1} \right\|_2 \right\},
\]
(ii) and the total sensitivity is bounded by
\[
    \sum_{p \in P} s(p) \leq 4\gamma d^{2 + \frac{3}{4} - \frac{1}{4}}.
\]

**Proof.** First, we observe that the level set \( X_{\text{Res}^z} \) (see Definition [1]) is contained in the level set
\[
    L = \left\{ x : x \in \mathbb{R}^d, \sum_{p \in P} w(p)f_{\text{Res}^z}(p, x) \leq 1 \right\}.
\]
By Theorem III of [36], the Löwner ellipsoid which contains the level set \( X_{\text{Res}^z} \) will also contain the level set \( L \), when setting the dilation factor, i.e., \( \alpha \) to \( \gamma \sqrt{d} \). In other words,
\[
    \frac{1}{\sqrt{d}} E \subseteq X_{\text{Res}^z} \subseteq L \subseteq \sqrt{d}\gamma E,
\]
where \( E \) denotes the Löwner ellipsoid of the level set \( X_{\text{Res}^z} \). Since \( L \) is contained in the ellipsoid \( \sqrt{d}\gamma E \), and contains the ellipsoid \( \frac{1}{\sqrt{d}} E \), using similar arguments to those established at the proof of Lemma [16], we obtain that there exists a diagonal matrix \( D \in \mathbb{R}^{d \times d} \) and an orthogonal matrix \( V \in \mathbb{R}^{d \times d} \) such that for every \( x \in \mathbb{R}^d \),
\[
    \left\| D'V^Tx \right\|_2 \leq \sum_{q \in P} w(q)f_{\text{Res}^z}(q, x) \leq \gamma \sqrt{d} \left\| D'V^Tx \right\|_2,
\]
where \( D' := \frac{1}{\sqrt{d}} D \).

With this, we proceed to bound the sensitivity of each point \( p \in P \).

**Proof of Claim [i].** Put \( p \in P \), and let \( U(q) := (VD')^{-1} \) for every \( q \in P \). Observe that,
\[
\begin{align*}
    &\sup_{x \in \mathbb{R}^d, f_{\text{Res}^z}(p, x) > 0} \frac{w(p)f_{\text{Res}^z}(p, x)}{\sum_{q \in P} w(q)f_{\text{Res}^z}(q, x)} \\
    &\leq \sup_{x \in \mathbb{R}^d, f_{\text{Res}^z}(p, x) > 0} \frac{w(p)f_{\text{Res}^z}(p, x)}{\left\| D'V^Tx \right\|_2} \\
    &\leq \sup_{x \in \mathbb{R}^d, f_{\text{Res}^z}(p, x) > 0} w(p) \min \left\{ \frac{\|U(p)^TD'V^Tx\|_2}{\left\| D'V^Tx \right\|_2}, \frac{\left\| x \right\|_2}{\left\| D'V^Tx \right\|_2} \right\} \\
    &\leq w(p) \min \left\{ \|U(p)\|_2, d^{\frac{1}{2} - \frac{1}{4}} \left\| \left(DV^T\right)^{-1} \right\|_2 \right\}
\end{align*}
\]
where the first inequality is by (52), the equality is by definition of \( f_{\text{Res}^z} \), and the last inequality follows from combining Lemma [18] with the fact that \( \frac{D'V^Tx}{\left\| D'V^Tx \right\|_2} \) is a unit vector and \( (DV^T)^{-1} D'V^T = I_d \).

**Proof of Claim [ii].** In order to bound the total sensitivity, we first let \( \beta_z = d^{\frac{1}{2} - \frac{1}{4}} \), \( M \in \mathbb{R}^{d \times d} \) be an orthogonal matrix that corresponds to the matrix \( V \) of the SVD factorization of \( (DV^T)^{-1} \) (See Definition [31]), and let \( M_i \) denote the \( i \)-th column of \( M \) for every \( i \in [d] \). Thus,
\[
\begin{align*}
    &\min \left\{ \|U(p)\|_2, \beta_z \left\| (DV^T)^{-1} \right\|_2 \right\} = \min \left\{ \|MU(p)\|_2, \beta_z \left\| (DV^T)^{-1} M_i \right\|_2 \right\} \\
    &\leq \min \left\{ \|MU(p)\|_1, \beta_z \left\| (DV^T)^{-1} M_i \right\|_2 \right\} \\
    &\leq \min \left\{ 2d \|U(p)^T(M_i + M_j)\|_1, \beta_z \left\| (DV^T)^{-1} (M_i + M_j) \right\|_2 \right\} \\
    &\leq 2d \min \left\{ \|U(p)^T(M_i + M_j)\|_1, \left\| (DV^T)^{-1} (M_i + M_j) \right\|_2 \right\}.
\end{align*}
\]
where the equality holds by definition of $M$, the first inequality holds by Lemma 18, the second inequality holds by Lemma 34, and by Claim 33, and the last inequality follows from the fact that $\beta_z \leq 2d$.

By combining (53) and (54), we have that
\[
s(p) \leq 2d^{1+\frac{1}{2}-\frac{1}{z}} \|w(p)\| \sum_{j=1}^{d} \min \left\{ \left\| U(p)^T (M_{s_1} + M_{s_2}) \right\|, \left\| (D'V^T)^{-1}(M_{s_1} + M_{s_2}) \right\| \right\}. \tag{55}
\]
where the inequality follows from invoking Lemma 18.

Summing (55) over every $p \in P$, we obtain that
\[
\sum_{p \in P} s(p) \leq 2\beta_{d^2+\frac{1}{z}-\frac{1}{e}} \sum_{j=1}^{d} \|M_1 + M_j\|_2 \leq 4\beta_{d^2+\frac{1}{z}-\frac{1}{e}},
\]
where the first inequality is by (52) and the second inequality holds since $\|x + y\|_2 \leq 2$ for any pair of unit vectors $x, y \in \mathbb{R}^d$.

**Corollary 10 (Outlier resistant functions).** Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, and let $f_{\text{Reg}} : P \times \mathbb{R}^d \rightarrow [0, \infty)$ be loss function such that for every $x \in \mathbb{R}^d$, and $p \in P$, $f_{\text{Reg}}(p, x) = \min \left\{ \|p^T x\|, \|x\|_2 \right\}$.

Then, there exists an algorithm that gets the set $P$ as an input, and returns a pair $(S, v)$, such that (i) with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for $P$ with respect to $f_{\text{Reg}}$, and (ii) the size of the coreset is $O\left(\frac{\beta_{d^2+\frac{1}{z}-\frac{1}{e}}}{\varepsilon^2} \left(d \log \left(\gamma d^2+\frac{1}{z}-\frac{1}{e} \right) + \log \left(\frac{1}{\delta} \right) \right) \right)$, where $\gamma$ is defined in the proof.

**Proof.** First, observe that by Lemma 34 the total sensitivity of the query space $(P, w, \mathbb{R}^d, f_{\text{Reg}})$ is bounded by $O\left(\gamma d^2+\frac{1}{z}-\frac{1}{2}\right)$. Let $s(p)$ be the upper bound on the sensitivity of each point $p \in P$ as in Lemma 34, and let $t = \sum_{q \in P} s(q)$. Let $S$ be an i.i.d random sample of size $O\left(\frac{\beta_{d^2+\frac{1}{z}-\frac{1}{e}}}{\varepsilon^2} \left(d \log \left(\gamma d^2+\frac{1}{z}-\frac{1}{e} \right) + \log \left(\frac{1}{\delta} \right) \right) \right)$, where each point is sampled with probability $\frac{s(p)}{t}$, and let $v(P) = \frac{w(P)}{s(P)}$. Hence by Theorem 1, we get that with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for the query space $(P, w, \mathbb{R}^d, f_{\text{Reg}})$.

\[\]

**G “Easy” examples covered by our framework**

**G.1 $\ell_2$-Regression for $z \in [1, \infty)$**

**Lemma 35.** Let $z \in [1, \infty)$, $(P, w, \mathbb{R}^d, f_{\ell_z})$ be a query space, such that for every $x \in \mathbb{R}^d$ and $p \in P$ the loss function $f_{\ell_z} : P \times \mathbb{R}^d \rightarrow [0, \infty)$ is defined to be $f_{\ell_z}(p, x) = \|p^T x\|^z$. Let $(U, D, V)$ be the $f$-SVD of $(P, w)$ with respect to $f_{\ell_z}$ (see Definition 4). Then, claims (i)–(ii) hold as follows:

(i) for every $p \in P$, the sensitivity of $p$ with respect to the query space $(P, w, \mathbb{R}^d, f_{\ell_z})$ is bounded by
\[
s(p) \leq \begin{cases} \frac{\|w(p)\| U(p)\|^z}{\sqrt{d^2 w(p) \|U(p)\|^z}} & z \in [1, 2] \\ \frac{d^{z+1}}{d} & z = 2 \\ \frac{d^{z+1}}{d} & z \in \mathbb{R} \setminus [1, 2] \end{cases}
\]

(ii) and the total sensitivity is bounded by
\[
\sum_{p \in P} s(p) \leq \begin{cases} d^{z+1} & z \in [1, 2] \\ d & z \in [1, 2] \setminus \{2\} \\ d^{z+1} & z = 2 \\ d & z = 2 \setminus \{1\} \end{cases}
\]

**Proof.** Note the following:
As for the sum of sensitivities, Claim (ii) follows from Lemma 5.

This satisfies (i) as

\[ \varepsilon, \delta \]

Let \( \varepsilon, \delta \in (0, 1) \), and let \( (S, v) \) be the output of a call to CORESET \( (P, w, f_{\ell_1}, \varepsilon, \delta) \). Then, with probability at least \( 1 - \delta \), \( (S, v) \) is an \( \varepsilon \)-coreset for the query space \( (P, w, \mathbb{R}^d, f_{\ell_1}) \), and the size of the coreset is

\[ |S| \in \begin{cases} O \left( \frac{d^{\frac{d+1}{2}}}{\varepsilon^2} \left( d \log (d^{\frac{d+1}{2}}) + \log \left( \frac{1}{\delta} \right) \right) \right) & z \in [1, 2] \\ O \left( \frac{d^2}{\varepsilon^2} \left( d \log (d) + \log \left( \frac{1}{\delta} \right) \right) \right) & z = 2 \\ O \left( \frac{d^{d+1}}{\varepsilon^2} \left( d \log (d^{d+1}) + \log \left( \frac{1}{\delta} \right) \right) \right) & z \in (2, \infty) \end{cases} \]
Proof. First, observe that by Lemma 35, the total sensitivity is bounded by $t := \begin{cases} \frac{d^2 + 1}{2} & z \in [1, 2) \\ d & z = 2 \\ d^{2+1} & \text{otherwise} \end{cases}$.

Plugging $s(p)$ for every $p \in P$ from Lemma 35, $t := t$, $\varepsilon := \varepsilon$ and $\delta := \delta$ into Theorem 6 yields that with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset of size $O \left( \frac{1}{\varepsilon^2} (d \log (t) + \log \left( \frac{1}{\varepsilon} \right)) \right)$.

### G.2 Least squared errors

**Lemma 37.** Let $(P, w, \mathbb{R}^d, f_{LSE})$ be a query space, such that for every $x \in \mathbb{R}^d$ and $p \in P$, the loss function $f_{LSE}$ is defined to be $f_{LSE}(p, x) = \|p - x\|_2^2$. Let $P' = \left\{ p' = \begin{cases} \|p\|_2^2 & \|p\|_2^2 \leq 2p^T y + \|y\|_2, \text{ which enables us to rewrite the problem by reformulating the query space and the input space (R}^d \text{ and P respectively). Let } X' = \left\{ \begin{bmatrix} 1 \\ -x \\ \|x\|_2^2 \end{bmatrix} \right| x \in \mathbb{R}^d \right\}.$ Then, we obtain that for every $x \in X'$

$$\frac{w(p) f_{LSE}(p, x)}{\sum_{q \in P} w(q) f_{LSE}(q, x)} = \frac{w(p) \|p'^T x\|_2^2}{\sum_{q \in P} w(q) \|q'^T y\|_2^2} \leq \sup_{y \in \mathbb{R}^d} \frac{w(p) \|p'^T y\|_2^2}{\sum_{q \in P} w(q) \|q'^T y\|_2^2},$$

where the second inequality is by rewriting the cost function and setting $y \in X'$ and the last inequality follows from sup operator.

Finally, the upper bound on the sensitivity of each point $p \in P$ and an upper bound on the total sensitivity follows from plugging $P', \mathbb{R}^{d+2}$ as the query space, and $z := 1$ into Corollary 35.

**Corollary 38.** Let $(P, w, \mathbb{R}^d, f_{LSE})$ be a query space, such that for every $x \in \mathbb{R}^d$, and $p \in P$, the loss function $f_{LSE}$ is defined to be

$$f_{LSE}(p, x) = \|p - x\|_2^2.$$
By construction of $P'$, it holds that for every $p' \in S'$ and $x' = \begin{pmatrix} 1 \\ -x \end{pmatrix} / \|x\|_2^2$ where $x \in \mathbb{R}^d$,

$$v(p') g_{\text{lse}}(p', x') = v(p') \|p' x'\| = v(p) \|p - x\|_2^2 = v(p) \|p - x\|_2^2 = v(p) f_{\text{lse}}(p, x).$$

Thus we obtain that for every $x \in \mathbb{R}^d$

$$\left| \sum_{p \in P} w(p) \|p - x\|_2^2 - \sum_{p \in S} w(p) \|p - x\|_2^2 \right| \leq \varepsilon \sum_{p \in P} w(p) \|p - x\|_2^2,$$

hold with probability at least $1 - \delta$, i.e., $(S, v)$ is an $\varepsilon$-coreset for the query space $(P, w, \mathbb{R}^d, f_{\text{lse}})$. \qed

**H Experimental setup**

**Preprocessing step.** We applied a standardization step, i.e., each input point has zero mean and unit variance. In addition, specifically for the problem of SVM and Logistic regression, the points were normalized such that the maximal norm of a point in the dataset will be 1.

**Faster algorithms for computing the $f$-SVD** Problems which can be reduced to the $\ell_2$-regression problem, are easier to deal with, since the $f$-SVD can be computed using the SVD factorization which is can be computed in $O(n^2 d)$, e.g., we showed that both logistic regression and SVM can be reduced to $\ell_2$-regression as discussed in Lemma 25 and Lemma 28.

As for our aforementioned problems, we shown a reduction to $\ell_1$ regression, which using [15], we can compute the $f$-SVD in roughly $O(n d + \text{poly}(d))$ time (worst case scenario).

Note that [15] can accelerate the computation time of the $f$-SVD if the problem can be reduced to $\ell_z$ regression for any $z \geq 1$, due to the fact that it computes an approximated Löwner ellipsoid using randomized algorithm. For other problems, the time needed for computing the $f$-SVD is mentioned at Theorem 6.