# 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 $\ell_z$-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 $f : P \times \mathbb{R}^d \to [0, \infty)$, 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 [1] and

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

got increasing attention [3, 4, 5, 6, 8, 20, 26, 27, 38] over the recent years; see surveys in [23, 47, 50]. Informally, coreset is a small weighted subset of the input points (unlike e.g. sketches, or dimension-reduction techniques) that approximates the loss of the input set $P$ for *every* feasible query $x$, up to a provable bound of $1 \pm \varepsilon$ for a given error parameter $\varepsilon \in (0, 1)$. The size of the coreset is usually polynomial in $1/\varepsilon$ but independent or near-logarithmic in the size of the input. Since such a coreset approximates every query (and not just the optimal one), it supports constraint optimization, and the above computation models using merge-and-reduce trees; see details in [23]. Moreover, coresets may be computed in time that is near-linear in the input, even for NP-hard optimization problems. Existing heuristic or inefficient algorithms may then be applied many times on the small coreset to obtain improved or faster models in such cases.

Example coresets in machine learning include *SVM* [33, 57, 58, 59, 60], $\ell_z$-regression [18, 21, 54], 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, 43, 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, $\ell_z$-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 $\frac{h(p,x)^z}{\sum_{q\in P} h(q,x)^z} \leq \frac{2}{n}$.*

*(iv) The set* $\mathcal{X}_g = \left\{ x \in \mathbb{R}^d \middle| \sum_{p \in P} g(p,x)^{\max\{1,z\}} \leq 1 \right\}$ *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* $B(0_d, y)$ *denotes a ball of radius* $y > 0$*, centered at* $0_d$*.*

*We denote by* $\mathcal{F}$*, the union of all functions* $f$ *with the above properties.*

**The intuition behind Definition 1.** Properties (i)-(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 (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 (i)-(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 [35], [56], 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 [35] 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 [56], 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 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).

**$\ell_z$-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 $\ell_z$-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 $\ell_z$-regression for any $z \in (0, 1)$.

**Outlier resistant functions (similar to $M$-estimators).** Among such functions, is the $\ell_z$-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 $\ell_z$-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 $\ell_z$-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 $\sqrt{d} \left\| DV^T x \right\|_2$ and from below by $\left\| DV^T x \right\|_2$. In some sense, this can be thought of as a $\left(1 - 1/\sqrt{d}\right)$-coreset (or a sketch) since it approximates the total loss for any query in $\mathbb{R}^d$ up to a multiplicative factor of $\left(1 - 1/\sqrt{d}\right)$. 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 $\sqrt{\sum_{p \in P} \left| p^T x \right|^2}$ 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 $\{1, \cdots, n\}$. 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 $A^T A = AA^T = I_d$. Finally, throughout the paper, vectors are addressed as column vectors, and $\mathbb{1} : \mathbb{R}^d \to 1$ is a weight function.

In what follows, we provide formally 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 $x \in \mathbb{R}^d$, $\left| 1 - \frac{\sum_{q \in S} v(q) f(q,x)}{\sum_{p \in P} f(p,x)} \right| \leq \varepsilon$.*

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

**Theorem 3** (Restatement of Theorem 5.5 in [7])**.** *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*

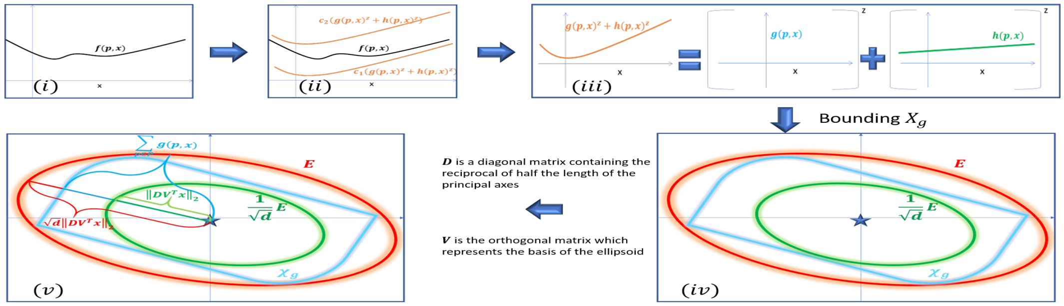



Figure 1: **How to compute $f$-SVD**: (i) Given a set $P \subseteq \mathbb{R}^2$, and a function $f : P \times \mathbb{R}^2 \to [0, \infty)$, (ii) find a function which can bound $f(p, \cdot) \times \mathbb{R}^2 \to [0, \infty)$ from above and below for every $p \in P$, (iii) decompose this function into two functions $g$ and $h$ where for every $p, q \in P$ and $x \in \mathbb{R}^2$, $g(p, \cdot)$ is a convex function (e.g., $g(p, x) = |p^T x|^4$), and $h(p, x) \approx h(q, x)$ (e.g., $h(p, x) = \|x\|_1 + 10$), here $z = 1$. (iv) Since $g$ is convex, we find the Löwner ellipsoid $E$ which contains $\mathcal{X}_g = \{x \in \mathbb{R}^2 | \sum_{p\in P} g(p, x) \le 1\}$, and the contracted ellipsoid $1/\sqrt{d}E$ is inscribed in $\mathcal{X}_g$. Following this, we compute a diagonal matrix $D \in [0, \infty)^{2\times 2}$ and an orthogonal matrix $V \in \mathbb{R}^{2\times 2}$ such that $E = \{x \in \mathbb{R}^2 | \left\|DV^T x\right\|_2 \le 1\}$. (v) By properties of the Löwener ellipsoid, we show that the total loss of $g$ (cyan line) for any query $x \in \mathbb{R}^2$ is in the range $[\left\|DV^T x\right\|_2, \sqrt{d}\left\|DV^T x\right\|_2]$ (green and red lines). When combined with the bounds on $f$, we obtain an upper bound on the sensitivity of each point in $P$ and on the total sensitivity.

*a function such that $s(p)$ is an upper bound on the sensitivity of $p$. Let $t = \sum_{p\in P} s(p)$ and $d'$ be the* VC dimension *of the quadruple $(P, \mathbb{1}, \mathbb{R}^d, f)$; see Definition 15. Let $c \ge 1$ be a sufficiently large constant, $\varepsilon, \delta \in (0, 1)$, and let $S$ be a random sample of $|S| \ge \frac{ct}{\varepsilon^2}\left(d' \log t + \log \frac{1}{\delta}\right)$ i.i.d points from $P$, such that every $p \in P$ is sampled with probability $s(p)/t$. Let $v(p) = \frac{t}{s(p)|S|}$ for every $p \in S$. Then, with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for $P$ with respect to $f$.*

## 5 Coreset for near-convex loss functions

For brevity purposes, proofs of the technical results have been omitted from this manuscript; we refer the reader to the supplementary material for the proofs. In addition, for simplicity of notation, we assume that the weight of each point in the input set is 1, while in the supplementary material, we handle the general case where each point may have any nonnegative weight. We also discuss generalized versions of Definition 1 and Definition 4.

### 5.1 Bounding the sensitivity

The following provides the generalization of the well-conditioned basis of [21], which will be used to bound the sensitivities.

**Definition 4** ($f$-SVD of $P$)**.** *Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, $f \in \mathcal{F}$ be a near-convex loss function (see Definition 1), and let $g, h, c_1, z$ be defined as in the context of Definition 1 with respect to $f$. Let $D \in [0, \infty)^{d\times d}$ be a diagonal matrix, and let $V \in \mathbb{R}^{d\times d}$ be an orthogonal matrix, such that for every $x \in \mathbb{R}^d$, $c_1 \left(\left(\left\|DV^T x\right\|_2\right)^z + \sum_{p\in P} h(p, x)^z\right) \le \sum_{p\in P} f(p, x)$, and let $\alpha \in \Theta\left(\sqrt{d}\right)$ such that for every $x \in \mathbb{R}^d$, $\sum_{p\in P} g(p, x)^{\max\{1,z\}} \le \left(\alpha \left\|DV^T x\right\|_2\right)^{\max\{1,z\}}$. Define $U : P \to \mathbb{R}^d$ such that $U(p) = (VD)^{-1} p$ for every $p \in P$. The tuple $(U, D, V)$ is the $f$-SVD of $P$.*

**Note** that (i) such factorization exists for any set of points $P$ and any near-convex loss function $f : P \times \mathbb{R}^d \to [0, \infty)$ satisfying Definition 1, and (ii) the matrix $VD$ 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 16 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 $f \in \mathcal{F}$. 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 $f \in \mathcal{F}$ be a near-convex loss function as in Definition 1. Let $g, h, c_1, c_2, z$ be defined as in the context of Definition 1 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 $p \in P$, $g\left(p, (DV^T)^{-1}y\right)^z \leq c \sum_{j=1}^{O(d)} g\left(p, (DV^T)^{-1}v_j\right)^z$. Then, for every $p \in P$, the sensitivity of $p$ is bounded by $s(p) \leq \frac{2c_2}{c_1 n} + \frac{cc_2}{c_1}\sum_{j=1}^{O(d)}\left(g\left(p, \left(DV^T\right)^{-1} v_j\right)\right)^z$, and the total sensitivity is bounded by $\sum_{p \in P} s(p) \in \frac{2c_2}{c_1} + \frac{cc_2}{c_1}\max\left\{n^{1-z}, 1\right\}\alpha^z O(d)$.*

**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 $\left(P, \mathbb{1}, \mathbb{R}^d, f\right)$; 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 $f : P \times \mathbb{R}^d \to [0, \infty)$, and a sample size $m \geq 1$.
**Output:** A pair $(S, v)$ that satisfies Theorem 6.

1 Set $d'$ := the VC dimension of quadruple $\left(P, \mathbb{1}, \mathbb{R}^d, f\right)$ `// See Definition 15`
2 Set $g$ and $\{z, c_1, c_2\}$ to be a function and a set of real positive numbers respectively, satisfying Property (i) and (ii) of Definition 1 with respect to $f$
3 Set $c > 0$ and $\{v_1, \cdots, v_d\}$ to be 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 $s(p) := \frac{cc_2}{c_1}\sum_{j=1}^{d} g\left(p, \left(DV^T\right)^{-1} v_j\right)^z + \frac{2c_2}{c_1 n}$
`// the bound of the sensitivity of` $p$ `as in Lemma 5`
7 Set $t := \sum_{p \in P} s(p)$
8 Set $\tilde{c} \geq 1$ to be a sufficiently large constant `// Can be determined from Theorem 6`
9 Pick an i.i.d sample $S$ of $m$ points from $P$, where each $p \in P$ is sampled with probability $\frac{s(p)}{t}$.
10 set $v : \mathbb{R}^d \to [0, \infty]$ to be a weight function such that for every $q \in S$, $v(q) = \frac{t}{s(q) \cdot m}$.
11 **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 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 $\left(P, f, \mathbb{R}^d\right)$. Let $t = \frac{2c_2}{c_1} + \frac{cc_2}{c_1}\max\left\{n^{1-z}, 1\right\}\alpha^z d$, $m \in O\left(\frac{t}{\varepsilon^2}\left(d' \log(t) + \log\left(\frac{1}{\delta}\right)\right)\right)$, 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 $O\left(T(n, d) d^4 \log\left(\frac{R}{r}\right)\right)$, 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, $\mathbf{nnz}(P)$ denotes the total number of nonzero entries in the set $P$, $\tilde{\boldsymbol{C}}$ denotes the ratio between the number of positive and negative labeled points (in practice, it's a constant number), $\boldsymbol{\lambda} = \sqrt{n}$ is the given regularization parameter for the problems, $\boldsymbol{\gamma} \geq 1$ is defined as in Corollary 10, $\boldsymbol{\varepsilon}$ is the error parameter, and $\boldsymbol{\delta}$ is the probability of failure.

| Problem type | Coreset's size | Construction time[1] |
|---|---|---|
| Logistic regression | $O\left(\frac{d\sqrt{n}}{\varepsilon^2}\left(d\log(d\sqrt{n}) + \log\left(\frac{1}{\delta}\right)\right)\right)$ | $O\left(nd^2\right)$ |
| $\ell_z$-Regression for $z \in (0,1)$ | $O\left(\frac{n^{1-z}d^{\frac{z}{2}+1}}{\varepsilon^2}\left(d\log\left(n^{1-z}d^{\frac{z}{2}+1}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ | $O\left(\mathrm{nnz}(P)\log n + d^{O(1)}\right)$ |
| SVM | $O\left(\frac{d\sqrt{n} + \frac{\tilde{C}^2+1}{\tilde{C}}}{\varepsilon^2}\left(d\log\left(d\sqrt{n} + \frac{\tilde{C}^2+1}{\tilde{C}}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ | $O\left(nd^2\right)$ |
| Restricted $\ell_z$-regression | $O\left(\frac{\gamma d^{2+\left\|\frac{1}{2}-\frac{1}{z}\right\|}}{\varepsilon^2}\left(d\log\left(\gamma d^{2+\left\|\frac{1}{2}-\frac{1}{z}\right\|}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ | $O\left(\mathrm{nnz}(P)\log n + d^{O(1)}\right)$ |

**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 $b \in \mathbb{R}$,* $f_{\text{LOG}}\left(p, \begin{bmatrix} x \\ b \end{bmatrix}\right) = \frac{1}{\lambda}\ln\left(1 + e^{p^T x + y(p)\cdot b}\right) + \frac{1}{2n}\|x\|_2^2$.

*Let $\varepsilon, \delta \in (0,1)$ be an error parameter and a probability of failure respectively, $m \in O\left(\frac{dn}{\lambda\varepsilon^2}\left(d\log\left(\frac{dn}{\lambda}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$, 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 $\varepsilon$-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 $f_{\text{NC}\ell_z} : P \times \mathbb{R}^d$ be a loss function such that for every $x \in \mathbb{R}^d$, and $p \in P$,* $f_{\text{NC}\ell_z}(p,x) = \left|p^T x\right|^z$.

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

We now show how our framework can be used to compute an $\varepsilon$-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 labelling function, $\lambda \geq 1$ be a regularization parameter such that for every $p \in P$, $x \in \mathbb{R}^d$, and $b \in \mathbb{R}$,* $f_{\text{SVM}}\left(p, \begin{bmatrix} x \\ b \end{bmatrix}\right) = \lambda \max\left\{0, 1 - \left(p^T x + y(p)\cdot b\right)\right\} + \frac{1}{2n}\|x\|_2^2$*. Let $P_+ = \{p | p \in P, y(p) = 1\}$, $P_- = P \setminus P_+$, $\tilde{C} = \frac{|P_+|}{|P_-|}$.*

*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 $\varepsilon$-coreset for $P$ with respect to $f_{\text{SVM}}$, and (ii) the size of the coreset is $|S| \in O\left(\frac{1}{\varepsilon^2}\left(\frac{dn}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right)\left(d\log\left(\frac{dn}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right) + \log\frac{1}{\delta}\right)\right)$.*

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

*Then, there exists an algorithm that gets the set $P$ as an in 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 $f_{\text{RES}\ell_z}$, and (ii) the size of the coreset is $O\left(\frac{\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|}}{\varepsilon^2}\left(d\log\left(\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$, 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 be dealt with in term of coreset construction time due to the existence of randomized algoritm of computing the Löwner ellipsoid by [15]; see Section H for detailed description.

# 7 Experimental Results

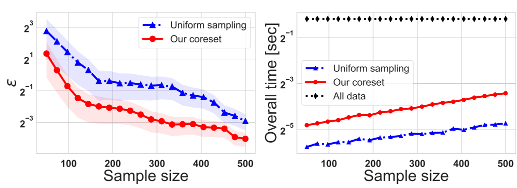
In what follows we evaluate our coreset against uniform sampling on real-world datasets, with respect to the SVM problem, Logistic regression problem and $\ell_z$-regression problem for $z \in (0, 1)$. Additional details of our setup can be found at Section H of the Supplementary Material.

**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.

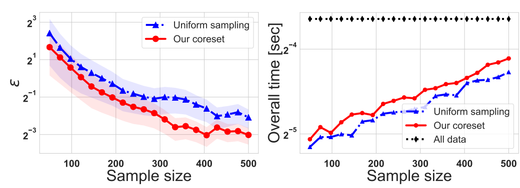
**Datasets.** The following datasets were used for our experiments mostly from UCI machine learning repository [22]: (i) **HTRU** [22] — $17,898$ radio emissions of the Pulsar star each consisting of $9$ features. (ii) **Skin** [22] — $245,057$ random samples of R,G,B from face images consisting of $4$ 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 records where each record is 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–2f, we have chosen $20$ sample sizes, starting from $50$ till $500$, at Figures 2g–2h, 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 $\left(\sum_{p\in P} f\left(p, x^*\right)\right) / \left(\min_{x\in\mathbb{R}^d} \sum_{p\in P} f(p,x)\right) - 1$. The results were averaged across $40$ trials, while the shaded regions correspond to the standard deviation.

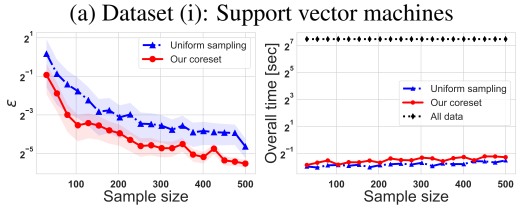
**Evaluation against prior work.** We can not have a fair comparison between our coreset to prior coresets for Logistic regression[47, 56] due to the fact that our formulation of the problem is different. As for support vector machines, we compared our efficacy against [60], the same way that we have 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 3.



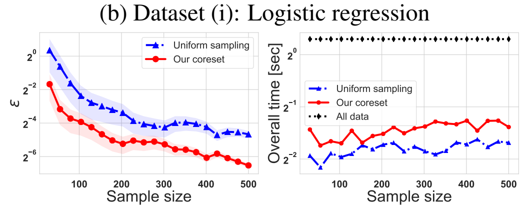
(a) Dataset (i): Support vector machines



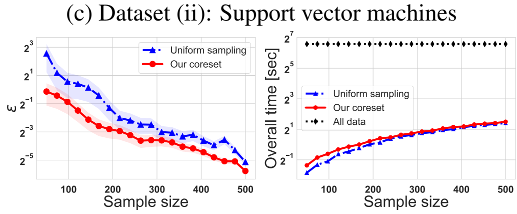
(b) Dataset (i): Logistic regression



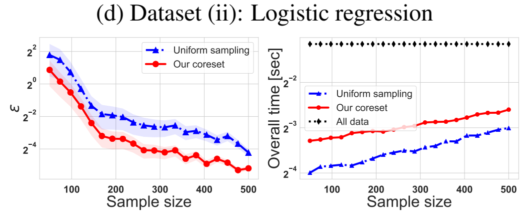
(c) Dataset (ii): Support vector machines



(d) Dataset (ii): Logistic regression

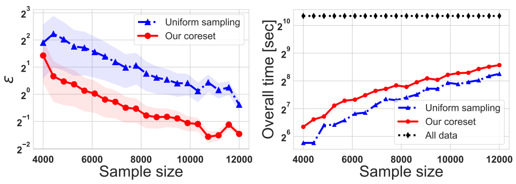


(e) Dataset (iii): Support vector machines

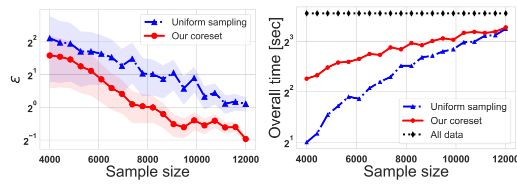


(f) Dataset (iii): Logistic regression

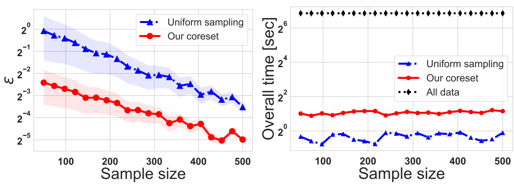
Figure 2: Experimental results



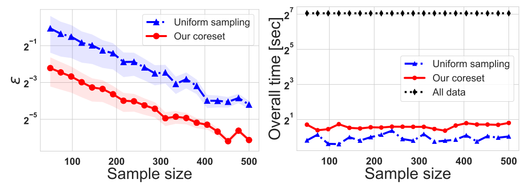

(g) Dataset (iv): Support vector machines




(h) Dataset (iv): Logistic regression

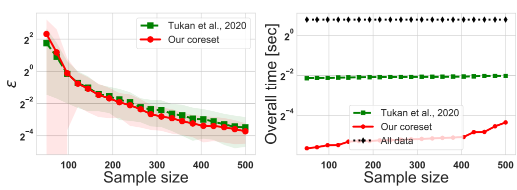



(i) Dataset (v): $\ell_{0.5}$-regression

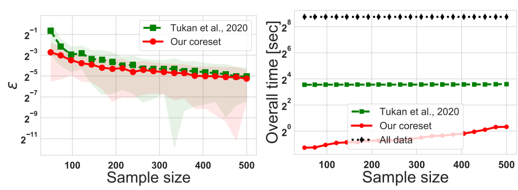



(j) Dataset (v): $\ell_{0.8}$-regression

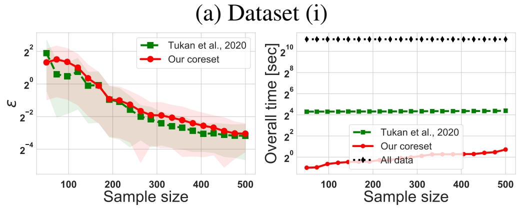
Figure 2: Experimental results



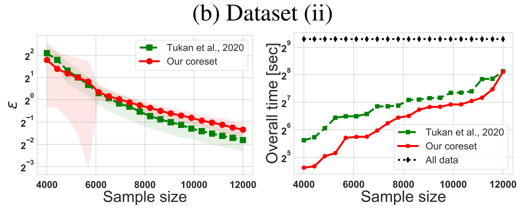

(a) Dataset (i)




(b) Dataset (ii)

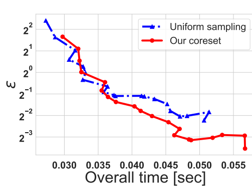



(c) Dataset (iii)

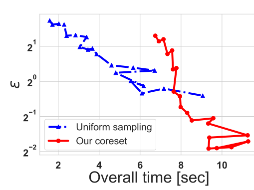



(d) Dataset (iv)

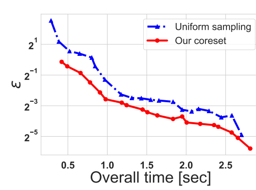
Figure 3: Comparison against prior work in the context of SVMs



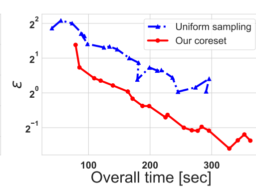

(a) Dataset (i): LR




(b) Dataset (iv): LR




(c) Dataset (iii): SVM




(d) Dataset (iv): SVM

Figure 4: Relative error as a function of the total running time. Here, LR stands for Logistic regression.

## 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 Löwner ellipsoid into the sensitivity bound. This is crucial step for generalizing the framework towards a much broader family of functions, e.g., loglog-Lipschitz functions [29]. We are aware that for $\ell_z$-regression problems where $z \geq 1$, 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*, $\ell_z$-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


[1] P. Agarwal, S. Har-Peled, and K. Varadarajan. Approximating extent measures of points. *Journal of the ACM*, 51(4):606–635, 2004.

[2] O. Bachem, M. Lucic, and A. Krause. Coresets for nonparametric estimation-the case of dp-means. In *ICML*, pages 209–217, 2015.

[3] O. Bachem, M. Lucic, and A. Krause. Scalable k-means clustering via lightweight coresets. In *KDD'18 Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1119–1127. ACM, 2018.

[4] O. Bachem, M. Lucic, and S. Lattanzi. One-shot coresets: The case of k-clustering. In A. Storkey and F. Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 784–792, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR.

[5] M. Bădoiu and K. L. Clarkson. Optimal core-sets for balls. *Computational Geometry*, 40(1):14–22, 2008.

[6] M.-F. F. Balcan, S. Ehrlich, and Y. Liang. Distributed $k$-means and $k$-median clustering on general topologies. In *Advances in Neural Information Processing Systems*, pages 1995–2003, 2013.

[7] V. Braverman, D. Feldman, and H. Lang. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

[8] V. Braverman, S. H.-C. Jiang, R. Krauthgamer, and X. Wu. Coresets for ordered weighted clustering. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 744–753, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[9] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[10] K. L. Clarkson. Subgradient and sampling algorithms for l 1 regression. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 257–266. Society for Industrial and Applied Mathematics, 2005.

[11] K. L. Clarkson. Coresets, sparse greedy approximation, and the frank-wolfe algorithm. *ACM Transactions on Algorithms (TALG)*, 6(4):63, 2010.

[12] K. L. Clarkson, R. Wang, and D. P. Woodruff. Dimensionality reduction for tukey regression. *arXiv preprint arXiv:1905.05376*, 2019.

[13] K. L. Clarkson and D. P. Woodruff. Sketching for m-estimators: A unified approach to robust regression. In *Proceedings of the twenty-sixth annual ACM-SIAM symposium on Discrete algorithms*, pages 921–939. SIAM, 2014.

[14] K. L. Clarkson and D. P. Woodruff. Input sparsity and hardness for robust subspace approximation. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 310–329. IEEE, 2015.

[15] K. L. Clarkson and D. P. Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM (JACM)*, 63(6):1–45, 2017.

[16] M. B. Cohen, S. Elder, C. Musco, C. Musco, and M. Persu. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 163–172, 2015.

[17] M. B. Cohen and R. Peng. L p row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 183–192. ACM, 2015.

[18] M. B. Cohen and R. Peng. Lp row sampling by lewis weights. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 183–192, 2015.

[19] R. M. Corless, G. H. Gonnet, D. E. Hare, D. J. Jeffrey, and D. E. Knuth. On the lambertw function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

[20] R. Curtain, S. Im, B. Moseley, K. Pruhs, and A. Samadian. On coresets for regularized loss minimization. *arXiv preprint arXiv:1905.10845*, 2019.

[21] A. Dasgupta, P. Drineas, B. Harb, R. Kumar, and M. W. Mahoney. Sampling algorithms and coresets for \ell_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

[22] D. Dua and C. Graff. UCI machine learning repository, 2017.

[23] D. Feldman. Core-sets: Updated survey. In *Sampling Techniques for Supervised or Unsupervised Tasks*, pages 23–44. Springer, 2020.

[24] D. Feldman, M. Faulkner, and A. Krause. Scalable training of mixture models via coresets. In *Advances in neural information processing systems*, pages 2142–2150, 2011.

[25] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 569–578. ACM, 2011.

[26] D. Feldman, M. Monemizadeh, C. Sohler, and D. P. Woodruff. Coresets and sketches for high dimensional subspace approximation problems. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 630–649. SIAM, 2010.

[27] D. Feldman, G. Rossman, M. Volkov, and D. Rus. Coresets for k-segmentation of streaming data. In *NIPS*, 2014.

[28] D. Feldman, M. Schmidt, and C. Sohler. Turning big data into tiny data: Constant-size coresets for k-means, pca and projective clustering. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1434–1453. SIAM, 2013.

[29] D. Feldman and L. J. Schulman. Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 1343–1354. SIAM, 2012.

[30] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter. Efficient and robust automated machine learning. In *Advances in neural information processing systems*, pages 2962–2970, 2015.

[31] L. Gu. A coreset-based semi-supverised clustering using one-class support vector machines. In *Control Engineering and Communication Technology (ICCECT), 2012 International Conference on*, pages 52–55. IEEE, 2012.

[32] I. Guyon, K. Bennett, G. Cawley, H. J. Escalante, S. Escalera, T. K. Ho, N. Macià, B. Ray, M. Saeed, A. Statnikov, et al. Design of the 2015 chalearn automl challenge. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[33] S. Har-Peled, D. Roth, and D. Zimak. Maximum margin coresets for active and noise tolerant learning. In *IJCAI*, pages 836–841, 2007.

[34] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[35] J. Huggins, T. Campbell, and T. Broderick. Coresets for scalable bayesian logistic regression. In *Advances In Neural Information Processing Systems*, pages 4080–4088, 2016.

[36] F. John. Extremum problems with inequalities as subsidiary conditions. In *Traces and emergence of nonlinear programming*, pages 197–215. Springer, 2014.

[37] I. Jubran, M. Tukan, A. Maalouf, and D. Feldman. Sets clustering. *arXiv preprint arXiv:2003.04135*, 2020.

[38] Z. Karnin and E. Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993, 2019.

[39] J.-C. Kuang. Applied inequalities. *Shandong Science and Technology Press, Jinan, China*, 3, 2004.

[40] M. Langberg and L. J. Schulman. Universal $\varepsilon$-approximators for integrals. In *Proceedings of the twenty-first annual ACM-SIAM symposium on Discrete Algorithms*, pages 598–607. SIAM, 2010.

[41] L. Lovász. *An algorithmic theory of numbers, graphs and convexity*. SIAM, 1986.

[42] M. Lucic, O. Bachem, and A. Krause. Strong coresets for hard and soft bregman clustering with applications to exponential family mixtures. In A. Gretton and C. C. Robert, editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1–9, Cadiz, Spain, 09–11 May 2016. PMLR.

[43] M. Lucic, M. Faulkner, A. Krause, and D. Feldman. Training gaussian mixture models at scale via coresets. *The Journal of Machine Learning Research*, 18(1):5885–5909, 2017.

[44] A. Maalouf, I. Jubran, and D. Feldman. Fast and accurate least-mean-squares solvers. In *Advances in Neural Information Processing Systems*, pages 8305–8316, 2019.

[45] A. Maalouf, A. Statman, and D. Feldman. Tight sensitivity bounds for smaller coresets. *arXiv preprint arXiv:1907.01433*, 2019.

[46] C. D. Meyer. *Matrix analysis and applied linear algebra*, volume 71. Siam, 2000.

[47] A. Munteanu, C. Schwiegelshohn, C. Sohler, and D. Woodruff. On coresets for logistic regression. In *Advances in Neural Information Processing Systems*, pages 6561–6570, 2018.

[48] T. E. Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[50] J. M. Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.

[51] W. Rudin. *Functional Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1991.

[52] T. Sarlos. Improved approximation algorithms for large matrices via random projections. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 143–152. IEEE, 2006.

[53] M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means. In *International Workshop on Approximation and Online Algorithms*, pages 232–251. Springer, 2019.

[54] C. Sohler and D. P. Woodruff. Subspace embeddings for the l1-norm with applications. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pages 755–764, 2011.

[55] C. Sohler and D. P. Woodruff. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 802–813. IEEE, 2018.

[56] E. Tolochinsky and D. Feldman. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. *arXiv preprint arXiv:1802.07382*, 2018.

[57] I.-H. Tsang, J.-Y. Kwok, and J. M. Zurada. Generalized core vector machines. *IEEE Transactions on Neural Networks*, 17(5):1126–1140, 2006.

[58] I. W. Tsang, J. T. Kwok, and P.-M. Cheung. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(Apr):363–392, 2005.

[59] I. W. Tsang, J. T.-Y. Kwok, and P.-M. Cheung. Very large svm training using core vector machines. In *AISTATS*, 2005.

[60] M. Tukan, C. Baykal, D. Feldman, and D. Rus. On coresets for support vector machines. *arXiv preprint arXiv:2002.06469*, 2020.

[61] M. Tukan, A. Maalouf, and D. Feldman. Open source code for all the algorithms presented in this paper, 2019. Link for open-source code.

[62] A. V. Uzilov, J. M. Keegan, and D. H. Mathews. Detection of non-coding rnas on the basis of predicted secondary structure formation free energy change. *BMC bioinformatics*, 7(1):173, 2006.

[63] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[64] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020.

[65] D. Woodruff and Q. Zhang. Subspace embeddings and\ell_p-regression using exponential random variables. In *Conference on Learning Theory*, pages 546–567, 2013.

# 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 $x, y \in \mathbb{R}^d$*

$$f\left(p, \theta x + (1 - \theta)y\right) \leq \theta f(p, x) + (1 - \theta) f(p, y).$$

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 $x \in \mathbb{R}^d$,*

$$c_1 \left(g(p, x)^z + h(p, x)^z\right) \leq f(p, x) \leq c_2 \left(g(p, x)^z + h(p, x)^z\right).$$

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

$$c_3 b g(p, x) \leq g(p, bx) \leq c_4 b g(p, x).$$

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

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

*(iv) The set $\mathcal{X}_g = \left\{ x \in \mathbb{R}^d \middle| \sum_{p \in P} w(p)^{\max\left\{1, \frac{1}{z}\right\}} g(p, x)^{\max\{1, z\}} \leq 1 \right\}$ 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 1 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 $x \in \mathbb{R}^d$,*

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

*and*

$$\sum_{p \in P} w(p)^{\max\left\{1, \frac{1}{z}\right\}} g(p, x)^{\max\{1, z\}} \leq \left(c_4 \alpha \left\|DV^T x\right\|_2\right)^{\max\{1, z\}}.$$

*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*

$$\left|\left\{S \cap \text{ranges}(x, r) \mid x \in \mathbb{R}^d, r \geq 0\right\}\right| = 2^{|S|},$$

*where $|A|$ denotes the number of points in $A$ for every $A \subseteq \mathbb{R}^d$.*

## C Existence of $f$-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 $x \in \mathbb{R}^d$,*

$$\sum_{p \in P} w(p)^{\max\left\{1, \frac{1}{z}\right\}} g(p, x)^\beta \leq \left(c_4 \alpha \left\|DV^T x\right\|_2\right)^\beta, \tag{1}$$

*and*

$$c_1 \left(\left(c_3 \left\|DV^T x\right\|_2\right)^z + \sum_{p \in P} w(p) h(p, x)^z\right) \leq \sum_{p \in P} w(p) f(p, x). \tag{2}$$

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

**Using Löwner ellipsoid.** Let $\mathcal{X}_g = \left\{x \in \mathbb{R}^d \middle| \left(\sum_{p \in P} w(p)^{\max\left\{1, \frac{1}{z}\right\}} g(p, x)^\beta\right)^{\frac{1}{\beta}} \leq 1\right\}$. Since $f \in F$ (see Definition 13), and $g$ is convex, we have that (i) $\mathcal{X}_g$ is a convex set, and (ii) $\mathcal{X}_g$ 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

$$\frac{1}{\sqrt{d}} E \subseteq X_g \subseteq E, \tag{3}$$

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 $B(0_d, r) \subseteq \mathcal{X}_g \subseteq E$. 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) $E = \left\{y \in \mathbb{R}^d \mid \left\|DV^T y\right\|_2 \leq 1\right\}$, and (ii) $V^T DDV$ 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 (i) 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

$$\frac{1}{\alpha} y \in \frac{1}{\alpha} E \subseteq X_g, \tag{4}$$

which consequently leads to

$$\sum_{p \in P} w(p)^{\max\left\{1, \frac{1}{z}\right\}} g\left(p, \frac{1}{\alpha} y\right)^\beta \leq 1, \tag{5}$$

where the inequality holds by Property (iv) of Definition 13 with respect to $g$. Hence,

$$\begin{aligned}\sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g(p,x)^\beta &\leq \left(c_4\alpha\left\|DV^Tx\right\|_2\right)^\beta \sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g\left(p, \frac{x}{\alpha\left\|DV^Tx\right\|_2}\right)^\beta \\ &\leq \left(c_4\alpha\left\|DV^Tx\right\|_2\right)^\beta,\end{aligned} \tag{6}$$

where the first inequality is by substituting $b := \alpha\left\|DV^Tx\right\|_2$ and $x := \frac{x}{b}$ in Property (ii) of $g$ (see Definition 13), and the second inequality is by combining the fact that $y = \frac{1}{\left\|DV^Tx\right\|_2}x$ with (5).

**Proving** (2). Let $b' = \alpha\left\|DV^Tx\right\|_2$. By (5), we get that $\sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g\left(p, \frac{1}{b'}x\right)^\beta \geq 1$. 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 $\sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g\left(p, \frac{1}{b}x\right)^\beta = 1$.

By (3), we have that $\left\|DV^Tx\right\|_2 = b\left\|DV^Tz\right\|_2 \leq b$ where $z = \frac{1}{b}x$. Hence, by plugging $x := z$ and $b := b$ in Property (ii) of Definition 13, we obtain that

$$\sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g(p,x)^\beta \geq c_3 b \sum_{p\in P} w(p)^{\max\left\{1,\frac{1}{z}\right\}} g(p,z)^\beta = (c_3 b)^\beta \geq \left(c_3\left\|DV^Tx\right\|_2\right)^\beta. \tag{7}$$

By combining Property (i) 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 $B\left(0_d, \sqrt{d}R\right)$, using similar arguments to those at (3). Then there exist a point $\tilde{c}$ on the ball $B(0_d, R)$ such that on the ball $B(\tilde{c}, 2d\max\{R, 1/R\})$ there exist a set of $d+1$ vectors $\{\tilde{v}_i\}_{i=1}^{d+1}$ each of norm $2dR$ such that any unit vector $x \in \mathbb{R}^d$ is a convex combination of them. For instance, the set $\{\tilde{v}_i\}_{i=1}^{d+1}$ 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

$$g(p, Ax)^z \leq \left(\sum_{i=1}^{d+1} g(p, A\tilde{v}_i)\right)^z \leq d^z \sum_{i=1}^{d+1} g(p, A\tilde{v}_i)^z \leq \left(2c_4 d^2 R\right)^z \sum_{i=1}^{d+1} g(p, Av_i)^z,$$

where $v_i = \frac{1}{2dR}\tilde{v}_i$ for every $i \in [d+1]$.

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 F.

# 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 $2l$ of streamed points, we generate an $\varepsilon$-coreset on

**Algorithm 2:** STREAMING-CORESET$(P, w, f, l, \varepsilon, \delta)$

**Input:** A set $P \subseteq \mathbb{R}^d$ of $n$ points, a weight function $w : \mathbb{R}^d \to [0, \infty)$, a leaf size $\ell > 0$ a convex loss function function $f : \mathbb{R}^d \to [0, \infty)$, 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     $(T, v) :=$ CORESET$(Q, w, f, \frac{\varepsilon}{2 \log n}, \frac{\delta}{2 \log n})$
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             $(T_1, v_1), (T_2, v_2) :=$ 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             $(T, v) :=$ CORESET$(T_1 \cup T_2, v', f, \frac{\varepsilon}{2 \log n}, \frac{\delta}{2 \log n})$
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 3–6. Lines 7–13 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 5. Suppose that there exists some $\beta \in (0.1, 0.8)$ such that $t \in \Theta(n^\beta)$ and let $\ell \geq 2^{\frac{\beta}{1-\beta}}$. 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 $w$

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

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 $\varepsilon' = \frac{\varepsilon}{2 \log n}$ in the above inequality as done at Lines 4 and 9 of Algorithm 2, we obtain that

$$(1 + \varepsilon')^{\log n} \leq e^{\frac{\varepsilon}{2}} \leq 1 + \varepsilon, \tag{8}$$

where the second inequality holds since $\varepsilon \in [\frac{1}{\log n}, \frac{1}{2}]$.

As for the lower bound, observe that

$$(1-\varepsilon)^{\log n} \geq 1 - \varepsilon \log n,$$

where the inequality holds since $\varepsilon \in [\frac{1}{\log n}, \frac{1}{2}]$.

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 $\ell \geq \sqrt[1-\beta]{2^{\beta}}$ 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,

$$2^{\sum\limits_{j=1}^{i} \beta^j} \cdot \ell^{\beta^i} \leq \frac{2^{\sum\limits_{k=0}^{i-1} \beta^k} \cdot \ell^{\beta^{i-1}}}{2} = 2^{\sum\limits_{k=1}^{i-1} \beta^k} \cdot \ell^{\beta^{i-1}},$$

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 $\ell \geq 2^{\frac{\beta}{1-\beta}}$ as a leaf size in the merge and reduce tree, we obtain an $\varepsilon$-coreset of size poly-logarithmic in $n$. $\square$

# F 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.

## F.1 Generalization of Lemma 5

**Lemma 18** (Equivalence of norms, [51]). *Let $a, b > 0$ such that $a \leq b$. Then for every $x \in \mathbb{R}^d$,*

$$\|x\|_b \leq \|x\|_a \leq d^{\frac{1}{a} - \frac{1}{b}} \|x\|_b.$$

**Claim 19.** *[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

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

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 11) such that $f \in \mathcal{F}$ as in Definition 13. Let $g, h, c_1, c_2, c_3, c_4, c_5, z$ be defined as in the context of Definition 13 with respect to $f$, $(U, D, V)$ be the $f$-SVD of $(P, w)$, and let $\alpha \in \Theta\left(\sqrt{d}\right)$ 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 $p \in P$,*

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

*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*

$$s(p) \le \left(\frac{2c_2c_5 w(p)}{c_1 \sum_{q\in P} w(q)}\right)^z + \frac{cc_2}{c_1 c_3^{2z}} \sum_{j=1}^{O(d)} w(p) \left(g\left(p, \left(DV^T\right)^{-1} v_j\right)\right)^z,$$

*and the total sensitivity is bounded by*

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

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

$$\frac{f(p,x)}{\sum\limits_{q\in P} w(q) f(q,x)} \le \frac{f(p,x)}{c_1 \left( \left(c_3 \left\|DV^T x\right\|_2\right)^z + \sum\limits_{q\in P} w(q) h(q,x)^z \right)} \tag{9}$$

$$\le \frac{c_2 g(p,x)^z + c_2 h(p,x)^z}{c_1\left(c_3\left\|DV^Tx\right\|_2\right)^z + c_1 \sum\limits_{q\in P} w(q) h(q,x)^z} \tag{10}$$

$$\le \frac{c_2 g(p,x)^z}{c_1\left(c_3\left\|DV^Tx\right\|_2\right)^z} + \frac{c_2 h(p,x)^z}{c_1 \sum\limits_{q\in P} h(q,x)^z}, \tag{11}$$

where (9) holds by Lemma 16, (10) holds by Property (i) of Definition 1 with respect to $f$, and the last inequality follows from plugging $a_1 := c_2 g(p,x)$, $r_1 := c_2 h(p,x)$, $a_2 := c_1 c_3 \left\|DV^T x\right\|_2$ and $r_2 := c_3 \sum_{q\in P} h(q,x)$ 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,

$$\frac{1}{\left\|DV^Tx\right\|_2^z} g(p,x)^z \le \frac{1}{c_3^z} g\left(p, \frac{x}{\left\|DV^Tx\right\|_2}\right)^z \tag{12}$$

$$= \frac{1}{c_3^z} g\left(p, \left(DV^T\right)^{-1} y\right)^z \tag{13}$$

$$\le \frac{c}{c_3^z} \sum_{j=1}^{d} g\left(p, \left(DV^T\right)^{-1} v_j\right)^z, \tag{14}$$

where (12) follows from substituting $b := \frac{1}{\|DV^T x\|_2}$ and $x := \frac{x}{b}$ in Property (ii) of $f$ (see Definition 1), (13) holds since $x = (DV^T)^{-1}(DV^T)x$, and (14) is by the assumption of Lemma 5.

By combining (9)–(14) with Property (iii) of $f$, the sensitivity of $p$ is bounded by

$$s(p) \leq \frac{c_2 c_5 w(p)}{c_1 \sum_{q \in P} w(q)} + \frac{cc_2}{c_1 c_3^{2z}} \sum_{j=1}^{d} w(p) g\left(p, \left(DV^T\right)^{-1} v_j\right)^z. \tag{15}$$

As for the total sensitivity, we first observe that if $z \in (0,1)$

$$\begin{aligned}
\sum_{q \in P} \sum_{j=1}^{O(d)} w(q) g\left(q, \left(DV^T\right)^{-1} v_j\right)^z &= \sum_{j=1}^{O(d)} \sum_{q \in P} w(q) g\left(q, \left(DV^T\right)^{-1} v_j\right)^z && (16)\\
&= \sum_{j=1}^{O(d)} \sum_{q \in P} \left(w(q)^{\frac{1}{z}} g\left(q, \left(DV^T\right)^{-1} v_j\right)\right)^z && (17)\\
&\leq n^{1-z} \sum_{j=1}^{O(d)} \left(\sum_{q \in P} w(q)^{\frac{1}{z}} g\left(q, \left(DV^T\right)^{-1} v_j\right)\right)^z && (18)\\
&\leq n^{1-z} \left(c_4 \alpha\right)^z \sum_{j=1}^{O(d)} \left\| DV^T \left(DV^T\right)^{-1} v_j \right\|_2^z && (19)\\
&= n^{1-z} \left(c_4 \alpha\right)^z \sum_{j=1}^{O(d)} \|v_j\|_2 && (20)\\
&= n^{1-z} \left(c_4 \alpha\right)^z O(d), && (21)
\end{aligned}$$

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 $z := z$, $n := n$, $a_i := w(q)^{\frac{1}{z}} g\left(q, \left(DV^T\right)^{-1} v_j\right)$ 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 $DV^T \left(DV^T\right)^{-1} = I_d$, and finally (21) follows from the assumption of Lemma 5.

Similarly for the case of $z \geq 1$,

$$\sum_{q \in P} \sum_{j=1}^{O(d)} w(q) g\left(q, \left(DV^T\right)^{-1} v_j\right)^z \in \left(c_4 \alpha\right)^z O(d). \tag{22}$$

Hence, Lemma 5 holds as

$$\begin{aligned}
\sum_{p \in P} s(p) &\leq \sum_{p \in P} \frac{c_2 c_5 w(p)}{c_1 \sum_{q \in P} w(q)} + \sum_{p \in P} \frac{cc_2}{c_1 c_3^{2z}} \sum_{j=1}^{O(d)} w(p) g\left(p, \left(DV^T\right)^{-1} v_j\right)^z \\
&\in \frac{c_2 c_5}{c_1} + \frac{cc_2 c_4^z}{c_1 c_3^{2z}} \max\left\{n^{1-z}, 1\right\} \alpha^z O(d),
\end{aligned}$$

where the first inequality holds by (15), and the second inequality holds by combining (16)–(22). $\square$

### 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 $\left(P, f, \mathbb{R}^d\right)$. Let $t = \frac{2c_2}{c_1} + \frac{cc_2}{c_1} \max\left\{n^{1-z}, 1\right\} \alpha^z d$, $m \in O\left(\frac{t}{\varepsilon^2}\left(d' \log(t) + \log\left(\frac{1}{\delta}\right)\right)\right)$, 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 $O\left(T(n,d)d^4 \log\left(\frac{R}{r}\right)\right)$, 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 $x \in \mathbb{R}^d$.*

*Proof.* In algorithm 1, we first compute the sensitivity bounds $s(p)$ for every $p \in P$ with respect to the query space $(P, w, \mathbb{R}^d, f)$. This is done based on Lemma 5; See Line 6. We then sample a sufficiently large number of points based on those sensitivity bound as Theorem 3 states; See Line 9. Hence, By plugging $P, w, f, \mathbb{R}^d, \varepsilon, \delta$ and $s(p)$ for every $p \in P$ into Theorem 3, we obtain that with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset (see Definition 2) of size $|S| \in O\left(\frac{t}{\varepsilon^2}\left(d' \log t + \log\left(\frac{1}{\delta}\right)\right)\right)$.

The overall time is dominated by computing the $f$-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,*

$$\frac{a_1 + r_1}{a_2 + r_2} \leq \frac{a_1}{a_2} + \frac{r_1}{r_2}.$$

*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,*

$$\frac{\max\{a_1, a_2, \cdots, a_N\}}{\max\{b_1, b_2, \cdots, b_N\}} \leq \max\left\{\frac{a_1}{b_1}, \frac{a_2}{b_2}, \cdots, \frac{a_N}{b_N}\right\}.$$

*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,

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

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

□

**Claim 23.** *For every $z, b \in \mathbb{R}$,*

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

*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

$$1 + e^{z+b} \leq 1 + 2e^{z^2+b} + e^{z^2+2b}.$$

Applying the logarithm function on both sides of the inequality above proves Claim 23 as

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

$\square$

**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)* $\frac{1}{c}\ln\left(1 + e^{p^Tx+b}\right) + \frac{1}{2N}\|x\|_2^2 \leq \frac{4}{c}(p^Tx)^2 + 4\max\left\{\frac{1}{c}\ln\left(1 + e^b\right), \frac{1}{2N}\|x\|_2^2\right\}$,

*(ii) and* $\frac{1}{c}\ln\left(1 + e^{p^Tx+b}\right) + \frac{1}{2N}\|x\|_2^2 \geq \frac{c}{8N}\left(\frac{1}{c}(p^Tx)^2 + \max\left\{\frac{1}{c}\ln\left(1 + e^b\right), \frac{1}{2N}\|x\|_2^2\right\}\right)$.

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

**Proof of Claim (i).** By plugging $z := p^Tx$ and $b := b$ into Claim 23, we obtain that

$$\ln\left(1 + e^{p^Tx+b}\right) \leq 2\ln\left(1 + e^{\left(p^Tx\right)^2+b}\right) \leq 2\ln\left(e^{\left(p^Tx\right)^2}\left(1 + e^b\right)\right) = 2\left(p^Tx\right)^2 + 2\ln\left(1 + e^b\right), \tag{23}$$

where the second inequality holds since $e^{\left(p^Tx\right)^2} \geq 1$, and the equality follows from properties of the logarithm function.

Thus, Claim (i) holds since

$$\begin{aligned}\frac{1}{c}\ln\left(1 + e^{p^Tx+b}\right) + \frac{1}{2N}\|x\|_2^2 &\leq \frac{2}{c}\left(p^Tx\right)^2 + \frac{2}{c}\ln\left(1 + e^b\right) + \frac{1}{2N}\|x\|_2^2 \\ &\leq \frac{4}{c}\left(p^Tx\right)^2 + 4\max\left\{\frac{1}{c}\ln\left(1 + e^b\right), \frac{1}{2N}\|x\|_2^2\right\},\end{aligned}$$

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

$$\|x\|_2 \geq \left|p^Tx\right|, \tag{24}$$

which consequently leads to

$$\frac{1}{c}\ln\left(1 + e^{p^Tx+b}\right) \geq \frac{1}{c}\ln\left(1 + e^{-\|x\|_2+b}\right) \geq \frac{1}{2N}\ln\left(1 + e^{-\|x\|_2+b}\right), \tag{25}$$

where the second inequality holds since $c \leq N$.

We show that

$$\ln\left(1 + e^{-\|x\|_2+b}\right) + \|x\|^2 \geq \frac{1}{2}\ln\left(1 + e^b\right), \tag{26}$$

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 $r \in \mathbb{R}$, $q(r) = \ln\left(1 + e^{-|r|+b}\right) + r^2$.

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 $x \in \mathbb{R}^d$

$$\ln\left(1 + e^{-\|x\|_2+b}\right) + \|x\|_2^2 \geq \ln\left(1 + e^{-\|x^*\|_2+b}\right) + \|x^*\|_2^2 \geq \ln\left(1 + e^{-\|x^*\|_2+b}\right) \geq \ln\left(1 + e^{-W(1)+b}\right),$$

where the first inequality holds by the definition of $x^*$, the second inequality holds since $\|x^*\|_2^2 \geq 0$, and the last inequality follows from the observation that $\|x^*\| \in [0, W(1)]$.

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

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

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

$$\frac{1}{2N}\|x\|_2^2 \geq \frac{1}{2N}\left|p^Tx\right|^2 = \frac{c}{2N}\left(\frac{1}{c}\left|p^Tx\right|^2\right), \tag{27}$$

where the first inequality holds by (24), and the equality holds since $\frac{c}{2N} \cdot \frac{1}{c} = \frac{1}{2N}$.

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

$$\frac{1}{c}\ln\left(1+e^{p^Tx+b}\right) + \frac{1}{2N}\|x\|_2^2 \geq \frac{c}{4N}\left(\frac{1}{2c}\left|p^Tx\right|^2 + \max\left\{\frac{1}{2c}\ln\left(1+e^b\right), \frac{1}{2N}\|x\|_2^2\right\}\right).$$

□

**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 $x \in \mathbb{R}^d$,*

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

*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*

$$s(p) = \frac{32}{\lambda}\left(\frac{2w(p)}{\sum\limits_{q\in P_{y(p)}} w(q)} + w(p)\|U(p)\|_2^2\right)\sum_{q\in P_{y(p)}} w(q),$$

*(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 $q \in P_p$

$$\sup_{(x,b)\in\mathbb{R}^d\times\mathbb{R}} \frac{w(p)f_{\text{LOG}}(p,(x\mid b))}{\sum\limits_{q\in P} w(q)f_{\text{LOG}}(q,(x\mid b))} \leq \sup_{(x,b)\in\mathbb{R}^d\times\mathbb{R}} \frac{w(p)f_{\text{LOG}}(p,(x\mid b))}{\sum\limits_{q\in P_{y(p)}} w(q)f_{\text{LOG}}(q,(x\mid b))},$$

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 $\left|q^T\gamma x\right| = \gamma\left|q^Tx\right|$.

(b) Since $\left|q^T x\right|$ is convex function, it also holds that $\sum_{q \in P} \left|q^T x\right|^2$ is convex due to the fact that sum of convex functions is also convex,

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

(d) For every $x \in \mathbb{R}^d$, and $b \in \mathbb{R}$ we have that

$$\frac{w(q) \max \left\{ \frac{1}{\lambda} \ln\left(1 + e^b\right), \frac{1}{2 \sum\limits_{q \in P_{y(p)}} w(q)} \|x\|_2^2 \right\}}{\sum\limits_{\tilde{q} \in P_{y(p)}} w(\tilde{q}) \max \left\{ \frac{1}{\lambda} \ln\left(1 + e^b\right), \frac{1}{2 \sum\limits_{q \in P_{y(p)}} w(q)} \|x\|_2^2 \right\}} \leq 2 \frac{w(q)}{\sum\limits_{\tilde{q} \in P_{y(p)}} w\left(\tilde{q}\right)},$$

where the inequality holds by plugging $a_1 := w(q) \frac{1}{\lambda} \ln\left(1 + e^b\right)$, $a_2 := \frac{1}{2 \sum\limits_{q \in P_{y(p)}} w(q)} \|x\|_2^2$, $b_1 := \frac{\sum\limits_{\tilde{q} \in P_{y(p)}} w(\tilde{q})}{\lambda} \ln\left(1 + e^b\right)$, $b_2 := \frac{1}{2} \|x\|_2^2$ into Claim 22.

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

- $f(p, (x \mid b)) := f_{\text{LOG}}(p, (x \mid b))$, $g(p, (x \mid b)) := \frac{1}{\sqrt{\lambda}} \left|p^T x\right|$ and $h(p, (x \mid b)) := \max \left\{ \sqrt{\frac{1}{\lambda} \ln\left(1 + e^b\right)}, \sqrt{\frac{1}{2 \sum\limits_{q \in P_{y(p)}} w(q)} \|x\|_2^2} \right\}$, for every $p \in P$, $x \in \mathbb{R}^d$, $b \in \mathbb{R}$,
- $\alpha = 1$,
- $c_1 := \frac{\lambda}{8N}$ 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

$$s(q) = \frac{32}{\lambda} \left( 2 \frac{w(q)}{\sum\limits_{\tilde{q} \in P_{y(p)}} w\left(\tilde{q}\right)} + w(q) \sum_{j=1}^{d} \left|U(q)^T e_j\right|^2 \right) \sum_{\tilde{q} \in P_{y(p)}} w(\tilde{q}).$$

Claim (i) now holds since for every $q \in P$,

$$\sum_{j=1}^{d} \left|U(q)^T e_j\right|^2 = \|U(q)\|_2^2,$$

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

As for the total sensitivity, we have by Lemma 5,

$$\sum_{q \in P_{y(p)}} s(q) \leq \frac{32}{\lambda} \left( 2 \frac{w(q)}{\sum\limits_{\tilde{q} \in P_{y(p)}} w\left(\tilde{q}\right)} + d \right) \sum_{q \in P_{y(p)}} w(p),$$

and

$$\sum_{q \in P \setminus P_{y(p)}} s(q) \leq \frac{32}{\lambda} \left( 2 \frac{w(q)}{\sum\limits_{\tilde{q} \in P \setminus P_{y(p)}} w(\tilde{q})} + d \right) \sum_{q \in P \setminus P_{y(p)}} w(p).$$

Hence, Claim (ii) holds as

$$\sum_{q \in P} s(q) \leq \frac{32}{\lambda} (2 + d) \sum_{q \in P} w(q).$$

□

**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 $b \in \mathbb{R}$, $f_{\text{LOG}} \left( p, \begin{bmatrix} x \\ b \end{bmatrix} \right) = \frac{1}{\lambda} \ln \left( 1 + e^{p^T x + y(p) \cdot b} \right) + \frac{1}{2n} \|x\|_2^2$.*

*Let $\varepsilon, \delta \in (0, 1)$ be an error parameter and a probability of failure respectively, $m \in O\left( \frac{dn}{\lambda \varepsilon^2} \left( d \log \left( \frac{dn}{\lambda} \right) + \log \left( \frac{1}{\delta} \right) \right) \right)$, 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 $\varepsilon$-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 $t := \frac{32}{\lambda} (2 + d) \sum\limits_{q \in P} w(p)$. Hence, plugging $s(p)$ for every $p \in P$ from Lemma 26, $t := t$, $\varepsilon := \varepsilon$ and $\delta := \delta$ into Theorem 6, yields that $(S, v)$ is an $\varepsilon$-coreset of size $O\left( \frac{dC}{\lambda \varepsilon^2} \left( d \log \left( \frac{dC}{\lambda} \right) + \log \left( \frac{1}{\delta} \right) \right) \right)$. □

### F.4 Proof of Corollary 8

**Lemma 27.** *Let $(P, w, \mathbb{R}^{d+1}, f_{\text{NC}\ell_z})$ be a query space, such that for every $p \in P$ and $x \in \mathbb{R}^d$,*

$$f_{\text{NC}\ell_z}(p, x) = \left| p^T x \right|^z.$$

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

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

$$s(p) = w(p) \|U(p)\|_z^z,$$

*(ii) and the total sensitivity is bounded by*

$$\sum_{p \in P} s(p) \leq n^{1-z} d^{\frac{z}{2}+1}.$$

*Proof.* Let $g : P \to [0, \infty)$ such that for every $p \in P$ and $x \in \mathbb{R}^d$, $g(p, x) = \left| p^T x \right|$, 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 $\sum\limits_{q \in P} w(q)^{\frac{1}{z}} g(q, x)$ is convex due to the fact that sum of convex functions is also convex.

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

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

$$\left| p^T y \right|^z \leq \|p\|_2^z \leq \|p\|_z^z = \sum_{i=1}^{d} \left| p^T e_i \right|^z,$$

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

- $f(q,x) := f_{\text{NC}\ell_z}(q,x)$, $g(q,x) := \left|q^T x\right|$ 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,
- $c := 1$, and
- $z := z$

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

$$s(p) = \sum_{i=1}^{d} \left|U(p)^T e_i\right|_z^z,$$

and the total sensitivity is bounded by

$$\sum_{q \in P} s(q) \leq n^{1-z} d^{\frac{z}{2}+1}.$$

□

**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 $f_{\text{NC}\ell_z} : P \times \mathbb{R}^d$ be a loss function such that for every $x \in \mathbb{R}^d$, and $p \in P$, $f_{\text{NC}\ell_z}(p,x) = \left|p^T x\right|^z$.*

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

*Proof.* First, observe that by Lemma 27, the total sensitivity is bounded by $t := n^{1-z}d^{\frac{z}{2}+1}$. 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 $O\left(\frac{n^{1-z}d^{\frac{z}{2}+1}}{\varepsilon^2}\left(d\log\left(n^{1-z}d^{\frac{z}{2}+1}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$. □

### 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 $X = \left\{(x,b) \in \mathbb{R}^d \times \mathbb{R} \mid \|x\|_2 \geq \gamma, |b| \leq 9\|x\|_2\right\}$. Then, for every $(x,b) \in X$, claims (i) – (ii) hold as follows:*

*(i)* $\frac{\|x\|_2^2}{N} + \frac{1}{c}\max\left\{0, 1 + p^T x + b\right\} \leq \frac{2}{c}\left|p^T x\right|^2 + 2\max\left\{\frac{1}{c}, \frac{b}{c}, \frac{\|x\|_2^2}{N}\right\}$,

*(ii)* $\frac{\|x\|_2^2}{N} + \frac{1}{c}\max\left\{0, 1 + p^T x + b\right\} \geq \frac{c\gamma^2}{(1+10\gamma)N}\left(\frac{1}{c}\left|p^T x\right|^2 + \max\left\{\frac{1}{c}, \frac{b}{c}, \frac{\|x\|_2^2}{N}\right\}\right)$.

*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

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

   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,

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

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

$$\begin{aligned}\frac{1}{c}\max\left\{0, 1 + p^T x + b\right\} + \frac{\|x\|_2^2}{N} &= \frac{1}{c}\left(1 + p^T x + b\right) + \frac{\|x\|_2^2}{N} \\ &\geq \frac{c}{2N}\left(\frac{1}{c}\left|p^T x\right|^2 + \max\left\{1, b, \frac{1}{2N}\|x\|_2^2\right\}\right),\end{aligned} \tag{28}$$

where the equality holds by the assumption of this case, and the inequality holds since $\left|p^T x\right|^2 \leq \|x\|_2^2$ due to the assumption that $\|p\|_2 \leq 1$.

2. Otherwise,

$$\frac{\left|p^T x\right|^2}{c} + \max\left\{\frac{1}{c}, \frac{b}{c}, \frac{\|x\|_2^2}{N}\right\} \leq \frac{1}{c} + \frac{10\|x\|_2^2}{\gamma c}, \tag{29}$$

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,

$$\frac{1}{c}\max\left\{0, 1 + p^T x + b\right\} + \frac{\|x\|_2^2}{N} \geq \frac{\|x\|_2^2}{N}, \tag{30}$$

holds since the $\max$ term is non-negative.

Let $l = \frac{c\gamma^2}{N(1+10\gamma)}$. Observe that

$$\frac{l}{c}\left(1 + \frac{10\|x\|^2}{\gamma}\right) \leq \frac{\|x\|_2^2}{N}, \tag{31}$$

since $\|x\|_2 \geq \gamma$.

Hence, we obtain that

$$\frac{1}{c}\max\left\{0, 1 + p^T x + b\right\} + \frac{\|x\|_2^2}{N} \geq \frac{c\gamma^2}{N(1 + 10\gamma)}\left(\frac{\left|p^T x\right|^2}{c} + \max\left\{\frac{1}{c}, \frac{b}{c}, \frac{1}{2N}\|x\|_2^2\right\}\right),$$

where the inequality holds by combining (29), (30) and (31).

Combining both cases proves Claim (ii).

□

**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 $b \in \mathbb{R}$,*

$$f_{\text{SVM}}(p, (x \mid b)) = \frac{1}{2\sum_{q \in P} w(q)}\|x\|_2^2 + \lambda\max\left\{0, 1 - \left(p^T x + y(p)b\right)\right\}.$$

*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*

$$s(p) = \max\left\{ \frac{9w(p)}{\sum\limits_{q \in P_{y(p)}} w(q)}, \frac{2w(p)}{\sum\limits_{q \in P \setminus P_{y(p)}} w(q)} \right\} + \frac{13w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)}$$
$$+ \frac{125 \sum\limits_{q \in P} w(q)}{4\lambda} \cdot \left( w(q) \left\| U(p) \right\|_2^2 + \frac{w(p)}{\sum\limits_{q \in P} w(q)} \right),$$

*(ii) and the total sensitivity is bounded by*

$$\sum_{p \in P} s(p) \leq 25 + \frac{\sum\limits_{p \in P_{y(p)}} w(p)}{\sum\limits_{q \in P \setminus P_{y(p)}} w(q)} + \frac{\sum\limits_{q \in P \setminus P_{y(p)}} w(p)}{\sum\limits_{p \in P} w(q)} + \frac{125 \sum\limits_{q \in P} w(q)}{4C} \cdot (d+2)\,.$$

*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 $X = \left\{(x, b) \mid x \in \mathbb{R}, b \in \mathbb{R}, \left\| x \right\|_2 \leq \gamma \right\}$. We have that

$$\begin{aligned} &\sup_{(x,b) \in \mathbb{R}^d \times \mathbb{R}} \frac{w(p) f_{\text{SVM}}(p, (x \mid b))}{\sum\limits_{q \in P} w(q) f_{\text{SVM}}(q, (x, b))} \\ &\leq \sup_{(x,b) \in X} \frac{w(p) f_{\text{SVM}}(p, (x \mid b))}{\sum\limits_{q \in P} w(q) f_{\text{SVM}}(q, (x \mid b))} + \sup_{(x,b) \in \mathbb{R}^d \times \mathbb{R} \setminus X} \frac{w(p) f_{\text{SVM}}(p, (x \mid b))}{\sum\limits_{q \in P} w(q) f_{\text{SVM}}(q, (x \mid b))}, \end{aligned} \tag{32}$$

**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 32. For that, we first proceed to bound the left hand side of (32).

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

$$-\gamma \leq -\gamma \left\| q \right\|_2 \leq - \left\| x \right\|_2 \left\| q \right\|_2 \leq q^T x \leq \left\| x \right\|_2 \left\| q \right\|_2 \leq \gamma \left\| q \right\|_2 \leq \gamma, \tag{33}$$

where the first and last inequalities hold since $\left\| q \right\| \leq 1$ for every $q \in P$, the second and fifth inequalities hold since $\left\| x \right\|_2 \leq \gamma$, and the third and forth inequalities hold by Cauchy-Schwartz's inequality.

In addition,

$$\sup_{(x,b) \in X} \frac{w(p) f_{\text{SVM}}(p, (x \mid b))}{\sum\limits_{q \in P} w(q) f_{\text{SVM}}(q, (x \mid b))} \leq \sup_{(x,b) \in X} \frac{w(p) \left\| x \right\|_2^2}{\sum\limits_{q \in P} w(q) \left\| x \right\|_2^2} + \sup_{(x,b) \in X} \frac{w(p) \max\left\{0, 1 + p^T x + y(p) b\right\}}{\sum\limits_{q \in P} w(q) \max\left\{0, 1 + q^T x + y(q) b\right\}}, \tag{34}$$

where the inequality holds by plugging $a_1 := \frac{w(p)}{2 \sum_{q \in P} w(q)} \left\| x \right\|_2^2$, $r_1 := w(p) \max\left\{0, 1 + p^T x + y(p) b\right\}$, $a_2 := \frac{1}{2} \left\| x \right\|_2^2$ and $r_2 := \sum\limits_{q \in P} w(q) f_{\text{SVM}}(q, (x, b)) - \frac{1}{2} \left\| x \right\|_2$ into Claim 21.

Bounding the rightmost term of (34) requires carefully checking three cases:

(a) If $y(p)b > 0$, then we have

$$\frac{w(p)\max\left\{0, 1+p^T x + y(p)b\right\}}{\sum_{q\in P} w(q)\max\left\{0, 1+q^T x + y(q)b\right\}} \leq \frac{w(p)\max\left\{0, 1+p^T x + y(p)b\right\}}{\sum_{q\in P_{y(p)}} w(q)\max\left\{1+q^T x + y(q)b\right\}}$$
$$= \frac{w(p)\left(1+p^T x + y(p)b\right)}{\sum_{q\in P_{y(p)}} w(q)\left(1+q^T x + y(q)b\right)}$$
$$\leq \frac{w(p)\left(1+p^T x\right)}{\sum_{q\in P_{y(p)}} w(q)\left(1+q^T x\right)} + \frac{y(p)w(p)b}{\sum_{q\in P_{y(p)}} y(q)w(q)b}, \tag{35}$$

where the first inequality holds since $P_{y(p)} \subseteq P$, the equality follows from combining the assumption that $\gamma \in (0,1)$ 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 $a_1 := w(p)\left(1+p^T x\right)$, $r_1 := w(p)y(p)b$, $a_2 := \sum_{q\in P_{y(p)}} w(q)\left(1+q^T x\right)$ and $r_2 := \sum_{q\in P_{y(p)}} y(q)w(q)b$ into Claim 21.

We also have

$$\frac{w(p)\max\left\{0, 1+p^T x\right\}}{\sum_{q\in P} w(q)\max\left\{0, 1+q^T x\right\}} \leq \frac{w(p)\max\left\{0, 1+\gamma\left\|p\right\|_2\right\}}{\sum_{q\in P} w(q)\max\left\{0, 1-\gamma\left\|q\right\|_2\right\}}$$
$$\leq \frac{w(p)\left(1+\gamma\right)}{\sum_{q\in P} w(q)\left(1-\gamma\right)},$$

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

$$\frac{w(p)\max\left\{0, 1+p^T x + y(p)b\right\}}{\sum_{q\in P} w(q)\max\left\{0, 1+q^T x + y(q)b\right\}} \leq \frac{w(p)\max\left\{0, 1+\gamma\left\|p\right\|_2 + \gamma\right\}}{\sum_{q\in P_{y(p)}} w(q)\max\left\{0, 1-\gamma\left\|q\right\|_2 - \gamma\right\}} \leq \frac{w(p)\max\left\{0, 1+2\gamma\right\}}{\sum_{q\in P_{y(p)}} w(q)\left(1-2\gamma\right)},$$

where the first inequality holds since $|y(p)b| \leq \gamma$ and $P_{y(p)} \subseteq P$, and the second inequality holds since $\gamma \in \left(0, \frac{1}{2}\right)$.

(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,

$$\frac{w(p)\max\left\{0, 1+p^T x + y(p)b\right\}}{\sum_{q\in P} w(q)\max\left\{0, 1+q^T x - \gamma\right\}} \leq \frac{w(p)\max\left\{0, 1+p^T x + y(p)b\right\}}{\sum_{q\in P\setminus P_{y(p)}} w(q)\max\left\{0, 1+q^T x - \gamma\right\}}$$
$$\leq \frac{w(p)\max\left\{0, 1+\gamma\left\|p\right\|_2 + \gamma\right\}}{\sum_{q\in P\setminus P_{y(p)}} w(q)\max\left\{0, 1-\left\|q\right\|_2\gamma + \gamma\right\}}$$
$$\leq \frac{w(p)}{\sum_{q\in P\setminus P_{y(p)}} w(q)},$$

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 $\|q\|_2 \leq$ for every $q \in P$.

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

$$\frac{w(p) \max\{0, 1+\gamma\}}{\sum\limits_{q \in P} w(q)\,(1-\gamma)} \leq \frac{w(p) \max\{0, 1+2\gamma\}}{\sum\limits_{q \in P} w(q)\,(1-2\gamma)}. \tag{36}$$

Combining the cases above with (36), yields that

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

**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 $|b| \leq 9 \left\|x\right\|_2$, then we note the following:

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

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

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

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

$$\left|q^T y\right|^2 \leq \|q\|_2^2 = \sum_{i=1}^{d} \left|q^T e_i\right|^2,$$

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 $c := \lambda$, $N := 2 \sum\limits_{q \in P} w(q)$ and $\gamma := 0.4$ into Lemma 28, we get that we can plug

- $f(p, (x \mid b)) := f_{\text{SVM}}(p, (x \mid b))$, $g(p, (x \mid b)) := \frac{1}{\lambda} \left|p^T x\right|$, and $h(p, (x \mid b)) := \max\left\{ \frac{1}{\lambda}, \frac{b}{\lambda}, \frac{\|x\|_2^2}{N} \right\}$
- $\alpha := 1$,
- $c_1 := \frac{\lambda \gamma^2}{(1+10\gamma) \sum\limits_{q \in P} w(q)}$ 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

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

with respect to any query in $\mathbb{R}^d \times \mathbb{R} \setminus X$.

(b) If $y(p)b \geq 9 \|x\|_2$ then we have that

$$\frac{w(p) f_{\text{SVM}}(p,(x \mid b))}{\sum_{q \in P} w(q) f_{\text{SVM}}(q,(x \mid b))} \leq \frac{w(p)\|x\|_2^2}{\|x\|_2^2 \sum\limits_{q \in P} w(q)} + \frac{w(p) \max\left\{0, 1 + p^T x + y(p)b\right\}}{\sum\limits_{q \in P} w(q) \max\left\{0, 1 + q^T x + y(q)b\right\}}$$
$$= \frac{w(p)}{\sum\limits_{q \in P} w(q)} + \frac{w(p) \max\left\{0, 1 + p^T x + y(p)b\right\}}{\sum\limits_{q \in P} w(q) \max\left\{0, 1 + q^T x + y(q)b\right\}}, \tag{39}$$

where the inequality holds by plugging $a_1 := \frac{w(p)\|x\|_2^2}{2 \sum\limits_{q \in P} w(q)}$, $r_1 := w(p)\max\left\{0, 1 + p^T x + y(p)b\right\}$, $a_2 := \frac{1}{2}\|x\|_2^2$ and $r_2 = \sum\limits_{q \in P} w(q) \max\left\{0, 1 + q^T x + y(q)b\right\}$ into Claim 21, and the equality holds since $\|x\|_2 \geq \gamma$.

In addition, we observe that

$$\begin{aligned}
\frac{w(p) \max\left\{0, 1 + p^T x + y(p)b\right\}}{\sum\limits_{q \in P} w(q) \max\left\{0, 1 + q^T x + y(q)b\right\}} &\leq \frac{w(p) \max\left\{0, 1 + p^T x + y(p)b\right\}}{\sum\limits_{q \in P_{y(p)}} w(q) \max\left\{0, 1 + q^T x + y(q)b\right\}} \\
&\leq \frac{w(p) \max\left\{0, 1 + \|x\|_2 + y(p)b\right\}}{\sum\limits_{q \in P_{y(p)}} w(q) \max\left\{0, 1 - \|x\|_2 + y(q)b\right\}} \\
&\leq \frac{w(p) \max\left\{0, 1 + \frac{10y(p)}{9} b\right\}}{\sum\limits_{q \in P_{y(p)}} w(q) \max\left\{0, 1 + \frac{8y(q)}{9} b\right\}} \\
&= \frac{w(p)\left(1 + \frac{10y(p)}{9} b\right)}{\sum\limits_{q \in P_{y(p)}} w(q)\left(1 + \frac{8y(q)}{9} b\right)} \\
&\leq \frac{w(p)}{\sum\limits_{q \in P_{y(p)}} w(q)} + \frac{5w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)} \\
&= \frac{9w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)}
\end{aligned} \tag{40}$$

where the first inequality holds since $P \subseteq P_{y(p)}$, 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 $a_1 := w(p)$, $r_1 := w(p)\frac{11y(p)}{10} b$, $a_2 := \sum\limits_{q \in P} w(q)$, and $r_2 := \frac{8}{9} \sum\limits_{q \in P} w(q)$ into Claim 21.

Combining (39) and (40), yields that

$$\frac{w(p) f_{\text{SVM}}(p,(x \mid b))}{\sum_{q \in P} w(q) f_{\text{SVM}}(q,(x \mid b))} \leq \frac{13w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)}$$

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

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

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

Thus,

$$\frac{w(p) f_{\text{SVM}}(p,(x \mid b))}{\sum\limits_{q \in P} f_{\text{SVM}}(p,(x \mid b))} \leq \frac{w(p) f_{\text{SVM}}(p,(x \mid b))}{\sum\limits_{q \in P_{y(p)}} w(q) f_{\text{SVM}}(q,(x \mid b))} = \frac{w(p)}{\sum\limits_{q \in P_{y(p)}} w(q)}. \tag{41}$$

By combining the three cases above, we obtain that

$$
\begin{aligned}
s(p) \leq & 2 \max \left\{ \frac{w(p)\,(1+2\gamma)}{\sum\limits_{q \in P_{y(p)}} w(q)\,(1-2\gamma)}, \frac{w(p)}{\sum\limits_{q \in P \setminus P_{y(p)}} w(q)} \right\} + \frac{13 w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)} \\
& + \frac{(1+10\gamma) \sum\limits_{q \in P} w(q)}{\gamma^2 \lambda} \cdot \left( w(q) \left\| U(p) \right\|_2^2 + \frac{w(p)}{\sum\limits_{q \in P} w(q)} \right),
\end{aligned} \tag{42}
$$

Claim (i) now holds as

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

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

$$
\sum_{q \in P_{y(p)}} \frac{w(q)}{\sum\limits_{q' \in P_{y(p)}} w(q')} = 1. \tag{43}
$$

In addition, by Lemma 5,

$$
\sum_{q \in P} w(q) \left\| U(q) \right\|_2^2 \leq d. \tag{44}
$$

Hence,

$$
\begin{aligned}
\sum_{q \in P} s(q) \leq & \sum_{q \in P} 2 \left( \frac{w(p)\,(1+2\gamma)}{\sum\limits_{q' \in P_{y(q)}} w(q')\,(1-2\gamma)} + \frac{w(q)}{\sum\limits_{q' \in P \setminus P_{y(q)}} w(q')} \right) + \frac{13 w(p)}{4 \sum\limits_{q \in P_{y(p)}} w(q)} \qquad (45) \\
& + \frac{(1+10\gamma) \sum\limits_{q \in P} w(q)}{\gamma^2 C} \sum_{q \in P} \left( w(q) \left\| U(p) \right\|_2^2 + \frac{w(p)}{\sum\limits_{q \in P} w(q)} \right) \\
\leq & \frac{4\,(1+2\gamma)}{1-2\gamma} + \frac{\sum\limits_{q \in P_{y(p)}} w(q)}{\sum\limits_{q' \in P \setminus P_{y(p)}} w(q')} + \frac{\sum\limits_{q' \in P \setminus P_{y(p)}} w(q')}{\sum\limits_{q \in P_{y(p)}} w(q)} + \frac{13}{2} \qquad (46) \\
& + \frac{(1+10\gamma) \sum\limits_{q \in P} w(q)}{\gamma^2 \lambda} \cdot (d+2)
\end{aligned}
$$

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 $\gamma = 0.4$. $\square$

**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 $b \in \mathbb{R}$, $f_{\text{SVM}}\left(p, \begin{bmatrix} x \\ b \end{bmatrix}\right) = \lambda \max\left\{0, 1 - \left(p^T x + y(p) \cdot b\right)\right\} + \frac{1}{2n} \|x\|_2^2$. Let $P_+ = \{p | p \in P, y(p) = 1\}$, $P_- = P \setminus P_+$, $\tilde{C} = \frac{|P_+|}{|P_-|}$.*

*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 $\varepsilon$-coreset for $P$ with respect to $f_{\text{SVM}}$, and (ii) the size of the coreset is $|S| \in O\left(\frac{1}{\varepsilon^2}\left(\frac{dn}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right)\left(d \log\left(\frac{dn}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right) + \log\frac{1}{\delta}\right)\right)$.*

*Proof.* First, observe that by Lemma 29 the total sensitivity of the query space $(P, w, \mathbb{R}^d, f_{\text{SVM}})$ is bounded by $O\left(\left(\frac{dC}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right)\right)$. Let $s(p)$ be the upper bound on the sensitivity of each point $p \in P$ as in Lemma 29, and let $t = \sum_{q \in P} s(q)$. Let $S$ be an i.i.d random sample of size $O\left(\frac{1}{\varepsilon^2}\left(\frac{dC}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right)\left(d \log\left(\frac{dC}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right) + \log\frac{1}{\delta}\right)\right)$, where each point $p \in P$ is sampled with probability $\frac{s(p)}{t}$, and let $v(p) = \frac{w(p)t}{s(p)|S|}$. Hence by Theorem 3, 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 $|S| \in O\left(\frac{1}{\varepsilon^2}\left(\frac{dC}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right)\left(d \log\left(\frac{dC}{\lambda} + \frac{\tilde{C}^2+1}{\tilde{C}}\right) + \log\frac{1}{\delta}\right)\right)$.

□

### F.6 Proof of Corollary 10

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,*

$$\|A\|_z = \max_{\substack{x \in \mathbb{R}^d \\ \|x\|_z = 1}} \|Ax\|_z .$$

**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*

$$A = \mathcal{U}\Sigma\mathcal{V}^T,$$

*where $\mathcal{U} \in \mathbb{R}^{d \times d}$ 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 $\mathcal{V} \in \mathbb{R}^{d \times d}$ 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*

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

*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 $x_1 = 0$,

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

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 $i \in [d]$,

$$|x_i| = |1 + x_i - 1| \leq |1 + x_i| + 1 \leq |1 + x_j| + 1 = 2 + a, \tag{47}$$

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

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

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

$$\frac{\|x\|_1}{|x_1+x_j|} \leq \frac{1+b(d-1)}{b-1}. \tag{49}$$

The right hand side is decreasing with $b$ since the enumerator 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 $A = \mathcal{U}\Sigma\mathcal{V}$ be the* SVD factorization *of $A$ (see Definition 31). Then for every $i \in [2,d]$,*

$$\|A\|_2 \leq \|A(\mathcal{V}_1+\mathcal{V}_i)\|_2\,,$$

*where $\mathcal{V}_j$ denotes the $j$th column of $\mathcal{V}$ for every $j \in [d]$.*

*Proof.* First, put $i \in [2,d]$, and note that by [46], we have that

$$\|A\|_2 = \|A\mathcal{V}_1\|_2\,.$$

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

$$\|A\mathcal{V}_1 + A\mathcal{V}_i\|_2^2 = \|A\mathcal{V}_1\|_2^2 + 2\mathcal{V}_1^T A^T A\mathcal{V}_i + \|A\mathcal{V}_i\|_2^2\,. \tag{50}$$

By orthogonality of $\mathcal{V}$ and $\mathcal{U}$,

$$\mathcal{V}_1^T A^T A\mathcal{V}_i = \mathcal{V}_1^T\mathcal{V}\Sigma^T\mathcal{U}^T\mathcal{U}\Sigma\mathcal{V}^T\mathcal{V}_i = \mathcal{V}_1^T\mathcal{V}\Sigma^T\Sigma\mathcal{V}^T\mathcal{V}_i = e_1^T\Sigma^T\Sigma e_i = \Sigma_{1,1}\Sigma_{i,i}e_1e_i^T = 0, \tag{51}$$

where the first equality holds by Definition 31, the second equality is by orthogonality of $\mathcal{U}$, the third equality is by orthogonality of $\mathcal{V}$, 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

$$\|A\mathcal{V}_1 + A\mathcal{V}_i\|_2 = \sqrt{\|A\mathcal{V}_1\|_2^2 + \|A\mathcal{V}_i\|_2^2} \geq \|A\mathcal{V}_1\|_2 = \|A\|_2\,.$$

$\square$

**Lemma 34.** *Let $(P, w, \mathbb{R}^d, f_{\mathrm{RES}\ell_z})$ be a query space as in Definition 11, such that for every $x \in \mathbb{R}^d$, and $p \in P$, the loss function $f_{\mathrm{RES}\ell_z}$ is defined to be*

$$f_{\mathrm{RES}\ell_z}(p,x) = \min\left\{\left|p^Tx\right|, \|x\|_z\right\}.$$

*Let $g_{\mathrm{RES}\ell_z} \in \mathcal{F}$ such that for every $x \in \mathbb{R}^d$ and $p \in P$, $g_{\mathrm{RES}\ell_z}(p,x) = \left|p^Tx\right|$. Let $(U,D,V)$ be the $F$-SVD of $P$ with respect to $g_{\mathrm{RES}\ell_z}$. Let $\gamma = \max\left\{1, \frac{2\pi d^{\left|\frac{1}{2}-\frac{1}{z}\right|}}{\|DV^T\|_2}\right\}$. Then claims (i) – (ii) hold as follows:*

*(i) For every $p \in P$, its sensitivity with respect to the query space $(P, w, \mathbb{R}^d, f_{\mathrm{RES}\ell_z})$ is bounded by*

$$s(p) = w(p) \min \left\{ \|U(p)\|_2, d^{\left|\frac{1}{2} - \frac{1}{z}\right|} \left\| \left(DV^T\right)^{-1} \right\|_2 \right\},$$

*(ii) and the total sensitivity is bounded by*

$$\sum_{p \in P} s(p) \leq 4\gamma d^{2+\left|\frac{1}{2} - \frac{1}{z}\right|}.$$

*Proof.* First, we observe that the level set $\mathcal{X}_{g_{\mathrm{RES}\ell_z}}$ (see Definition 1) is contained in the level set $L = \left\{ x \middle| x \in \mathbb{R}^d, \sum\limits_{p \in P} w(p) f_{\mathrm{RES}\ell_z}(p, x) \leq 1 \right\}$. By Theorem III of [36], the Löwner ellipsoid which contains the level set $\mathcal{X}_{g_{\mathrm{RES}\ell_z}}$ will also contain the level set $L$, when setting the dilation factor, i.e., $\alpha$ to $\gamma\sqrt{d}$. In other words,

$$\frac{1}{\sqrt{d}} E \subseteq \mathcal{X}_{g_{\mathrm{RES}\ell_z}} \subseteq L \subseteq \sqrt{d}\gamma E,$$

where $E$ denotes the Löwner ellipsoid of the level set $\mathcal{X}_{g_{\mathrm{RES}\ell_z}}$. 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 $x \in \mathbb{R}^d$,

$$\left\|D'V^T x\right\|_2 \leq \sum_{q \in P} w(q) f_{\mathrm{RES}\ell_z}(q, x) \leq \gamma\sqrt{d} \left\|D'V^T x\right\|_2, \tag{52}$$

where $D' := \frac{1}{2\gamma} D$.

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

**Proof of Claim (i).** Put $p \in P$, and let $U(q) := \left(VD'\right)^{-1} q$ for every $q \in P$. Observe that,

$$\begin{aligned}
\sup_{\substack{x \in \mathbb{R}^d \\ f_{\mathrm{RES}\ell_z}(p,x) > 0}} \frac{w(p) f_{\mathrm{RES}\ell_z}(p, x)}{\sum\limits_{q \in P} w(q) f_{\mathrm{RES}\ell_z}(q, x)} &\leq \sup_{\substack{x \in \mathbb{R}^d, \\ f_{\mathrm{RES}\ell_z}(p,x) > 0}} \frac{w(p) f_{\mathrm{RES}\ell_z}(p, x)}{\|D'V^T x\|_2} \\
&= \sup_{\substack{x \in \mathbb{R}^d, \\ f_{\mathrm{RES}\ell_z}(p,x) > 0}} w(p) \min \left\{ \frac{\left|U(p)^T D'V^T x\right|}{\|D'V^T x\|_2}, \frac{\|x\|_z}{\|D'V^T x\|_2} \right\} \\
&\leq w(p) \min \left\{ \|U(p)\|_2, d^{\left|\frac{1}{2} - \frac{1}{z}\right|} \left\| \left(D'V^T\right)^{-1} \right\|_2 \right\}
\end{aligned} \tag{53}$$

where the first inequality is by (52), the equality is by definition of $f_{\mathrm{RES}\ell_z}$, and the last inequality follows from combining Lemma 18 with the fact that $\frac{D'V'^T x}{\left\|D'V'^T x\right\|_2}$ is a unit vector and $\left(D'V^T\right)^{-1} D'V^T = I_d$.

**Proof of Claim (ii).** In order to bound the total sensitivity, we first let $\beta_z = d^{\left|\frac{1}{2} - \frac{1}{z}\right|}$, $M \in \mathbb{R}^{d \times d}$ be an orthogonal matrix that corresponds to the matrix $\mathcal{V}$ of the *SVD* factorization of $\left(D'V^T\right)^{-1}$ (See Definition 31), and let $M_i$ denote the $i$th column of $M$ for every $i \in [d]$. Thus,

$$\begin{aligned}
\min \left\{ \|U(p)\|_2, \beta_z \left\| \left(D'V^T\right)^{-1} \right\|_2 \right\} &= \min \left\{ \|MU(p)\|_2, \beta_z \left\| \left(D'V^T\right)^{-1} M_{1*} \right\|_2 \right\} \\
&\leq \min \left\{ \|MU(p)\|_1, \beta_z \left\| \left(D'V^T\right)^{-1} M_{1*} \right\|_2 \right\} \\
&\leq \min \left\{ 2d \left|U(p)^T (M_1 + M_j)\right|, \beta_z \left\| \left(D'V^T\right)^{-1} (M_1 + M_j) \right\|_2 \right\} \\
&\leq 2d \min \left\{ \left|U(p)^T (M_1 + M_j)\right|, \left\| \left(D'V^T\right)^{-1} (M_1 + M_j) \right\|_2 \right\},
\end{aligned} \tag{54}$$

where the equality holds by definition of $M$, the first inequality holds by Lemma 18, the second inequality holds by Lemma 32 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

$$s(p) \leq 2d^{1+\left|\frac{1}{2}-\frac{1}{z}\right|} w(p) \sum_{j=1}^{d} \min \left\{ \left| U(p)^T (M_{*1} + M_{*j}) \right|, \left\| \left(D'V^T\right)^{-1} (M_{*1} + M_{*j}) \right\|_z \right\}, \quad (55)$$

where the inequality follows from invoking Lemma 18.

Summing (55) over every $p \in P$, we obtain that

$$\sum_{p \in P} s(p) \leq 2\gamma d^{1+\left|\frac{1}{2}-\frac{1}{z}\right|} \sum_{j=1}^{d} \left\| M_1 + M_j \right\|_2 \leq 4\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|},$$

where the first inequality is by (52) and the second inequality holds since $\|x+y\|_2 \leq 2$ for any pair of unit vectors $x, y \in \mathbb{R}^d$. $\square$

**Corollary 10** (Outlier resistant functions)**.** *Let $P \subseteq \mathbb{R}^d$ be a set of $n$ points, and let $f_{\mathrm{RES}\ell_z} : P \times \mathbb{R}^d \to [0, \infty)$ be loss function such that for every $x \in \mathbb{R}^d$, and $p \in P$, $f_{\mathrm{RES}\ell_z}(p, x) = \min \left\{ \left| p^T x \right|, \|x\|_z \right\}$. Then, there exists an algorithm that gets the set $P$ as an in 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 $f_{\mathrm{RES}\ell_z}$, and (ii) the size of the coreset is $O\left( \frac{\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|}}{\varepsilon^2} \left( d \log \left( \gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|} \right) + \log\left(\frac{1}{\delta}\right) \right) \right)$, where $\gamma$ is defined in the proof.*

*Proof.* First, observe that by Lemma 34 the total sensitivity of the query space $(P, w, \mathbb{R}^d, f_{\mathrm{RES}\ell_z})$ is bounded by $O\left(\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|}\right)$. 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 $O\left( \frac{\gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|}}{\varepsilon^2} \left( d \log \left( \gamma d^{2+\left|\frac{1}{2}-\frac{1}{z}\right|} \right) + \log \frac{1}{\delta} \right) \right)$, where each point is sampled with probability $\frac{s(p)}{t}$, and let $v(P) = \frac{w(p)t}{s(p)|S|}$. Hence by Theorem 3, we get that with probability at least $1-\delta$, $(S, v)$ is an $\varepsilon$-coreset for the query space $(P, w, \mathbb{R}^d, f_{\mathrm{RES}\ell_z})$.

$\square$

# G "Easy" examples covered by our framework

## G.1 $\ell_z$-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 fo every $x \in \mathbb{R}^d$ and $p \in P$ the loss function $f_{\ell_z} : P \times \mathbb{R}^d \to [0, \infty)$ is defined to be $f_{\ell_z}(p, x) = \left| p^T x \right|^z$. 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*

$$s(p) \leq \begin{cases} w(p) \left\| U(p) \right\|_z^z & z \in [1, 2] \\ \sqrt{d^z} w(p) \left\| U(p) \right\|_z^z & \textit{otherwise} \end{cases},$$

*(ii) and the total sensitivity is bounded by*

$$\sum_{p \in P} s(p) \leq \begin{cases} d^{\frac{z}{2}+1} & z \in [1, 2) \\ d & z = 2 \\ d^{z+1} & \textit{otherwise} \end{cases}.$$

*Proof.* Note the following:

(a) For every $q \in P$, $x \in \mathbb{R}^d$ and $b \geq 0$ we have $\left|q^T bx\right| = b\left|q^T x\right|$.

(b) Since $\left|q^T x\right|$ is convex function, it also holds that $\sum\limits_{q \in P} w(q)\left|q^T x\right|^z$ is convex due to the fact that sum of convex functions is also convex,

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

(d) For any unit vector $x \in \mathbb{R}^d$ and $q \in P$,

$$\left|U(q)^T x\right|^z \leq \|U(q)\|_2^z \|x\|_2 = \|U(q)\|_2^z \leq \begin{cases} \|U(q)\|_z^z & z \in [1,2] \\ d^{\frac{1}{2}-\frac{1}{z}} \|U(q)\|_z^z & z > 2 \end{cases},$$

where the first inequality holds by Cauchy Schwartz's inequality, the equality is by the assumption that $x$ is a unit vector, and the last inequality holds by plugging $a := 2$ and $b := z$ for $z > 2$ and $a := z$ and $b := 2$ for $z \in [1,2]$ into Lemma 18.

Hence, plugging

- $f(p,x) := f_{\ell_z}(p,x)$, $g(p,x) := f_{\ell_z}(p,x)$, $h(p,x) := 0$ for every $p \in P$, and $x \in \mathbb{R}^d$,
- $c_i := 1$ for every $i \in [5]$,
- $z := z$,
- $\alpha := \sqrt{d}$ for $z \neq 2$ and $\alpha := 1$ for $z = 2$,
- $v_i := e_i$ where $e_i$ denotes a vector which at its $i$th entry there is 1, and 0's elsewhere,
- and $c := \begin{cases} 1 & z \in [1,2] \\ d^{\frac{1}{2}-\frac{1}{z}} & z > 2 \end{cases}$,

into Lemma 5, yields that

$$s(p) = w(p) \sum_{i \in [d]} \left|U(p)^T D V^T e_i\right|^z \cdot \begin{cases} 1 & z \in [1,2] \\ d^{\frac{1}{2}-\frac{1}{z}} & z > 2 \end{cases}$$

This satisfies (i) as

$$\sum_{i \in [d]} \left|U(q)^T e_i\right|^z = \|U(q)\|_z^z,$$

holds for every $q \in P$ by definition of norms.

As for the sum of sensitivities, Claim (ii) follows from Lemma 5. $\square$

**Corollary 36.** *Let $(P, w, \mathbb{R}^d, f_{\ell_z})$ be a query space, such that for every $x \in \mathbb{R}^d$, and $p \in P$, the loss function $f_{\text{NC}\ell_z}$ is defined to be*

$$f_{\ell_z}(p,x) = \left|p^T x\right|^z.$$

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

$$|S| \in \begin{cases} O\left(\frac{d^{\frac{z}{2}+1}}{\varepsilon^2}\left(d\log\left(d^{\frac{z}{2}+1}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) & z \in [1,2) \\ O\left(\frac{d}{\varepsilon^2}\left(d\log(d) + \log\left(\frac{1}{\delta}\right)\right)\right) & z = 2 \\ O\left(\frac{d^{z+1}}{\varepsilon^2}\left(d\log\left(d^{z+1}\right) + \log\left(\frac{1}{\delta}\right)\right)\right) & z \in (2,\infty) \end{cases}.$$

*Proof.* First, observe that by Lemma 35, the total sensitivity is bounded by $t := \begin{cases} d^{\frac{z}{2}+1} & z \in [1,2) \\ d & z = 2 \\ d^{z+1} & \text{otherwise} \end{cases}$.

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 $O\left(\frac{t}{\varepsilon^2}\left(d \log(t) + \log\left(\frac{1}{\delta}\right)\right)\right)$. $\square$

### G.2 Least squared errors

**Lemma 37.** *Let $(P, w, \mathbb{R}^d, f_{\text{LSE}})$ be a query space, such that for every $x \in \mathbb{R}^d$ and $p \in P$, the loss function $f_{\text{LSE}}$ is defined to be $f_{\text{LSE}}(p, x) = \|p - x\|_2^2$. Let $P' = \left\{ p' = \begin{bmatrix} \|p\|_2^2 \\ 2p \\ 1 \end{bmatrix} \middle| p \in P \right\}$ and let $g_{\text{LSE}} : P' \times \mathbb{R}^{d+2} \to [0, \infty)$ such that for every $y \in \mathbb{R}^{d+2}$ and $p \in P'$, $g_{\text{LSE}}(p, y) = \left|p^T y\right|$. Let $(U, D, V)$ be the $f$-SVD of $P'$ with respect to $g_{\text{LSE}}$. Then, claims (i) – (ii) hold as follows:*

*(i) for every $p \in P$, the sensitivity of $p$ with respect to the query space $(P, w, \mathbb{R}^d, f_{\text{LSE}})$ is bounded by $s(p) \leq w(p)\left\|U(p')\right\|_1$,*

*(ii) and the total sensitivity is bounded by $\sum\limits_{p \in P} s(p) \in O\left(d^{1.5}\right)$.*

*Proof.* Put $p \in P$, and observe that for every $y \in \mathbb{R}^d$, $\|p - y\|_2^2 = \|p\|_2^2 - 2p^T y + \|y\|_2$, which enables us to rewrite the problem by reformulating the query space and the input space ($\mathbb{R}^d$ and $P$ respectively). Let $X' = \left\{ \begin{bmatrix} 1 \\ -x \\ \|x\|_2^2 \end{bmatrix} \middle| x \in \mathbb{R}^d \right\}$. Then, we obtain that for every $x \in X'$

$$\frac{w(p) f_{\text{LSE}}(p, x)}{\sum\limits_{q \in P} w(q) f_{\text{LSE}}(q, x)} = \frac{w(p)\left|{p'}^T x\right|}{\sum\limits_{q \in P} w(q)\left|{q'}^T y\right|} \leq \sup_{y \in \mathbb{R}^{d+2}} \frac{w(p)\left|{p'}^T y\right|}{\sum\limits_{q \in P} w(q)\left|{q'}^T y\right|},$$

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. $\square$

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

$$f_{\text{LSE}}(p, x) = \|p - x\|_2^2.$$

*Let $P' = \left\{ \begin{bmatrix} p' = \|p\|_2^2 \\ 2p \\ 1 \end{bmatrix} \middle| p \in P \right\}$ and let $g_{\text{LSE}} : P' \times \mathbb{R}^{d+2} \to [0, \infty)$ such that for every $x \in \mathbb{R}^{d+2}$ and $p \in P'$, $g_{\text{LSE}}(p, x) = \left|p^T x\right|$. For every $p \in P$ and $p' = (\|p\|_2^2 \mid 2p \mid 1)$ we define $w'(p') = w(p)$. Let $\varepsilon, \delta \in (0, 1)$, and let $(S', v')$ be a coreset for the query space $\left(P', w', \mathbb{R}^{d+2}, g_{\text{LSE}}\right)$ by Corollary 36. Let $S = \left\{ p \mid (\|p\|_2^2 \mid 2p \mid 1) \in S' \right\}$, and for every $p \in S$, and $p' = (\|p\|_2^2 \mid 2p \mid 1) \in S'$ let $v(p) = v'(p')$. Then, with probability at least $1 - \delta$, $(S, v)$ is an $\varepsilon$-coreset for the query space $(P, w, \mathbb{R}^d, f_{\text{LSE}})$, and the size of the coreset is $|S| \in O\left(\frac{(d+2)^{2.5}}{\varepsilon^2}\left(d \log\left((d+2)^{2.5}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$.*

*Proof.* First, observe that by Lemma 37, the total sensitivity is bounded by $t := (d+2)^{1.5}$ of $\left(P', w', \mathbb{R}^{d+2}, g_{\text{LSE}}\right)$. Plugging $P := P'$, $t := t$, $\varepsilon := \varepsilon$ and $\delta := \delta$ into Corollary 36, yields that $S', v'$ is an $\varepsilon$-coreset of size $O\left(\frac{(d+2)^{2.5}}{\varepsilon^2}\left(d \log\left((d+2)^{2.5}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$ for the query space $\left(P', w', \mathbb{R}^{d+2}, g_{\text{LSE}}\right)$.

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

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

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 $\left(P, w, \mathbb{R}^d, f_{\text{LSE}}\right)$. $\square$

# 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 $O\left(n^2 d\right)$, 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 $O\left(nd + poly(d)\right)$ 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.