No Dimensional Sampling Coresets for Classification

Meysam Alishahi [1] **Jeff M. Phillips** [1 2]

Abstract

We refine and generalize what is known about coresets for classification problems via the sensitivity sampling framework. Such coresets seek the smallest possible subsets of input data, so one can optimize a loss function on the coreset and ensure approximation guarantees with respect to the original data. Our analysis provides the first no dimensional coresets, so the size does not depend on the dimension. Moreover, our results are general, apply for distributional input and can use iid samples, so provide sample complexity bounds, and work for a variety of loss functions. A key tool we develop is a Radamacher complexity version of the main sensitivity sampling approach, which can be of independent interest.

1. Introduction

In machine learning, coresets (Phillips, 2016) are small subsets of input data that act as proxy for the full set while *guaranteeing* not to deviate much in accuracy. By reducing data size, they improve scalability and efficiency. A common approach for constructing coresets involves sampling data points with probabilities proportional to their so-called *sensitivity score*, a bound on the worst-case impact that a point can have on the property of interest. In classification problems, the property optimized is a loss function ℓ which is typically a smooth and convex approximation of the misclassification rate. However, the ultimately goal is not just to optimize the loss function over the observed data, but to understand how well the classifier will perform using new data drawn iid from the same (unknown) distribution P.

To formalize this, consider a probability distribution P over a set $\mathcal{X}$. Let $\mathcal{W}$ be the parameter space of potential models, and then $\ell : \mathcal{X} \times \mathcal{W} \longrightarrow [0, \infty)$ is a loss (or cost) function defined on a single $x \in \mathcal{X}$. Let $\int_{x\in\mathcal{X}} \ell(x, w)\mathrm{d}P(x)$ be the full loss function evaluated at $w \in \mathcal{W}$, which we will seek to minimize. Now a (finite) set $Y \subseteq \mathcal{X}$ accompanied by a measure (or weight function) ν, is termed an ε-*coreset* for $(\mathcal{X}, P, \mathcal{W}, \ell)$ for $\varepsilon \in (0, 1)$ when it approximates P in the following way for each $w \in \mathcal{W}$:

$$\Big| \int_{\mathcal{X}} \ell(x, w)\mathrm{d}P(x) - \sum_{y\in Y} \nu(y)\ell(y, w) \Big| \leq \varepsilon \int_{\mathcal{X}} \ell(x, w)\mathrm{d}P(x)$$

Algorithms for coresets are often described with respect to a finite sample $X \subset \mathcal{X}$ where we then assume P has support limited to and uniform over X. We will return later to the integral of P definition to state more general bounds. Then a very common approach to create a coreset is via sensitivity sampling (Feldman & Langberg, 2011). This samples points from X proportional to their sensitivity score $s(x)$. Via an importance sampling argument, one can show the size of the sample Y needed to induce an ε-coreset grows proportionally to the total sensitivity $S = \sum_{x\in X} s(x)$.

For classification, consider a non-increasing function $\phi : \mathbb{R} \longrightarrow (0, \infty)$ and a set $\mathcal{W} \subseteq \mathbb{R}^d$. Define $\ell_\phi(x, w) = \phi(\langle x, w\rangle)$; this aligns with scenarios like Logistic regression when $\phi(t) = \log(1 + e^{-t})$. However, in (Munteanu et al., 2018), it was proven that certain types of non-increasing functions ϕ, specifically those where $\frac{\phi(x)}{\phi(-x)}$ approaches zero as x goes to infinity, can lead to ε-coresets that encompass all points in X (see Theorem 3.1 in (Tolochinksy et al., 2022) which refines (Munteanu et al., 2018)). To construct a smaller coreset, some additional problem structure needs to be introduced. We mostly focus on the natural way of regularizing the problem (Shalev-Shwartz & Ben-David, 2014), initiated in the context of coresets by Tolochinksy et al. (2022). For a positive integer k, define $\ell_{k,\phi}(x, w) = \phi(\langle x, w\rangle) + \frac{1}{k}\|w\|_2^2$. An ε-coreset for $(X, p, \mathcal{W}, \ell_{k,\phi})$ was termed a *coreset for a monotonic function* ϕ (Tolochinksy et al., 2022; Curtin et al., 2020). Others approached this issue in varied ways. For example, Munteanu et al. (2018) introduced a complexity measure $\mu(X)$ that quantifies the hardness of compressing a dataset for logistic regression, and showed sublinear-sized coresets can be founding depending on this parameter. Their results were polynomially improved by Mai et al. (2021) in terms

*Equal contribution [1]Kahlert School of Computing, University of Utah, Salt Lake City, Utah, USA [2]visiting ScaDS.AI, University of Leipzig and MPI for Math in the Sciences, Leipzig, Germany. Correspondence to: Meysam Alishahi <meysam.alishahi@utah.edu>, Jeff M. Phillips <jeffp@cs.utah.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

of $\mu(X)$ and d. Other work (Mirzasoleiman et al., 2020) considers greedily selected coresets for learning classifiers under incremental gradient descent where regularization is again needed, but now to ensure strong convexity.

The most studied examples of ϕ are:

- *sigmoid function*: $\sigma(t) = \frac{1}{1+e^t}$,
- *logistic function*: $\text{logistic}(t) = \log(1 + e^{-t})$,
- *svm loss function*: $\text{svm}(t) = \max(0, 1 - t)$.
- *ReLU function*: $\text{ReLU}(t) = \max(0, t)$

Main Results. Before we describe our results in more technical detail, we first summarize our main contributions.

- We provide the first *no dimensional* coreset for monotone functions, of size $O(k^3/\varepsilon^2)$; it has no dependence on dimension d. A special case of our results are in Table 1 with comparisons to prior work; this includes a $(\sqrt{\log k})$ factor improvement for logistic loss.
- These bounds use a uniform sample from X, *and* apply to iid samples from unknown distribution P; so this provides a true sample complexity bound for regularized classification.
- We provide a new general sensitivity sampling bound based on Radamacher complexity (Theorem 1.1), which leads to these no dimensional bounds, and we believe will be of general interest.
- We provide a, in our opinion, simpler proof of the main sensitivity sampling claim using VC dimension (Feldman et al., 2020; Braverman et al., 2021), which can handle issues of weighted loss functions needed for these results.

1.1. Preliminaries

Throughout the paper, we maintain the assumption that P is a probability measure over $\mathcal{X}$. If $f : \mathcal{X} \longrightarrow [0, \infty)$ is P-integrable, then the *expectation of f with respect to* P is defined as $\mathbb{E}_{x\sim P}(f(x)) = \int_{\mathcal{X}} f(x)\mathrm{d}P$. While this framework is quite broad, the most natural instances stem from the following two examples.

Discrete: Let p be a discrete distribution defined over a countable set $\mathcal{X} = \{x_1, x_2, \ldots\} \subseteq \mathbb{R}^d$. For any subset $A \subseteq \mathcal{X}$, the probability measure $P(A)$ is computed as the sum of probabilities of elements in A, given by $P(A) = \sum_{a\in A} p(a)$. In this scenario, given a function $f : \mathcal{X} \longrightarrow [0, \infty)$, the expectation $\mathbb{E}_{x\sim P}(f(x))$ is expressed as $\int_{\mathcal{X}} f(x)\mathrm{d}P = \sum_{i=1}^{\infty} p(x_i)f(x_i)$. The case that $\mathcal{X}$ is finite falls under this framework when the distribution p possesses finite support.

Continuous: Let p be a continuous probability density distribution defined over $\mathbb{R}^d$, and $\boldsymbol{m}$ represents the Lebesgue measure (or Borel measure) over $\mathbb{R}^d$. Then the probability measure $P(A)$ for a measurable set A is computed as $P(A) = \int_A p(x)\mathrm{d}\boldsymbol{m}$.

When $p(x)$ represents a probability density function, the notation $x \sim P$ denotes sampling based on p.

If $\mathcal{F}$ is a set of P-integrable functions from $\mathcal{X}$ to $[0, \infty)$ such that $\int f(x)\mathrm{d}P(x) > 0$ for all $f \in \mathcal{F}$, then the tuple $(\mathcal{X}, P, \mathcal{F})$ is termed as *positive definite*. A P-integrable function $s\colon \mathcal{X} \longrightarrow (0, \infty)$ is called an *upper sensitivity function* for a positive definite tuple $(\mathcal{X}, P, \mathcal{F})$ if

$$\sup_{f\in\mathcal{F}} \frac{f(x)}{\int f(z)\mathrm{d}P(z)} \leq s(x) \qquad \text{for almost all } x \in \mathcal{X}. \quad (1)$$

The value $S = \int s(x)\mathrm{d}P(x)$ will be referred to as the *total sensitivity* for $(\mathcal{X}, P, \mathcal{F})$ with respect to the function $s(\cdot)$. The *sensitivity normalized probability measure* for $(\mathcal{X}, P, \mathcal{F})$ is defined as $\mathrm{d}Q = \frac{s(x)}{S}\mathrm{d}P$ or equivalently $Q(A) = \int_A \frac{s(x)}{S}\mathrm{d}P(x)$ for each P-measurable A. Given that $S = \int s(x)\mathrm{d}P(x)$, it follows that Q represents a probability measure over $\mathcal{X}$ since $\int_{\mathcal{X}} \mathrm{d}Q(x) = \int_{\mathcal{X}} \frac{s(x)}{S}\mathrm{d}P(x) = 1$. An *$s$-sensitivity sample* from $\mathcal{X}$ with a size of m is a set of m iid draws $x_1 \ldots, x_m$ from $\mathcal{X}$ based on Q which will be expressed as as $x_{1:m} \sim Q$. To simplify, one can imagine iid sampling from $\mathcal{X}$ using the probability distribution $q(x) = \frac{s(x)}{S}p(x)$. Define the *$s$-augmented family of* $(\mathcal{X}, P, \mathcal{F})$ as

$$\mathcal{T}_{\mathcal{F}} = \left\{ T_f(\cdot) = \frac{Sf(\cdot)}{s(\cdot)\int_{\mathcal{X}} f(x)\mathrm{d}P(x)} : f \in \mathcal{F} \right\}. \quad (2)$$

For each $f \in \mathcal{F}$, $\sup_{x\in\mathcal{X}} T_f(x) \leq S$ and

$$\mathop{\mathbb{E}}_{x\sim Q}(T_f(x)) = \int_{\mathcal{X}} \frac{Sf(x)}{s(x)\int_{\mathcal{X}} f(z)\mathrm{d}P(z)} \frac{s(x)}{S}\mathrm{d}P(x) = 1.$$

Thus, for any $f \in \mathcal{F}$,

$$\begin{aligned} \Big| \int_{\mathcal{X}} f(x)\mathrm{d}P(x) - \sum_{i=1}^{m} \frac{Sf(x_i)}{ms(x_i)} \Big| \leq \varepsilon \int_{\mathcal{X}} f(x)\mathrm{d}P(x) \\ \iff \quad \left| 1 - \frac{1}{m}\sum_{i=1}^{m} T_f(x_i) \right| \leq \varepsilon. \end{aligned} \quad (3)$$

The *m-th Rademacher Complexity* of a family $\mathcal{F}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ with respect to a probability measure P over $\mathcal{X}$ is given by

$$\mathcal{R}_m^P(\mathcal{F}) = \mathop{\mathbb{E}}_{x_{1:m}\sim P} \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}} \sup_{f\in\mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^{m} \sigma_i f(x_i) \right],$$

were $\sigma_{1:m} \sim \{-1, 1\}$ denotes m iid samples $\sigma_1, \ldots, \sigma_m$ uniformly drawn from $\{-1, 1\}$. Rademacher Complexity somehow serves as a more detailed dimension than VC-dimension.

Table 1. Comparison with prior state-of-the-art results. The notation $O^*(\cdot)$ hides some logarithmic factor in terms of given parameters.

Function ϕ	Assumption	Sampling	Coreset size	Reference
logistic	bounded $\mu(X)$	sqrt lev. scores	$O^*\left(\frac{d^3\mu^3(X)}{\varepsilon^4}\right)$ $O^*\left(\frac{\sqrt{n}d^{\frac{3}{2}}\mu(X)}{\varepsilon^2}\right)$	(Munteanu et al., 2018)
logistic, svm, ReLU	bounded $\mu(X)$	ℓ_l Lewis	$O\left(\frac{\mu^2(X)d\log(d/\varepsilon)}{\varepsilon^2}\right)$	(Mai et al., 2021)
logistic, svm	$\|x_i\|_2 \le 1\ \forall i$ reg. term: $\frac{1}{k}\|w\|_2^2$ $\|w\|_2 \le 1$	sensitivity	$O^*\left(\frac{kd^2\log n\log k}{\varepsilon^2}\right)$	(Tolochinksy et al., 2022)
σ	reg. term: $\frac{1}{k}\|w\|_2^2$	sensitivity	$O^*\left(\frac{kd^2\log n\log k}{\varepsilon^2}\right)$	(Tolochinksy et al., 2022)
logistic, svm	$\|x_i\|_2 \le 1\ \forall i$ reg. term: $\frac{1}{k}\|w\|_2^2$ $\frac{1}{k}\|w\|_2$, or $\frac{1}{k}\|w\|_1$	uniform	$O\left(\frac{dk\log k}{\varepsilon^2}\right)$	(Curtin et al., 2020)
σ, svm, ReLU	$\|x_i\|_2 \le 1\ \forall i$ reg. term: $\frac{1}{k}\|w\|_2^2$	uniform	$O\left(\frac{dk\log k}{\varepsilon^2}\right)$ $O\left(\frac{k^3}{\varepsilon^2}\right)$	Theorems D.6,D.7, D.18, D.20, and Section D.4
logistic	$\|x_i\|_2 \le 1\ \forall i$ reg. term: $\frac{1}{k}\|w\|_2^2$	uniform	$O\left(\frac{dk\sqrt{\log k}}{\varepsilon^2}\right)$ $O\left(\frac{k^3}{\varepsilon^2\sqrt{\log k}}\right)$	Theorems 2.7, 2.8

1.2. Our Contributions

Our first main result is the following theorem.

Theorem 1.1. *Let $(\mathcal{X}, P, \mathcal{F})$ be a positive definite tuple and $s(\cdot)$ be an upper sensitivity function with the total sensitivity S. For any $t > 0$, an s-sensitivity sample from $\mathcal{X}$ of size m, with probability at least $1 - 2\exp\left(-\frac{2mt^2}{S}\right)$, satisfies*

$$\left|\int f(x)\mathrm{d}P(x) - \sum_{i=1}^{m}\frac{Sf(x_i)}{ms(x_i)}\right| \le (2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}) + t)\int f(x)\mathrm{d}P(x) \quad \forall f \in \mathcal{F}.$$

This theorem, which is pivotal in our new results, is an adaptation of the central result from (Feldman & Langberg, 2011; Braverman et al., 2021; Feldman et al., 2020), extensively employed in coreset construction results, but now centered on Rademacher Complexity instead of VC-dimension.

Given a function $f : \mathcal{X} \longrightarrow [0, \infty)$, for each $r \ge 0$, the set $\mathbf{range}(f, \succ, r)$ is defined as:

$$\mathbf{range}(f, \succ, r) = \{x \in \mathcal{X} \colon f(x) > r\}.$$

Now, considering a family of non-negative P-measurable functions $\mathcal{F}$, the *$\mathcal{F}$-linked range space* (see Joshi et al. 2011) is defined as $(\mathcal{X}, \mathbf{Ranges}(\mathcal{F}, \succ))$ where

$$\mathbf{Ranges}(\mathcal{F}, \succ) = \{\mathbf{range}(f, \succ, r) \colon f \in \mathcal{F}, r \ge 0\}.$$

Consider a range space $(\mathcal{X}, \mathcal{R})$, where $\mathcal{X}$ is the ground set and $\mathcal{R}$ is a collection of subsets of $\mathcal{X}$. A subset $Y \subseteq \mathcal{X}$ is said to be shattered by $\mathcal{R}$ if every subset $Z \subseteq Y$ can be expressed as $Z = Y \cap R$ for some $R \in \mathcal{R}$. The *VC-dimension of* $(\mathcal{X}, \mathcal{R})$ is defined as the cardinality of the largest subset $Y \subseteq \mathcal{X}$ that can be shattered by R.

Theorem 1.2. *Let $(\mathcal{X}, P, \mathcal{F})$ be a positive definite tuple, and $s \colon \mathcal{X} \longrightarrow (0, \infty)$ be an upper sensitivity function for it with the total sensitivity $S = \int s(x)\mathrm{d}P(x)$. There is a universal constant C such that, for any $0 < \varepsilon, \delta < 1$, if $m \ge \frac{CS}{\varepsilon^2}\left(\mathsf{VC}\log S + \log\frac{1}{\delta}\right)$, where $\mathsf{VC} = \mathrm{VCdim}(\mathbf{Ranges}(\mathcal{T}_\mathcal{F}, \succ))$, then, with probability at least $1 - \delta$, any s-sensitivity sample $x_1 \ldots, x_m$ from $\mathcal{X}$ satisfies*

$$\left|\int f(x)\mathrm{d}P(x) - \frac{1}{m}\sum_{i=1}^{m}\frac{Sf(x_i)}{s(x_i)}\right| \le \varepsilon\int f(x)\mathrm{d}P(x) \qquad \forall f \in \mathcal{F}.$$

This theorem, with variants previously shown in (Braverman et al., 2021; Feldman et al., 2020), are usually constrained by the assumption that $\mathcal{X}$ is finite, which has been extensively used to construct coresets. However, in this work we remove this restriction. This general case was also handled in (Langberg & Schulman, 2010). Section 2.2 showcases a novel and simple proof for this theorem, leveraging Fubini's Theorem (Fubini, 1907)[1]. Moreover, in instances

[1]Fubini's Theorem is applicable in our context, given we are

where Rademacher complexity refines VC-dimension, our new result Theorem 1.1 provides improved bounds. Next we describe several applications stemming from these two theorems.

Well-behaved distributions. We next describe essential problem parameters, generalizing ones used elsewhere.

Definition 1.3. Let P be a probability measure over $\mathbb{R}^d$, $\phi\colon \mathbb{R} \longrightarrow [0,\infty)$ a non-increasing Lipschitz function. We say P is *well-behaved* (w.r.t. ϕ) if

$$\mathop{\mathbb{E}}_{x\sim P} \|x\|_2^2 \le E_1^2 \quad \text{and} \quad \mathop{\mathbb{E}}_{x\sim P} \phi\left(\frac{\|x\|_2}{2E_1}\right) \ge E_2, \tag{4}$$

for positive constants E_1, E_2. For constant $k > 0$ define

$$\mathcal{L}_{\phi,k} = \left\{ \ell_{\phi,k}(w,\cdot) = \phi(\langle w,\cdot\rangle) + \frac{\|w\|_2^2}{k} \colon w \in \mathcal{X} \right\}. \tag{5}$$

Given $\mathbb{E}(t^2) \ge \mathbb{E}^2(t)$, we can conclude that $\mathbb{E}_{x\sim P}(\|x\|_2) \le E_1$. The assumption on the expectation of $\|x\|_2^2$ is fairly standard, such as in bounded or normalized data where it is an absolute constant. On the other hand, when p is the Gaussian distribution $\mathcal{N}_d(\mathbf{0}, \boldsymbol{I})$, then $E_1 = d$. The lower bound for $\mathop{\mathbb{E}}_{x\sim P} \phi\left(\frac{\|x\|_2}{2E_1}\right)$ is a constant bounded by $\phi(1/2)$ when ϕ is convex; then we can apply Jensen's inequality so

$$\mathop{\mathbb{E}}_{x\sim P} \phi\left(\frac{\|x\|_2}{2E_1}\right) \ge \phi\left(\mathop{\mathbb{E}}_{x\sim P} \frac{\|x\|_2}{2E_1}\right) \ge \phi\left(\frac{1}{2}\right) = E_2. \tag{6}$$

Connecting to coresets. Assume that $s(\cdot) : \mathcal{X} \longrightarrow (0,\infty)$ is an upper sensitivity function for a family of functions $\mathcal{F}$ with total value $S = \int_{\mathcal{X}} s(x)\mathrm{d}P$. Since $\frac{f(x)}{\int f(z)\mathrm{d}P(z)} \le s(x)$ for each $f \in \mathcal{F}$, we have $1 = \int \frac{f(x)}{\int f(z)\mathrm{d}P(z)}\mathrm{d}P(x) \le \int s(x)\mathrm{d}P(x) = S$. If we define $s'(\cdot) = s(\cdot) + 1$, then $s' : \mathcal{X} \longrightarrow (1,\infty)$ is also an upper sensitivity function for that family with total value $S' = 1 + S = O(S)$, since $S \ge 1$.

Theorem 1.4. *Let P be a well-behaved probability measure over $\mathbb{R}^d$ and ϕ be an L-Lipschitz function. If $s(\cdot) : \mathbb{R}^d \longrightarrow (1,\infty)$ is an upper sensitivity function for $\mathcal{L}_{\phi,k}$ with total sensitivity S and $m \ge \frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right)$, then any s-sensitivity sample $x_1,\ldots,x_m$ from $\mathcal{X}$ with weights $u_i = \frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathbb{R}^d,\ P,\ \mathbb{R}^d,\ \ell_{\phi,k})$ with probability at least $1-\delta$, where $C = (2LE_1 + \phi(0))\max\left(2E_1k, \frac{1}{E_2}\right) + 8LkE_1 + 1$.*

When E_1 is independent of dimension, such as bounded or normalized data, this theorem gives an ε-coreset for $(\mathbb{R}^d,\ P,\ \mathbb{R}^d,\ \ell_{\phi,k})$ with size *independent of dimension d and input size n*. In contrast, the coresets for (Tolochinksy et al., 2022), also considering bounded data, used VC-dimension (via Theorem 1.2), so their size depends on d and n.

Bounds for specific ϕ. Before describing our bounds for common choices of ϕ we examine the context of prior work by Tolochinksy et al. (2022). They considered absolutely bounded data with $\|x\|_2^2 \le 1$; so $E_1 = 1$ and $E_2 = \phi(1)$. For simpler notation we define R so that $\mathcal{W}_R = \{x \in \mathbb{R}^d \colon \|w\|_2 \le R\}$. Up to a few small typos, we summarize their results:

Theorem 1.5 (Tolochinksy et al., 2022)**.** *Let $X = \{x_1,\ldots,x_n\}$ be a set of n points in the unit ball of $\mathbb{R}^d$, $\boldsymbol{u}$ be the uniform probability measure over X, and $\varepsilon,\delta \in (0,1)$, $R,k > 0$ where k is a sufficiently large constant. Then $m = O(\frac{t}{\varepsilon^2}(d^2\log t + \log\frac{1}{\delta})$ sensitivity samples is, with probability at least $1-\delta$, an ε-coreset for*

- $(X, \boldsymbol{u}, \mathbb{R}^d, \ell_{k,\sigma})$ *with* $t = (1+k)\log n$.
- $(X, \boldsymbol{u}, \mathcal{W}_R, \ell_{k,\mathrm{logistic}})$ *with* $t = R^2(1+Rk)\log n$.
- $(X, \boldsymbol{u}, \mathcal{W}_R, \ell_{k,\mathrm{svm}})$ *with* $t = R(1+Rk)\log n$.

It should be emphasized that the last two results in Theorem 1.5 was successfully improved to $O(\frac{k}{\varepsilon^2}(d\log k + \log\frac{1}{\delta})$ for $\mathcal{W} = \mathbb{R}^d$ and $k = \Theta(n^{1-\kappa})$ for some $\kappa \in (0,1)$ by Curtin et al. (2020). The strategy utilized by Tolochinksy et al. (2022) and Curtin et al. (2020) to demonstrate this theorem can be encapsulated as: (1) find an upper bound for VC-dimension, (2) approximate a sensitivity upper bound and thus the total sensitivity, (3) sample m points form X according to an appropriate distribution, and (4) use Theorem 1.2 to complete the argument. The argument by Tolochinksy et al. (2022), however, encounters a critical issue. To apply Theorem 1.2, we must replace d with the VC-dimension of the linked-range space of the s-augmented family of $(\mathcal{X}, P, \mathcal{L}_{\phi,k})$, i.e. $\mathrm{VCdim}(\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\phi,k}}, \succ))$, while they directly used the VC-dimension of linked-range space of $\mathcal{F}$ (instead of $\mathcal{T}_{\mathcal{L}_{\phi,k}}$). Our work addresses this issue in that it leads to sensitivity functions for which we can compute the VC-dimension of the augmented function space. Furthermore, by instead bounding Radamacher complexity in step (1), we can then use our new Theorem 1.1 in step (4) to obtain new bounds. These are summarized in Table 2 and the next theorem; details are in appendix.

Theorem 1.6. *Let P be a probability measure over $\mathbb{R}^d$. Given the conditions outlined in the initial 5 columns of Table 2, any s-sensitivity sample size m (sampling according to the distribution provided in Column 6) guarantees an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \mathcal{L}_{\phi,k})$ with a probability of at least $1-\delta$.*

dealing with σ-finite measures.

Table 2. Coreset size for $(\mathbb{R}^d, P, \mathbb{R}^d, \mathcal{L}_{\phi,k})$; it is common for $E_1, A = 1$.

function ϕ	data assumption	total sens. S	constant C	sampling	Coreset size	reference
σ	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1^2$	$O(E_1^2 k)$	$O(E_1^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.6
σ	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1^2$	$O(E_1^2 k)$	$O(1)$	p	$\frac{CS}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)$	Thm D.7
logistic	$P(\|x\|_2 \ge A) = 0$	$O(\frac{A^2 k}{\sqrt{\log(A^2 k)}})$	$O(A^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.10
logistic	$P(\|x\|_2 \ge A) = 0$	$O(\frac{A^2 k}{\sqrt{\log(A^2 k)}})$	$O(1)$	p	$\frac{CS}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)$	Thm 2.8
logistic	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O(\frac{E_1^2 k}{\sqrt{\log(E_1^2 k)}})$	$O(E_1^2 k)$	$\frac{s}{S}p$	$\frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right))$	Thm D.13
logistic	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O(\frac{E_1^2 k}{\sqrt{\log(E_1^2 k)}})$	$O(1)$	$\frac{s}{S}p$	$\frac{CS}{\varepsilon^2}\left(d^2\log S + \log\frac{1}{\delta}\right)$	Thm D.15
svm	$P(\|x\|_2 \ge A) = 0$	$O(A^2 k)$	$O(A^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.18
svm	$P(\|x\|_2 \ge A) = 0$	$O(A^2 k)$	$O(1)$	p	$\frac{CS}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)$	Thm D.20
svm	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O(E_1^2 k)$	$O(E_1^2 k)$	$\frac{s}{S}p$	$\frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right))$	Thm D.22
svm	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O(E_1^2 k)$	$O(1)$	$\frac{s}{S}p$	$\frac{CS}{\varepsilon^2}\left(d^2\log S + \log\frac{1}{\delta}\right)$	Thm D.23
ReLU	$P(\|x\|_2 \ge A) = 0$	$O((1+A)^2 k)$	$O((1+A)^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Sec. D.4
ReLU	$P(\|x\|_2 \ge A) = 0$	$O((1+A)^2 k)$	$O(1)$	p	$\frac{CS}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)$	Sec. D.4
ReLU	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O((1+E_1)^2 k)$	$O((1+E_1)^2 k)$	$\frac{s}{S}p$	$\frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right))$	Sec. D.4
ReLU	$\mathbb{E}_{x\sim p}\|x\|_2^2 \le E_1$	$O((1+E_1)^2 k)$	$O(1)$	$\frac{s}{S}p$	$\frac{CS}{\varepsilon^2}\left(d^2\log S + \log\frac{1}{\delta}\right)$	Sec. D.4

Sample complexity. Note that if the sampling column of Table 2 uses p, that indicates the total sensitivity has a constant upper bound, and we can use iid samples from the unknown distribution P. Not only can we avoid the computational issue of estimating s, but these provide sample complexity bounds for the full (continuous) setting.

2. Proofs of Main Results

2.1. Proof of Theorem 1.1

The proof follows by McDiarmid's inequality (see lemma C.1), which is a strong concentration inequality that bounds the difference between the sampled mean and the true mean of a function satisfying the bounded differences property: the effect of changing a single observation. Here we outline the proof of Theorem 1.1; full proof in Appendix C.1.

Sketch of proof of Theorem 1.1. In view of Equation (3), we need to show that, for any iid sample $x_1, \dots, x_m \in \mathcal{X}$ according to the sensitivity normalized distribution $q = p(x)\frac{s(x)}{S}$, with probability at least $1 - 2\exp\left(-\frac{2mt^2}{S^2}\right)$,

$$\sup_{f\in\mathcal{F}}\left|1 - \sum_{i=1}^{m}\frac{T_f(x_i)}{m}\right| \le 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}) + t.$$

For every $x \in \mathcal{X}$ and $f \in \mathcal{F}$, we have $\frac{f(x)}{\int_\mathcal{X} f(z)\mathrm{d}P(z)} \le s(x)$ which implies $0 \le T_f(x) = \frac{Sf(x)}{s(x)\int_\mathcal{X} f(z)\mathrm{d}P(z)} \le S$. Therefore, $\sup_{x\in\mathcal{X}} T_f(x) \le S$. Now, define

$$g(x_1, \dots, x_m) = \sup_{f\in\mathcal{F}}\left[1 - \sum_{i=1}^{m}\frac{T_f(x_i)}{m}\right].$$

For each $i \in [m]$ and each $(x_1, \dots, x_m), (x_1', \dots, x_m') \in \mathcal{X}^m$ such that $x_j' = x_j$ for $j \ne i$, observe that

$$|g(x_1, \dots, x_m) - g(x_1', \dots, x_m')| \le \frac{S}{m}.$$

Therefore, using McDiarmid's Inequality, with a probability at least $1 - \exp\left(-\frac{2mt^2}{S^2}\right)$, we have

$$\sup_{f\in\mathcal{F}}\left|1 - \sum_{i=1}^{m}\frac{T_f(x_i)}{m}\right| \le \mathbb{E}_{x_{1:m}\sim q}\sup_{f\in\mathcal{F}}\left|1 - \sum_{i=1}^{m}\frac{T_f(x_i)}{m}\right| + t$$
$$\le 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}) + t. \qquad \square$$

2.2. Proof of Theorem 1.2

This approach to get relative error goes through VC-dimension, which exploits combinatorial properties of range spaces. A *P-range space* is a tuple $(\mathcal{X}, P, \mathcal{R})$ where $(\mathcal{X}, P)$ is a probability measure space and $\mathcal{R}$ is a subset of $2^\mathcal{X}$ whose members are P-measurable. Given a P-range space, the best relative error possible in general is conditioned on a small additive parameter $\eta > 0$ as follows. For an $\varepsilon, \eta \in (0, 1)$, the measure ν on $\mathcal{X}$ is called a *relative* (ε, η)-*approximation for* $(\mathcal{X}, P, \mathcal{R})$ if each $R \in \mathcal{R}$ is ν-measurable

and $|P(R) - \nu(R)| \leq \varepsilon \max(\eta, P(R))$. Using ideas from Li et al. (2001), then Har-Peled & Sharir (2011) showed that a sufficiently large sample according to P provides a relative (ε, η)-approximation for $(\mathcal{X}, P, \mathcal{R})$, if the VC-dimension is bounded as VC. Specifically, there is a universal constant C such that, for any $\eta > 0$, and $\varepsilon, \delta \in (0,1)$, with probability at least $1-\delta$, iid sample $X = \{x_1, \ldots, x_m\}$ according to P with $m \geq \frac{C}{\eta\varepsilon^2}\left(\mathsf{VC}\log\frac{1}{\eta} + \log\frac{1}{\delta}\right)$ satisfies

$$\left|P(R) - \frac{|R \cap X|}{m}\right| \leq \varepsilon \max\left(\eta, P(R)\right) \qquad \forall R \in \mathcal{R}.$$

In order to use this relative error conditioned on η to obtain an unconditioned relative error, the key insight is that using an s-sensitivity sample for an upper sensitivity function s with total sensitivity S, then by setting $\eta = 1/S$, we obtain unconditioned relative error. This is formalized in the following in the next lemma – a discrete version of which is implicit in the proof of Theorem 31 in (Feldman et al., 2020).

Lemma 2.1. *Let $(\mathcal{X}, P, \mathcal{F})$ be a positive definite tuple and $s(\cdot)$ be an upper sensitivity function for it with the total sensitivity S. If the measure ν is a relative (ε, η)-approximation for $(\mathcal{X}, Q, \mathbf{Ranges}(\mathcal{T}_{\mathcal{F}}, \succ))$, where $\mathrm{d}Q(x) = \frac{s(x)}{S}\mathrm{d}P(x)$, then for each $f \in \mathcal{F}$,*

$$\left|\int_{x\in\mathcal{X}} f(x)\mathrm{d}P(x) - \int_{x\in\mathcal{X}} \frac{Sf(x)}{s(x)}\mathrm{d}\nu(x)\right| \leq (\varepsilon + S\eta\varepsilon)\int_{x\in\mathcal{X}} f(x)\mathrm{d}P(x).$$

With this, the proof of Theorem 1.2 is straightforward.

Proof of Theorem 1.2. Since $(\mathcal{X}, Q, \mathbf{Ranges}(\mathcal{T}_{\mathcal{F}}, \succ))$ is a Q-range, for an $\varepsilon \in (0,1)$ and $\eta = \frac{1}{S}$, Har-Peled & Sharir (2011)'s sampling result implies that there is a universal constant C such that, any iid sample $X = \{x_1, \ldots, x_m\}$ according to Q with $m \geq \frac{4CS}{\varepsilon^2}\left(\mathsf{VC}\log S + \log\frac{1}{\delta}\right)$ satisfies

$$\left|Q(R) - \frac{|R\cap X|}{m}\right| \leq \frac{\varepsilon}{2}\max\left(\frac{1}{S}, Q(R)\right) \qquad \forall R \in \mathcal{R},$$

with probability at least $1-\delta$. This implies that the uniform probability measure over X serves as a relative $\left(\frac{\varepsilon}{2}, \frac{1}{S}\right)$-approximation for $(\mathcal{X}, Q, \mathbf{Ranges}(\mathcal{T}_{\mathcal{F}}, \succ))$. The proof concludes by using Lemma 2.1 to derive the error co-efficient on the right hand side as $(\frac{\varepsilon}{2} + S\frac{1}{S}\frac{\varepsilon}{2}) = \varepsilon$. □

We next provide what we believe is a simpler and more direct proof of Lemma 2.1. A key improvement is the use of Fubini's theorem to transform back and forth between integrating over $\mathcal{X}$ to over the range of the function. We first define some useful shorthand notation: $Q_{\succ}(g,t) = Q(\mathbf{range}(g, \succ, t)) = \int_{x\in\mathcal{X}} \mathbb{1}_{g(x)>t}\mathrm{d}Q$.

Lemma 2.2. *For any $g = T_f \in \mathcal{T}_{\mathcal{F}}$ and $r \geq 0$,*

$$\int_{x\in\mathcal{X}} \min\{r, g(x)\}\mathrm{d}Q(x) = \int_0^r Q_{\succ}(g,t))\mathrm{d}t.$$

Proof.

$$\begin{aligned}
&\int_{x\in\mathcal{X}} \min\{r, g(x)\}\mathrm{d}Q(x) \\
&\quad= \int_{x\in\mathcal{X}}\left(\int_{t=0}^{r} \mathbb{1}_{t<g(x)}\,\mathrm{d}t\right)\mathrm{d}Q(x) \\
&\quad= \int_{t=0}^{r}\left(\int_{x\in\mathcal{X}} \mathbb{1}_{t<g(x)}\mathrm{d}Q(x)\right)\mathrm{d}t \qquad (*) \\
&\quad= \int_{t=0}^{r} Q_{\succ}(g,t)\mathrm{d}t,
\end{aligned}$$

where $(*)$ is true due to Fubini's Theorem. □

Proof of Lemma 2.1. Dividing by $P(f) = \int_{\mathcal{X}} f(z)\mathrm{d}P(z)$, we can now restate the goal as proving

$$\frac{1}{P(f)}\left|P(f) - \int_{\mathcal{X}} \frac{Sf(x)}{s(x)}\mathrm{d}\nu(x)\right| \leq \varepsilon(1+S\eta).$$

For an arbitrary $g = T_f \in \mathcal{T}_{\mathcal{F}}$, using that $\int_{\mathcal{X}} g(x)\mathrm{d}\nu = \frac{1}{P(f)}\int_{\mathcal{X}} \frac{Sf(x)}{s(x)}\mathrm{d}\nu(x)$, and then $\int_{\mathcal{X}} g(x)\mathrm{d}Q(x) = 1$ we can transform the left-hand side

$$\begin{aligned}
&\frac{1}{P(f)}\left|P(f) - \int_{\mathcal{X}} \frac{Sf(x)}{s(x)}\mathrm{d}\nu(x)\right| \\
&\quad= \left|1 - \int_{x\in\mathcal{X}} g(x)\mathrm{d}\nu(x)\right| \\
&\quad= \left|\int_{x\in\mathcal{X}} g(x)\mathrm{d}Q(x) - \int_{x\in\mathcal{X}} g(x)\mathrm{d}\nu(x)\right| \\
&\quad= \Big|\int_{x\in\mathcal{X}} \min\{S, g(x)\}\mathrm{d}Q(x) \\
&\qquad\qquad - \int_{x\in\mathcal{X}} \min\{S, g(x)\}\mathrm{d}\nu(x)\Big|, \qquad (7)
\end{aligned}$$

where the last line follows $\max_x g(x) \leq S$. Now applying Lemma 2.2 twice on both Q and ν we have

$$\begin{aligned}
(7) &= \left|\int_{t=0}^{S} Q_{\succ}(g,t) - \int_{t=0}^{S} \nu_{\succ}(g,t)\mathrm{d}t\right| \\
&\leq \int_{t=0}^{S} |Q_{\succ}(g,t) - \nu_{\succ}(g,t)|\,\mathrm{d}t \\
&\leq \varepsilon\int_{t=0}^{S} \max\{\eta, Q_{\succ}(g,t)\}\mathrm{d}t, \qquad (8)
\end{aligned}$$

where the last line follows by ν being an (ε,η)-approximation of $(\mathcal{X}, Q, \mathbf{Ranges}(\mathcal{T}_{\mathcal{F}}, \succ))$. We can then split this integral on t from 0 to S into two parts at a

point $r_g = \sup\{r \geq 0 \colon Q_\succ(g,t) \geq \eta\}$. Note that for each $t \in [0, r_g)$, $Q_\succ(g,t) \geq \eta$ and for each $t > r_g$, $Q_\succ(g,t) < \eta$. Taking these observations into account, we obtain

$$\begin{aligned}\int_{t=0}^{r_g} \max\{\eta, Q_\succ(g,t)\}\mathrm{d}t &= \int_{t=0}^{r_g} Q_\succ(g,t)\mathrm{d}t \\ \text{(Lemma 2.2)} &= \int_{x\in\mathcal{X}} \min\{r_g, g(x)\}\mathrm{d}Q(x) \\ &\leq \int_{x\in\mathcal{X}} g(x)\mathrm{d}Q(x) = 1.\end{aligned}$$

and

$$\int_{t=r_g}^{S} \max\{\eta, Q_\succ(g,t)\}\mathrm{d}t = \int_{t=r_g}^{S} \eta \mathrm{d}t \leq S\eta.$$

Combining these two parts, we have

$$(8) \leq \varepsilon(1 + S\eta)$$

completing the proof. □

2.3. Proof of Theorem 1.4

When dealing with Rademacher complexity, we can leverage the helpful lemma known as Talagrand's Contraction Lemma (refer to Theorem 4.12 in (Ledoux & Talagrand, 1991)). This lemma, in particular, establishes that $\mathcal{R}_m(\phi \circ \mathcal{F}) \leq L\mathcal{R}_m(\mathcal{F})$, where $\phi \colon \mathbb{R} \longrightarrow \mathbb{R}$ is an L-Lipschitz function, and $\phi \circ \mathcal{F} = \{\phi \circ f \colon f \in \mathcal{F}\}$. Our requirement extends beyond this lemma to the following generalization (for the proof, see Appendix C.2).

Corollary 2.3. *Let $\boldsymbol{T} \subset \mathbb{R}^m$ be a bounded set, and $\varepsilon_1, \ldots, \varepsilon_m > 0$. For any L-Lipschitz functions $\phi_i \colon \mathbb{R} \longrightarrow \mathbb{R}$ with $\phi_i(0) = 0$ for $i \in [m]$, we have*

$$\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\boldsymbol{\varepsilon}} \sup_{\boldsymbol{t}\in\boldsymbol{T}} \left|\sum_{i=1}^m \varrho_i \phi_i(t_i)\right| \leq 2L \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\boldsymbol{\varepsilon}} \sup_{\boldsymbol{t}\in\boldsymbol{T}} \left|\sum_{i=1}^m \varrho_i t_i\right|,$$

where $\boldsymbol{\varepsilon} = \prod_{i=1}^m \{\pm\varepsilon_i\}$.

To establish the proof of Theorem 1.4, we first need to bound the Radamacher complexity.

Lemma 2.4. *For a well-behaved measure P (see Definition 1.3), if $s(\cdot) : \mathbb{R}^d \longrightarrow (1, \infty)$ is an upper sensitivity function for $\mathcal{L}_{\phi,k}$ with total sensitivity S, then*

$$\mathcal{R}_m^q(\mathcal{T}_{\mathcal{L}_{\phi,k}}) \leq C\sqrt{\frac{S}{m}}$$

for $C = (2LE_1 + \phi(0)) \max\left(4E_1k, \frac{1}{E_2}\right) + 8LkE_1 + 1$.

As the proof is detail-involved, we present a sketch of it here. The complete proof can be found in Appendix C.3.

Sketch of proof. we start with the two following observations

$$\begin{aligned}\mathop{\mathbb{E}}_{x_{1:m}\sim q} \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}} \left\|\sum_{i=1}^m \sigma_i \frac{x_i}{s(x_i)}\right\|_2 &\leq \sqrt{\frac{m}{S}} E_1 \\ \mathop{\mathbb{E}}_{x_{1:m}\sim q} \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}} \left|\sum_{i=1}^m \sigma_i \frac{1}{s(x_i)}\right| &\leq \sqrt{\frac{m}{S}}.\end{aligned} \tag{9}$$

Next, for $\alpha(w) = \frac{S}{\int_{\mathcal{X}} \ell_{\phi,k}(w,x)\mathrm{d}P(x)}$, we deduce that $\alpha(w)\|w\|_2^2 \leq Sk$ and

$$\alpha(w) \leq \begin{cases} \max\left(4SE_1k, \frac{S}{E_2}\right) & \|w\|_2 \leq 1 \\ \frac{kS}{2^{2n}} & 2^n \leq \|w\|_2 \leq 2^{n+1}. \end{cases}$$

For $\bar{\phi}(\cdot) = \phi(\cdot) - \phi(0)$, we obtain

$$\begin{aligned}\mathcal{R}_m^q(\mathcal{T}_{\mathcal{L}_{\phi,k}}) \leq\ & \underbrace{\mathop{\mathbb{E}}_{\substack{x_{1:m}\sim q \\ \sigma_{1:m}\sim\{-1,1\}}} \sup_{w\in\mathcal{X}} \alpha(w) \left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right|}_{=M_1} \\ &+ \underbrace{\mathop{\mathbb{E}}_{\substack{x_{1:m}\sim q \\ \sigma_{1:m}\sim\{-1,1\}}} \sup_{w\in\mathcal{X}} \alpha(w) \left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\phi(0)}{s(x_i)}\right|}_{=M_2} \\ &+ \frac{1}{k} \underbrace{\mathop{\mathbb{E}}_{\substack{x_{1:m}\sim q \\ \sigma_{1:m}\sim\{-1,1\}}} \sup_{w\in\mathcal{X}} \alpha(w)\|w\|_2^2 \left|\sum_{i=1}^m \frac{\sigma_i}{ms(x_i)}\right|}_{=N}\end{aligned}$$

Using the provided upper bound for $\alpha(w)$ and Equation (9), we conclude $M_1 \leq 2LE_1 \max\left(2E_1k, \frac{1}{E_2}\right)\sqrt{\frac{S}{m}} + 8LkE_1\sqrt{\frac{S}{m}}$ and $M_2 \leq \sqrt{\frac{S}{m}} \max\left(4E_1k, \frac{1}{E_2}\right)\phi(0)$. Utilizing the bound $\alpha(w)\|w\|_2^2 \leq Sk$ and Equation (9), we derive $N \leq \sqrt{\frac{S}{m}}k$ which completes the proof. □

The proof of Theorem 1.4 now follows immediately from Theorem 1.1 and Lemma 2.4.

2.4. Proof of results in Table 2

We here only outline the proofs for $\phi = \text{logistic}$. The proofs for other cases share similarities and are presented in full detail in Appendix D. As these results are yielded from Theorems 1.2 and 1.4, we need to know the upper sensitivity function. We start with a simple observation, providing us with a method to calculate an upper sensitivity function.

Lemma 2.5. *For a probability measure P over $\mathcal{X}$, assume that $\mathcal{W} \subseteq \mathcal{X}$, $\ell : \mathcal{X} \times \mathcal{X} \longrightarrow (0, \infty)$, and $\gamma : [0, \infty) \longrightarrow [0, \infty)$ such that $0 < \int_{\mathcal{X}} \gamma(\|x\|_2) \mathrm{d}P(x) < \infty$ and*

$$\gamma(\|x\|_2) \leq \frac{\ell(x, w)}{\ell(y, w)} \qquad \forall x, y \in \mathcal{X}, w \in \mathcal{W}. \tag{10}$$

Then $s(y) = \frac{1}{\int_{\mathcal{X}} \gamma(\|x\|_2) \mathrm{d}P(x)}$ is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}_{\mathcal{W}})$, where $\mathcal{L}_{\mathcal{W}} = \{\ell(\cdot, w) \colon w \in \mathcal{W}\}$.

Then we need the following extension of a lemma by Tolochinksy et al. (2022).

Lemma 2.6. *Let $\phi : \mathbb{R} \longrightarrow (0, \infty)$ be a non-increasing function such that*

$$\frac{\phi(-\alpha z) + \frac{z^2}{k}}{\phi(\alpha z) + \frac{z^2}{k}} \leq \beta(\alpha) \quad 0 \leq \alpha \leq B_1, 0 \leq z \leq B_2. \tag{11}$$

If we set $M = \phi(-B_1 B_2)$, then, for each $x, y, w \in \mathcal{X}$ with $\|x\|, \|y\| \leq B_1, \|w\| \leq B_2$, we have

$$\frac{\phi(0)}{M\beta(\|x\|)} \leq \frac{\ell_{\phi,k}(x, w)}{\ell_{\phi,k}(y, w)}.$$

If ϕ is universally bounded by M, then we do not need upper bounds B_1, B_2 for α, z and consequentially do not need upper bounds for $\|x\|, \|y\|$, and $\|w\|$, and the same statement holds.

With a detailed proof (see Lemma D.8), we observe that

$$\frac{\mathrm{logistic}(-\alpha z) + \frac{z^2}{k}}{\mathrm{logistic}(\alpha z) + \frac{z^2}{k}} \leq \begin{cases} \frac{85\alpha^2 k}{\log(\alpha^2 k)} & \alpha^2 k > e \\ 85 & \alpha^2 k \leq e, \end{cases}$$

for each $\alpha, z \geq 0$. Using Lemmas 2.5 and 2.6, this concludes $s(y) = 1 + \frac{170(1+kA^2)}{\sqrt{\max(1, \log(kA^2))}}$ is an upper sensitivity for $(\mathbb{R}^d,\ P,\ \mathcal{L}_{\mathrm{logistic},k})$ with probability measure P with $P(\{x \in \mathcal{X} \colon \|x\|_2 \geq A\}) = 0$ (see Lemma D.9).

Theorem 2.7. *Assume that P is probability measure over $\mathbb{R}^d$ such that $P(\{x \in \mathcal{X} \colon \|x\|_2 \geq A\}) = 0$, $S = 1 + \frac{170(1+kA^2)}{\sqrt{\max(1, \log(kA^2))}}$, and $C = (2A + 1)\max(4Ak, 2.5) + 8Ak + 1$. For $m \geq \frac{2S}{\varepsilon^2}(8C^2 + S \log \frac{2}{\delta})$, any iid sample $x_1, \ldots, x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\mathrm{logistic},k})$ with probability at least $1 - \delta$.*

Proof. We observed that $s(y) = 1 + \frac{170(1+kA^2)}{\sqrt{\max(1, \log(kA^2))}}$ serves as an upper sensitivity function for $(\mathbb{R}^d,\ P,\ \mathcal{L}_{\mathrm{logistic},k})$. Since it is a constant function, s-sensitivity sampling is equivalent to sampling according to P. It is worth noting that logistic is a 1-Lipschitz, convex, and decreasing function. For $E_1 = A$, we have

$$\mathop{\mathbb{E}}_{x \sim p} \mathrm{logistic}\left(\frac{\|x\|_2}{2E_1}\right) \geq \mathrm{logistic}\left(\frac{1}{2}\right) \geq \frac{2}{5} = E_2.$$

This implies that Definition 1.3 is satisfied for $\mathcal{L}_{\mathrm{logistic},k}$ and thus Theorem 1.4 concludes the statement for $m \geq \frac{2S}{\varepsilon^2}\left(8C^2 + S \log \frac{2}{\delta}\right)$. □

As the upper sensitivity function for $\mathcal{L}_{\mathrm{logistic},k}$ is constant, $\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\mathrm{logistic},k}}, \succ) = \mathbf{Ranges}(\mathcal{L}_{\mathrm{logistic},k}, \succ)$: every function in $\mathcal{T}_{\mathcal{L}_{\mathrm{logistic},k}}$ is a function in $\mathcal{L}_{\mathrm{logistic},k}$ scaled by a positive constant. Thus $\mathrm{VCdim}(\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\mathrm{logistic},k}}, \succ)) = \mathrm{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\mathrm{logistic},k}, \succ))$. For an $f_w \in \mathcal{L}_{\mathrm{logistic},k}$ and $r \geq 0$,

$$\begin{aligned}
\mathbf{range}(f_w, \succ, r) &= \left\{x \in \mathbb{R}^d \colon f_w(x) > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \mathrm{logistic}(\langle x, w\rangle) + \frac{\|w\|^2}{k} > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \log\left(1 + e^{-\langle x, w\rangle}\right) > \underbrace{r - \frac{\|w\|^2}{k}}_{=t}\right\} \\
&= \begin{cases} \mathbb{R}^d & t \leq 0 \\ \left\{x \in \mathbb{R}^d \colon \langle x, w\rangle < \log(e^t - 1)\right\} & t > 0, \end{cases}
\end{aligned}$$

which concludes that $\mathbf{Ranges}(\mathcal{L}_{\mathrm{logistic},k}, \succ)$ only includes half-spaces and the whole space $\mathbb{R}^d$. Therefore, by Radon's theorem, $\mathrm{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\mathrm{logistic},k}, \succ)) \leq d + 1$. Leveraging Theorem 1.2, we obtain the following theorem.

Theorem 2.8. *Let P be a probability measure over $\mathbb{R}^d$ such that $P(\{x \in \mathcal{X} \colon \|x\|_2 \geq A\}) = 0$ and $S = 1 + \frac{170(1+kA^2)}{\sqrt{\max(1, \log(kA^2))}}$. There is an $m = O\left(\frac{S}{\varepsilon^2}\left(d \log S + \log \frac{1}{\delta}\right)\right)$ such that any iid sample $x_1, \ldots, x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathbb{R}^d,\ P,\ \mathbb{R}^d,\ \ell_{\mathrm{logistic},k})$ with probability at least $1 - \delta$.*

3. Conclusion and Experimental results

This paper provides the first no dimensional sampling coresets for classification; they provides relative error for standard loss functions on linear classification with regularization, and an expectation bound on the data norm. Some results apply to iid samples from a continuous distribution, and hence imply sample complexity bounds. The key new ingredient is a Radamacher complexity bound for sensitivity sample coresets, which we expect will find further use.

The appendix shows how these results can be applied to kernelized versions of these problems, and also recover the best sampling bounds for KDE coresets.

While we do not provide new experimental evidence of the claims, our results are consistent with simulations in many previous papers. For example Tolochinksy et al. (2022) show non-uniform sensitivity sampling slightly outperforming uniform samples; but this does not contradict our results using iid samples since, for example, in those

experiments doubling the iid sample size improves upon the non-uniform sample results.

Acknowledgements

Thanks to support from NSF CCF-2115677, CDS&E-1953350, IIS-1816149, and IIS-2311954.

Impact Statement

This paper advances the field of Machine Learning, particularly in understanding the convergence rate and potential compression of classic problem formulations. While this has potential for a wide variety of impacts, including societal ones, there are none in particular that we feel demand to be highlighted here as specific consequences.

References

Anthony, M. and Bartlett, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, USA, 1st edition, 2009. ISBN 052111862X.

Aronszajn, N. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950. URL http://dx.doi.org/10.2307/1990404.

Assadi, S., Bateni, M., Bernstein, A., Mirrokni, V., and Stein, C. *Coresets Meet EDCS: Algorithms for Matching and Vertex Cover on Massive Graphs*, pp. 1616–1635. 2019. doi: 10.1137/1.9781611975482.98. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611975482.98.

Bachem, O., Lucic, M., and Krause, A. Practical coreset constructions for machine learning. *arXiv: Machine Learning*, 2017. URL https://api.semanticscholar.org/CorpusID:88517375.

Bobrowski, O., Mukherjee, S., and Taylor, J. E. Topological consistency via kernel estimation. *Bernoulli*, 23(1):288 – 328, 2017. doi: 10.3150/15-BEJ744. URL https://doi.org/10.3150/15-BEJ744.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. doi: 10.1145/130385.130401. URL https://doi.org/10.1145/130385.130401.

Braverman, V., Lang, H., Ullah, E., and Zhou, S. Improved Algorithms for Time Decay Streams. In Achlioptas, D. and Végh, L. A. (eds.), *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2019)*, volume 145 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 27:1–27:17, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-125-2. doi: 10.4230/LIPIcs.APPROX-RANDOM.2019.27. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.APPROX-RANDOM.2019.27.

Braverman, V., Feldman, D., Lang, H., Statman, A., and Zhou, S. Efficient coreset constructions via sensitivity sampling. In Balasubramanian, V. N. and Tsang, I. (eds.), *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, pp. 948–963. PMLR, 17–19 Nov 2021. URL https://proceedings.mlr.press/v157/braverman21a.html.

Bundefineddoiu, M., Har-Peled, S., and Indyk, P. Approximate clustering via core-sets. In *Proceedings of the Thiry-Fourth Annual ACM Symposium on Theory of Computing*, STOC '02, pp. 250–257, New York, NY, USA, 2002. Association for Computing Machinery. ISBN 1581134959. doi: 10.1145/509907.509947. URL https://doi.org/10.1145/509907.509947.

Charikar, M., Kapralov, M., and Waingarten, E. A quasi-monte carlo data structure for smooth kernel evaluations, 2024.

Clarkson, K., Wang, R., and Woodruff, D. Dimensionality reduction for tukey regression. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1262–1271. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/clarkson19a.html.

Cohen, M. B., Elder, S., Musco, C., Musco, C., and Persu, M. Dimensionality reduction for k-means clustering and low rank approximation. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '15, pp. 163–172, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335362. doi: 10.1145/2746539.2746569. URL https://doi.org/10.1145/2746539.2746569.

Curtin, R. R., Im, S., Moseley, B., Pruhs, K., and Samadian, A. Unconditional coresets for regularized loss minimization. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings*

of Machine Learning Research, pp. 482–492. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/samadian20a.html.

Dasgupta, A., Drineas, P., Harb, B., Kumar, R., and Mahoney, M. W. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing*, 38(5): 2060–2078, 2009. doi: 10.1137/070696507. URL https://doi.org/10.1137/070696507.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pp. 1127–1136, USA, 2006. Society for Industrial and Applied Mathematics. ISBN 0898716055.

Feldman, D. and Langberg, M. A unified framework for approximating and clustering data. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, STOC '11, pp. 569–578, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306911. doi: 10.1145/1993636.1993712. URL https://doi.org/10.1145/1993636.1993712.

Feldman, D. and Schulman, L. J. Data reduction for weighted and outlier-resistant clustering. In *Proceedings of the Twenty-Third Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '12, pp. 1343–1354, USA, 2012. Society for Industrial and Applied Mathematics.

Feldman, D., Schmidt, M., and Sohler, C. 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*, SODA '13, pp. 1434–1453, USA, 2013. Society for Industrial and Applied Mathematics. ISBN 9781611972511.

Feldman, D., Schmidt, M., and Sohler, C. Turning big data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM Journal on Computing*, 49(3):601–657, 2020. doi: 10.1137/18M1209854. URL https://doi.org/10.1137/18M1209854.

Frahling, G. and Sohler, C. Coresets in dynamic geometric data streams. In *Proceedings of the Thirty-Seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pp. 209–217, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1581139608. doi: 10.1145/1060590.1060622. URL https://doi.org/10.1145/1060590.1060622.

Frahling, G. and Sohler, C. A fast k-means implementation using coresets. *International Journal of Computational Geometry & Applications*, 18(06):605–625, 2008. doi: 10.1142/S0218195908002787. URL https://doi.org/10.1142/S0218195908002787.

Fubini, G. Sugli integrali multipli. *Reale Accademia dei Lincei, Rome*, 16(1):608–614, 1907.

Gärtner, B. and Jaggi, M. Coresets for polytope distance. In *Proceedings of the twenty-fifth annual symposium on Computational geometry*, pp. 33–42, 2009.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL http://jmlr.org/papers/v13/gretton12a.html.

Har-Peled, S. and Mazumdar, S. On coresets for k-means and k-median clustering. In *Proceedings of the Thirty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pp. 291–300, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138520. doi: 10.1145/1007352.1007400. URL https://doi.org/10.1145/1007352.1007400.

Har-Peled, S. and Sharir, M. Relative (p, ϵ)-approximations in geometry. *Discrete & Computational Geometry*, 45(3):462–496, 2011. doi: 10.1007/s00454-010-9248-1. URL https://doi.org/10.1007/s00454-010-9248-1.

Huang, L. and Vishnoi, N. K. Coresets for clustering in euclidean spaces: Importance sampling is nearly optimal. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2020, pp. 1416–1429, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450369794. doi: 10.1145/3357713.3384296. URL https://doi.org/10.1145/3357713.3384296.

Huang, L., Jiang, S. H.-C., Li, J., and Wu, X. Epsilon-coresets for clustering (with outliers) in doubling metrics. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 814–825, 2018. doi: 10.1109/FOCS.2018.00082.

Joshi, S., Kommaraji, R. V., Phillips, J. M., and Venkatasubramanian, S. Comparing distributions and shapes using the kernel distance. In *Proceedings of the Twenty-Seventh Annual Symposium on Computational Geometry*, SoCG '11, pp. 47–56, New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306829. doi: 10.1145/1998196.1998204. URL https://doi.org/10.1145/1998196.1998204.

Karnin, Z. and Liberty, E. Discrepancy, coresets, and sketches in machine learning. In Beygelzimer, A. and

Hsu, D. (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 1975–1993. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/karnin19a.html.

Lacoste-Julien, S., Lindsten, F., and Bach, F. Sequential Kernel Herding: Frank-Wolfe Optimization for Particle Filtering. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 38 of *JMLR Workshop and Conference Proceedings*, San Diego, United States, May 2015. URL https://hal.science/hal-01099197.

Langberg, M. and Schulman, L. J. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '10, pp. 598–607, USA, 2010. Society for Industrial and Applied Mathematics. ISBN 9780898716986.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces [electronic resource] : Isoperimetry and Processes / by Michel Ledoux, Michel Talagrand.* Classics in Mathematics. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 1991. edition, 1991. ISBN 3-642-20212-8.

Li, Y., Long, P. M., and Srinivasan, A. Improved bounds on the sample complexity of learning. *Journal of Computer and System Sciences*, 62(3):516–527, 2001. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.2000.1741. URL https://www.sciencedirect.com/science/article/pii/S0022000000917410.

Lopez-Paz, D., Muandet, K., Schölkopf, B., and Tolstikhin, I. Towards a learning theory of cause-effect inference. In Bach, F. and Blei, D. (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1452–1461, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/lopez-paz15.html.

Mai, T., Musco, C. N., and Rao, A. Coresets for classification – simplified and strengthened. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=qz0MLeaTP1C.

Matheny, M. and Phillips, J. M. Approximate maximum halfspace discrepancy. In *32nd International Symposium on Algorithms and Computation (ISAAC 2021)*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik, 2021.

Matousek, J. *Lectures on Discrete Geometry*. Springer-Verlag, Berlin, Heidelberg, 2002. ISBN 0387953744.

McDiarmid, C. *On the method of bounded differences*, pp. 148–188. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989. doi: 10.1017/CBO9781107359949.008.

Mercer, J. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446, 1909. ISSN 02643952. URL http://www.jstor.org/stable/91043.

Mirzasoleiman, B., Bilmes, J., and Leskovec, J. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Munteanu, A., Schwiegelshohn, C., Sohler, C., and Woodruff, D. P. On coresets for logistic regression. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6562–6571, Red Hook, NY, USA, 2018. Curran Associates Inc.

Munteanu, A., Omlor, S., and Woodruff, D. Oblivious sketching for logistic regression. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7861–7871. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/munteanu21a.html.

Mussay, B., Osadchy, M., Braverman, V., Zhou, S., and Feldman, D. Data-independent neural pruning via coresets. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gmHaEKwB.

Phillips, J. M. Coresets and sketches. In *Handbook of Discrete and Computational Geometry*, chapter 49. CRC Press, 2016.

Phillips, J. M. and Tai, W. M. *Improved Coresets for Kernel Density Estimates*, pp. 2718–2727. 2018a. doi: 10.1137/1.9781611975031.173. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611975031.173.

Phillips, J. M. and Tai, W. M. Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '18, pp. 2718–2727, USA, 2018b. Society for Industrial and Applied Mathematics. ISBN 9781611975031.

Phillips, J. M. and Tai, W. M. Near-optimal coresets of kernel density estimates. *Discrete & Computational Geometry*, 63(4):867–887, 2020. doi: 10.1007/

s00454-019-00134-6. URL https://doi.org/10.1007/s00454-019-00134-6.

Phillips, J. M., Wang, B., and Zheng, Y. Geometric Inference on Kernel Density Estimates. In Arge, L. and Pach, J. (eds.), *31st International Symposium on Computational Geometry (SoCG 2015)*, volume 34 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 857–871, Dagstuhl, Germany, 2015. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-939897-83-5. doi: 10.4230/LIPIcs.SOCG.2015.857. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SOCG.2015.857.

Rinaldo, A. and Wasserman, L. Generalized density clustering. *The Annals of Statistics*, 38(5):2678 – 2722, 2010. doi: 10.1214/10-AOS797. URL https://doi.org/10.1214/10-AOS797.

Schölkopf, B. and Smola, A. J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 06 2018. ISBN 9780262256933. doi: 10.7551/mitpress/4175.001.0001. URL https://doi.org/10.7551/mitpress/4175.001.0001.

Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. 2015.

Sener, O. and Savarese, S. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1aIuk-RW.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014. ISBN 1107057132.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

Sohler, C. and Woodruff, D. P. Strong coresets for k-median and subspace approximation: Goodbye dimension. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 802–813, 2018. doi: 10.1109/FOCS.2018.00081.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Tai, W. M. Optimal Coreset for Gaussian Kernel Density Estimation. In Goaoc, X. and Kerber, M. (eds.), *38th International Symposium on Computational Geometry (SoCG 2022)*, volume 224 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 63:1–63:15, Dagstuhl, Germany, 2022. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-227-3. doi: 10.4230/LIPIcs.SoCG.2022.63. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.SoCG.2022.63.

Taylor, C. C. Nonparametric Density Estimation: The L1 View. *Royal Statistical Society. Journal. Series A: General*, 148(4):392–393, 12 2018. ISSN 0035-9238. doi: 10.2307/2981908. URL https://doi.org/10.2307/2981908.

Tolochinksy, E., Jubran, I., and Feldman, D. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21520–21547. PMLR, 17–23 Jul 2022.

Tsang, I. W., Kwok, J. T., Cheung, P.-M., and Cristianini, N. Core vector machines: Fast svm training on very large data sets. *Journal of Machine Learning Research*, 6(4), 2005.

Zheng, Y. and Phillips, J. M. l_∞ error and bandwidth selection for kernel density estimates of large data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pp. 1533–1542, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783357. URL https://doi.org/10.1145/2783258.2783357.

A. Generalized Framework: Reproducing kernel Hilbert Space

We can establish the validity of our findings in a broader framework by employing reproducing kernels. In particular, the above-mentioned results are automatically derived when considering $\mathcal{X} = \mathbb{R}^d$ with the linear kernel $\mathsf{K}(x, w) = \langle x, w\rangle$ as a reproducing kernel Hilbert space H. For those less familiar with reproducing kernel Hilbert spaces, maintaining this latter setting throughout the entire paper is helpful.

A.1. Kernelization of the Results

The kernel method is known as a powerful technique in machine learning that enables the handling of non-linear relationships in data by implicitly mapping it to a higher-dimensional space. This flexibility and efficiency make it a widely used approach, especially in scenarios where linear methods may not be sufficient. Reproducing kernel Hilbert spaces (RKHS) are intimately connected to kernel methods. An RKHS is a type of Hilbert space associated with a positive definite kernel function. In the context of machine learning, we can think that the input space is mapped into an RKHS through a feature map. Broadly, the RKHS framework provides a mathematical foundation for understanding how non-linear relationships in data can be effectively captured through implicit mappings into higher-dimensional spaces. The key insight in the context of kernel methods is that the kernel function implicitly represents the inner product in the Hilbert space, for more see (Mercer, 1909; Aronszajn, 1950; Boser et al., 1992; Steinwart & Christmann, 2008; Schölkopf & Smola, 2018). The following definition extends Definition 1.3 to the context of RKHS cases.

Definition A.1. Let H be a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, P a probability measure over $\mathcal{X}$, $\phi\colon \mathbb{R} \longrightarrow [0, \infty)$ a non-increasing L-Lipschitz function, and $k > 0$ a constant. We say P is *well-behaved* (w.r.t ϕ and K) if

$$\mathop{\mathbb{E}}_{x\sim P} \mathsf{K}(x, x) \leq E_1^2 \qquad \text{and} \qquad \mathop{\mathbb{E}}_{x\sim P} \phi\left(\frac{\sqrt{\mathsf{K}(x, x)}}{2E_1}\right) \geq E_2, \tag{12}$$

where E_1, E_2 are two positive constants. Define

$$\mathcal{L}^H_{\phi,k} = \left\{\ell^H_{\phi,k}(w, \cdot) = \phi(\mathsf{K}(w, \cdot)) + \frac{1}{k}\mathsf{K}(w, w)\colon w \in \mathcal{X}\right\} \tag{13}$$

and

$$\bar{\mathcal{L}}^H_{\phi,k} = \left\{\bar{\ell}^H_{\phi,k}(w, \cdot) = \phi(-\mathsf{K}(w, \cdot)) + \frac{1}{k}\mathsf{K}(w, w)\colon w \in \mathcal{X}\right\} \tag{14}$$

To keep things simpler, we occasionally employ $\|X\|_H$ and $\langle\cdot, \cdot\rangle_H$ instead of $\sqrt{\mathsf{K}(x, x)}$ and $\mathsf{K}(\cdot, \cdot)$ respectively.

Remark A.2. Throughout the paper, we mainly focus on coresets for $\mathcal{L}^H_{\phi,k}$. However, we will see that our analysis effortlessly extends to $\bar{\mathcal{L}}^H_{\phi,k}$. This consideration is insignificant when dealing with the standard inner product as the kernel, given that $-\langle w, \cdot\rangle = \langle\cdot, -w\rangle$. However, this property does not hold for kernels in general.

Theorem A.3 (Theorem 1.4, Kernelized Representation)**.** *For a well-behaved probability measure P with respect to a Hilbert space H of real-valued functions defined on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, if $s(\cdot) : \mathcal{X} \longrightarrow (1, \infty)$ is an upper sensitivity function for $\mathcal{L}^H_{\phi,k}$ with total sensitivity S and $m \geq \frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right)$, then any s-sensitivity sample $x_1, \ldots, x_m$ from $\mathcal{X}$ with weights $u_i = \frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathcal{X},\ P,\ \mathcal{X},\ \ell^H_{\phi,k})$ with probability at least $1 - \delta$, where*

$$C = (2LE_1 + \phi(0)) \max\left(2E_1k, \frac{1}{E_2}\right) + 8LkE_1 + 1.$$

The statement remains true if we replace $\mathcal{L}^H_{\phi,k}$ and $(\mathcal{X},\ P,\ \mathcal{X},\ \ell^H_{\phi,k})$ by $\bar{\mathcal{L}}^H_{\phi,k}$ and $(\mathcal{X},\ P,\ \mathcal{X},\ \bar{\ell}^H_{\phi,k})$.

Considering $\mathbb{R}^d$ as a Hilbert space equipped with linear kernel, we derive Theorem 1.4 as a corollary of this theorem. In the following, we present various applications derived from this theorem.

Theorem A.4. *Let us consider H, a reproducing kernel Hilbert space of real-valued functions defined on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$, and P, a probability measure over $\mathcal{X}$. Given the conditions outlined in the initial 4 columns of Table 3, any s-sensitivity sample of size m (sampling according to the distribution provided in Column 5) guarantees an ε-coreset for $(\mathcal{X}, P, \mathcal{X}, \mathcal{L}^H_{\phi,k})$ with a probability of at least $1 - \delta$.*

Table 3. Coreset size for $(\mathcal{X}, P, \mathcal{X}, \mathcal{L}^H_{\phi,k})$ where P is a probability measure over $\mathcal{X}$ and H is an RKHS.

function ϕ	data assumption	total sens. S	constant C	sampling	Coreset size	reference
$\sigma(\pm\cdot)$	$\mathbb{E}_{x\sim P}\|x\|_H^2 \leq E_1^2$	$O(E_1^2 k)$	$O(E_1^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.6
logistic$(\pm\cdot)$	$P(\|x\|_H \geq A) = 0$	$O(\frac{A^2 k}{\sqrt{\log(A^2 k)}})$	$O(A^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.10
logistic$(\pm\cdot)$	$\mathbb{E}_{x\sim P}\|x\|_H^2 \leq E_1$	$O(\frac{E_1^2 k}{\sqrt{\log(E_1^2 k)}})$	$O(E_1^2 k)$	$\frac{s}{S}p$	$\frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right))$	Thm D.13
svm$(\pm\cdot)$	$P(\|x\|_H \geq A) = 0$	$O(A^2 k)$	$O(A^2 k)$	p	$\frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$	Thm D.18
svm$(\pm\cdot)$	$\mathbb{E}_{x\sim P}\|x\|_H^2 \leq E_1$	$O(E_1^2 k)$	$O(E_1^2 k)$	$\frac{s}{S}p$	$\frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right))$	Thm D.22

The first five columns of each row in Table 3 outline assumptions concerning H and P, while the last two columns detail sample complexity and the respective theorem in the paper where the result is introduced. Specifically, under the conditions specified in the initial four columns, a sample of size m (as per the sampling distribution in Column 5) produces an ε-coreset with a probability of $1-\delta$. It is worth emphasizing once more that the utilization of p in Column 5 signifies that sampling is directly conducted from the data distribution, specifically through uniform sampling from the provided data points. This eliminates the necessity for sensitivity computations in this context.

A.1.1. **Kernel Density Estimate**

Let P be a probability measure over $\mathbb{R}^d$ and $\mathsf{K} : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ be a kernel, for instance the Gaussian kernel $\mathsf{K}(x,w) = \exp(-\|x-w\|^2)$. At any point $w \in \mathbb{R}^d$, the kernel density estimate is defined ad $\mathrm{KDE}_P(w) = \int_{x\in\mathbb{R}^d} \mathsf{K}(x,w)\mathrm{d}P(x)$. When we have finite number of points $X = \{x_1, \ldots, x_n\}$ given as the data set, we can assume that P is a uniform probability measure over X, i.e., $P(x = x_i) = \frac{1}{n}$. In this scenario, we use KDE_X instead of KDE_P and and the expression for KDE_X is given by

$$\mathrm{KDE}_X(w) = \frac{1}{n}\sum_{x\in X} \mathsf{K}(x,w).$$

The evaluation of KDE_X demands O(n) time, which can become impractical for massive datasets. Therefore, a frequently employed approach is to substitute X with a significantly smaller dataset Y, allowing KDE_Y to function as an approximation for KDE_X. Formally, for a given $\varepsilon \in (0,1)$, we look for a small size Y such that

$$\sup_{w\in\mathbb{R}^d} |\mathrm{KDE}_X(w) - \mathrm{KDE}_Y(w)| \leq \varepsilon.$$

This challenge has been thoroughly explored and investigated in various studies such as (Phillips & Tai, 2020; Tai, 2022; Charikar et al., 2024; Bobrowski et al., 2017; Taylor, 2018; Gretton et al., 2012; Phillips et al., 2015; Rinaldo & Wasserman, 2010; Scott, 2015; Silverman, 1986; Zheng & Phillips, 2015). Our primary result in this context is presented in the following theorem, restated as Corollary D.19 along with its proof.

Theorem A.5. *Let $\mathsf{K} : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow (0,1]$ be a reproducing kernel, i.e., a kernel associated with an RKHS, and P be a probability measure over $\mathbb{R}^d$. For $\varepsilon, \delta \in (0,1)$, there exists a universal constant C (independent of d and K) such that if $m \geq \frac{C}{\varepsilon^2}\log\frac{1}{\delta}$, then, with probability at least $1-\delta$, for any iid random sample $X = \{x_1, \ldots, x_m\}$ based on P, we have*

$$\sup_{w\in\mathbb{R}^d} |\mathrm{KDE}_P(w) - \mathrm{KDE}_X(w)| \leq \varepsilon.$$

Note that here, we treat d as a variable, and importantly, our bound is independent of d. The Gaussian, Laplacian, and Exponential kernels are typically chosen for study. Compared to the state-of-the-art results, we can reproduce the findings of (Lopez-Paz et al., 2015) and (Lacoste-Julien et al., 2015). Nevertheless, there are some improvements in these results, such as works by (Phillips & Tai, 2018a; 2020; Karnin & Liberty, 2019), and more recent papers by (Tai, 2022; Charikar et al., 2024). It is worth noting that some of these results consider d as a constant, while others treat it as a variable.

B. Related Works

In machine learning, a common scenario involves having a set of n data points and the goal of minimizing a data-driven objective loss function. At times, for scalability reasons, it becomes necessary to choose a smaller subset of $m \ll n$ points.

The aim is to minimize the objective function on these points, potentially using non-uniform weights for the selected points. This approach aims to achieve a near-optimal solution for the entire data set. Coresets are an important tool in scalable machine learning, providing a means to achieve this objective. Coresets have found application in various problem domains, such as clustering (Bundefineddoiu et al., 2002; Har-Peled & Mazumdar, 2004; Frahling & Sohler, 2005; 2008; Feldman & Langberg, 2011; Feldman & Schulman, 2012; Feldman et al., 2013; Braverman et al., 2019; Huang & Vishnoi, 2020; Feldman et al., 2020), linear regression (Drineas et al., 2006; Dasgupta et al., 2009; Clarkson et al., 2019), principal component analysis (Cohen et al., 2015; Feldman et al., 2020), and more (Bachem et al., 2017; Sener & Savarese, 2018; Munteanu et al., 2018; Sohler & Woodruff, 2018; Phillips & Tai, 2018b; Huang et al., 2018; Assadi et al., 2019; Mussay et al., 2020).

Improving the prior result by (Feldman & Langberg, 2011) which suggests that common sensitivity sampling requires $O(dS^2)$ samples, (Braverman et al., 2021) and (Feldman et al., 2020) proved that $O(dS \log S)$ would be sufficient. Here, S represents the total sensitivity, and d is some VC-dimension associated with the function space. The proof of these results benefits from a connection between approximating a class of functions and approximating their linked-ranges spaces, the same concept established by (Joshi et al., 2011) in the study of ε-coresets for kernel density estimates. It is important to note that these results are subject to the assumption that the space is finite. We eliminate this constraint in Theorem 1.2 with a different and more straightforward proof outlined in Section 2.2. It is notable to highlight that the coreset bounds derived from this theorem are all dependent on the dimension. More importantly, we provide a counterpart of this theorem (see Theorem 1.1) using Rademacher complexity in place of VC-dimension, marking the first result of this kind. As a result, we are able to present several coreset bounds independent of the dimension – the first such results of this type (see Table 2). In the following, we focus more on the related works pertaining to these findings.

Our work shares some connections with the research conducted by (Munteanu et al., 2018; Mai et al., 2021; Tolochinksy et al., 2022), and (Curtin et al., 2020). (Munteanu et al., 2018) demonstrated the existence of datasets for which coresets of sublinear size do not exist. They introduced a complexity measure $\mu(X)$ for data points X, quantifying the difficulty associated with compressing a dataset for logistic regression. They demonstrated the existence of coresets with size $O^*(\frac{d^{3/2}\mu(X)\sqrt{|X|}}{\varepsilon^2})$ and $O^*(\frac{d^3\mu^2(X)}{\varepsilon^4})$ through a random sampling procedure ($\emptyset^*(\cdot)$ hides some logarithmic factors in terms of the problem's parameters). As a downside of their method, it is not clear how to compute $\mu(X)$ and also they conjectured that computing the value of $\mu(X)$ in general is hard. Moreover, their second bound is roughly quadratic in terms of $\mu(X)$ and $\frac{1}{\varepsilon^2}$ when the data possesses a small μ-complexity. Following their work, (Mai et al., 2021) improved these bounds to $O^*(\frac{d\mu^2(X)}{\varepsilon^2})$ which is linear in terms of $\frac{1}{\varepsilon^2}$. Their method relies on subsampling data points with probabilities proportional to their ℓ_1 Lewis weights. As they utilize the complexity measure $\mu(X)$ in their context, which cannot be directly interpreted in terms of our variables, presenting a meaningful comparison between our findings and theirs becomes challenging. Nevertheless, as an advantage, some of our bounds are independent of dimension, surpassing their results in this particular aspect.

The most closed works to ours are (Tolochinksy et al., 2022) and (Curtin et al., 2020). With an almost similar setting as ours, but restricted to Euclidean space and with a bounded parameter space, (Tolochinksy et al., 2022) found the coresets of size $O^*\left(\frac{kd^2 \log n \log k}{\varepsilon^2}\right)$ for $\phi = \text{logistic}, \sigma, \text{svm}$. Enhancing some of these results, (Curtin et al., 2020) conducted an analysis of sampling-based coreset constructions designed for regularized loss minimization problems (logistic regression or SVM). Their context is slightly distinct from ours. To rephrase their setting in the context of ours, we can express that in our setting, the regularization term is $\frac{1}{k}$ whereas they roughly use $\frac{1}{n^{1-\kappa}}$ for it (up to a constant). In the scenario where k is proportional to $n^{1-\kappa}$ for a fixed $\kappa \in (0,1)$, they have demonstrated that a uniform sample of size $O^*\left(\frac{dn^{1-\kappa}}{\epsilon^2}\right)$ functions as a coreset with high probability. Here, $n = |X|$ and d represents a type of VC-dimension linked to the function space (refer to the paper for the precise definition).They have further established the tightness of uniform sampling, accurate up to poly-logarithmic factors. A special case of our results not only reproduces their results but also improves upon theirs for logistic regression. It is still noteworthy that their bounds are dimension-dependent, whereas we present some bounds independent of dimension. For a comprehensive comparison, refer to Table 1.

A few other variants of coresets for classification exist; we briefly mention a few. Munteanu et al. (2021) applies sketching to create data compression for logistic regression. Their methods are data oblivious and show improvements for sparse high-dimensional vectors. These results are still polynomial in data parameter $\mu(X)$ and dimension d. Mirzasoleiman et al. (2020) considers coresets for regularized classification in the context of incremental gradient descent. They greedily select a coreset to be used in IGD, and their results use that the $\frac{1}{k}\|w\|^2$ regularizer makes loss function strongly convex as a

function of k. Coresets can also be found for the $0/1$ mis-classification function of size $O(d/\varepsilon^2)$ via classic VC-dimension arguments, but as this cost function is non-convex over the model parameters w, the best known algorithms (Matheny & Phillips, 2021) to solve for the approximately optimal solution are still exponential in dimension $O^*(1/\varepsilon^d)$. Or if a dataset is known to be linearly separable, then greedy coresets based on Frank-Wolfe can be found of size roughly $O(1/\varepsilon)$ which approximately preserve that separation margin (Tsang et al., 2005; Gärtner & Jaggi, 2009).

C. Main Tools

This section consists of our primary tools and their accompanying proofs essential for establishing our other key findings.

C.1. Proof of Theorem 1.1

We begin by revisiting McDiarmid's inequality, which serves as a concentration inequality that bounds the difference between the sampled mean and the true mean of a specific function. If $(\mathcal{X}_1, \Sigma_1, P_1), \ldots, (\mathcal{X}_n, \Sigma_1, P_n)$ are probability measure spaces, then $P = \prod_{i=1}^n P_i$ is a probability measure over $\prod_{i=1}^n \mathcal{X}_i$. A function $f : \prod_{i=1}^n \mathcal{X}_i \longrightarrow \mathbb{R}$ satisfies the bounded differences property if there are constants $c_1, c_2, \ldots, c_n$ such that for all $i \in [n]$, and for all $x_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$, $\ldots$, $x_n \in \mathcal{X}_n$

$$\sup_{x_i' \in \mathcal{X}_i} |f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i.$$

Lemma C.1 (McDiarmid 1989)**.** *Let $f : \prod_{i=1}^n \mathcal{X}_i \longrightarrow \mathbb{R}$ satisfy the bounded differences property with bounds $c_1, c_2, \ldots, c_n$. Consider independent random variables $X_1, X_2, \ldots, X_n$ where $X_i \in \mathcal{X}_i$ for all i. Then, for any $\varepsilon > 0$,*

$$P\left(f(X_1, X_2, \ldots, X_n) - \mathbb{E}[f(X_1, X_2, \ldots, X_n)] \geq \varepsilon\right) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

We are all set to demonstrate our findings for this subsection

Theorem C.2 (Theorem 1.1, Restated)**.** *Let $(\mathcal{X}, P, \mathcal{F})$ be a positive definite tuple and $s(\cdot)$ be an upper sensitivity function for it with the total sensitivity S. Any s-sensitivity sample from $\mathcal{X}$ of size m, with probability at least $1 - 2\exp\left(-\frac{2mt^2}{S^2}\right)$, satisfies*

$$\left|\int_x f(x)\mathrm{d}P(x) - \sum_{i=1}^m \frac{S}{ms(x_i)} f(x_i)\right| \leq (2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}) + t)\int_x f(x)\mathrm{d}P(x) \qquad \forall f \in \mathcal{F}.$$

Proof. In view of Equation (3), we need to show that, for any iid sample $x_1, \ldots, x_m \in \mathcal{X}$ according to the sensitivity normalized distribution $q(x) = \frac{s(x)}{S}p(x)$, with probability at least $1 - 2\exp\left(-\frac{2mt^2}{S^2}\right)$,

$$\sup_{f \in \mathcal{F}} \left|1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right| \leq 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}) + t.$$

For every $x \in \mathcal{X}$ and $f \in \mathcal{F}$, we have $\frac{f(x)}{\int_\mathcal{X} f(z)\mathrm{d}P(z)} \leq s(x)$ which implies $0 \leq T_f(x) = \frac{Sf(x)}{s(x)\int_\mathcal{X} f(z)\mathrm{d}P(z)} \leq S$. Therefore, $\sup_{x \in \mathcal{X}} T_f(x) \leq S$. Now, define

$$g(x_1, \ldots, x_m) = \sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right].$$

For each $i \in [m]$ and each $(x'_1, \ldots, x'_m) \in \mathcal{X}^m$ such that $x'_j = x_j$ for $j \neq i$, observe that

$$\begin{aligned}
g(x_1, \ldots, x_m) - g(x'_1, \ldots, x'_m) &= \sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right] - \sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x'_i)\right] \\
&\leq \sup_{f \in \mathcal{F}} \left[\left(1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right) - \left(1 - \frac{1}{m}\sum_{i=1}^m T_f(x'_i)\right)\right] \\
&\leq \sup_{f \in \mathcal{F}} \left(\frac{1}{m}\sum_{i=1}^m T_f(x_i) - \frac{1}{m}\sum_{i=1}^m T_f(x'_i)\right) \\
&\leq \sup_{f \in \mathcal{F}} \left|\frac{T_f(x_i) - T_f(x'_i)}{m}\right| \leq \frac{S}{m}.
\end{aligned}$$

We can repeat the above argument to show $g(x_1, \ldots, x_m) - g(x'_1, \ldots, x'_m) \leq \frac{S}{m}$. Thus $|g(x_1, \ldots, x_m) - g(x'_1, \ldots, x'_m)| \leq \frac{S}{m}$. Therefore, using McDiarmid's Inequality, with a probability at least $1 - \exp\left(-\frac{2mt^2}{S^2}\right)$, we have

$$\sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right] \leq \underbrace{\mathop{\mathbb{E}}_{x_{1:m} \sim q} \left(\sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right]\right)}_{=A} + t.$$

Similarly, we can prove that, with probability at least $1 - \exp\left(-\frac{2mt^2}{S}\right)$, we have

$$\sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m T_f(x_i) - 1\right] \leq \underbrace{\mathop{\mathbb{E}}_{x_{1:m} \sim q} \left(\sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m T_f(x_i) - 1\right]\right)}_{=B} + t.$$

Consequently, since $A, B \geq 0$, with probability at least $1 - 2\exp\left(-\frac{2mt^2}{S^2}\right)$ we have

$$\sup_{f \in \mathcal{F}} \left|1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right| \leq \max\{A, B\} + t.$$

Let us now deal with $\max\{A, B\}$. As $\mathbb{E}_{x \sim q}(T_f(x)) = 1$, we can write

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_{1:m} \sim q} \sup_{f \in \mathcal{F}} \left[1 - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right] \\
&\quad= \mathop{\mathbb{E}}_{x_{1:m} \sim q} \sup_{f \in \mathcal{F}} \left[\mathop{\mathbb{E}}_{x'_{1:m} \sim q} \frac{1}{m}\sum_{i=1}^m T_f(x'_i) - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right] \\
&\quad\leq \mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{x'_{1:m} \sim q} \sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m T_f(x'_i) - \frac{1}{m}\sum_{i=1}^m T_f(x_i)\right] \\
(*) \qquad &\quad= \mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{x'_{1:m} \sim q} \mathop{\mathbb{E}}_{\sigma_{1:m} \sim \{-1,1\}} \sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m \sigma_i \left(T_f(x'_i) - T_f(x_i)\right)\right] \\
&\quad\leq \mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{x'_{1:m} \sim q} \mathop{\mathbb{E}}_{\sigma_{1:m} \sim \{-1,1\}} \left(\sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m \sigma_i T_f(x'_i)\right] + \sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m -\sigma_i T_f(x_i)\right]\right) \\
&\quad= 2 \mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{\sigma_{1:m} \sim \{-1,1\}} \sup_{f \in \mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^m \sigma_i T_f(x_i)\right] \\
&\quad= 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}).
\end{aligned}$$

The equality marked by $(*)$ is true since σ_i indeed exchanges x_i and x_i' together, and since they are iid samples from q it cannot change the expectation. With the same approach, we would have

$$\mathop{\mathbb{E}}_{x_{1:m}\sim q} \sup_{f\in\mathcal{F}} \left[\frac{1}{m}\sum_{i=1}^{m} T_f(x_i) - 1\right] \leq 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F}),$$

which concludes $\max\{A, B\} \leq 2\mathcal{R}_m^q(\mathcal{T}_\mathcal{F})$ completing the proof. □

C.2. Contraction Lemma

When dealing with Rademacher complexity, we can leverage the helpful lemma known as Talagrand's Contraction Lemma (refer to Theorem 4.12 in (Ledoux & Talagrand, 1991)). This lemma, in particular, establishes that $\mathcal{R}_m(\phi\circ\mathcal{F}) \leq L\mathcal{R}_m(\mathcal{F})$, where $\phi\colon \mathbb{R}\longrightarrow\mathbb{R}$ is an L-Lipschitz function, and $\phi\circ\mathcal{F} = \{\phi\circ f\colon f\in\mathcal{F}\}$. Our requirement extends beyond this lemma to a generalization.

Lemma C.3. *Let $\Psi : \mathbb{R}^{\geq 0}\longrightarrow\mathbb{R}^{\geq 0}$ be convex and increasing, $\boldsymbol{T}\subset\mathbb{R}^m$ a bounded set, and $\varepsilon_1,\dots,\varepsilon_m > 0$. For any 1-Lipschitz functions $\phi_i\colon\mathbb{R}\longrightarrow\mathbb{R}$ with $\phi_i(0) = 0$ for each $i\in[m]$, we have*

$$\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\frac{1}{2}\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left|\sum_{i=1}^m \varrho_i\phi_i(t_i)\right|\right) \leq \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left|\sum_{i=1}^m \varrho_i t_i\right|\right),$$

where $\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}$ means m independent samples $\varrho_1,\dots,\varrho_m$ such that each ϱ_i is uniformly drawn from $\{-\varepsilon_i,\varepsilon_i\}$.

Before turning to the proof of this lemma, note that the only difference between this lemma and Theorem 4.12 in (Ledoux & Talagrand, 1991) is that ϱ_i here takes its value randomly and uniformly in $\{+\varepsilon_i,-\varepsilon_i\}$ rather than $\{+1,-1\}$. However, the proof is almost identical to that of Theorem 4.12 in (Ledoux & Talagrand, 1991).

Proof of Lemma C.3. We first prove

$$\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\sum_{i=1}^m \varrho_i\phi_i(t_i)\right) \leq \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\sum_{i=1}^m \varrho_i t_i\right). \tag{15}$$

To this end, it siffices to show

$$\mathop{\mathbb{E}}_{\varrho\in\{\pm\varepsilon\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}(t_1+\varrho\phi(t_2))\right) \leq \mathop{\mathbb{E}}_{\varrho\in\{\pm\varepsilon\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}(t_1+\varrho t_2)\right), \tag{16}$$

where $\boldsymbol{t} = (t_1,t_2)\subset\mathbb{R}^2$ is a bounded set and ϕ is a 1-Lipschitz function. It suffices to prove that for all $\boldsymbol{t} = (t_1,t_2)$, $\boldsymbol{s} = (s_1,s_2)\in\boldsymbol{T}$, we have

$$\mathop{\mathbb{E}}_{\varrho\in\{\pm\varepsilon\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}(t_1+\varrho t_2)\right) \geq \frac{1}{2}\Psi(t_1+\varepsilon_2\phi(t_2)) + \frac{1}{2}\Psi(s_1-\varepsilon_2\phi(s_2)) = I.$$

Since Ψ is increasing, without loss of generality, we may assume that

$$t_1+\varepsilon_2\phi(t_2) \geq s_1+\varepsilon_2\phi(s_2) \qquad \text{and} \qquad s_1-\varepsilon_2\phi(s_2) \geq t_1-\varepsilon_2\phi(t_2). \tag{17}$$

We distinguish between the following cases.

1. **Case $t_2, s_2 \geq 0$.**
   - If $s_2\leq t_2$, set $a = s_1-\varepsilon_2\phi(s_2)$, $b = s_1-\varepsilon_2 s_2$, $a' = t_1+\varepsilon_2 t_2$, $b' = t_1+\varepsilon_2\phi(t_2)$. Since ϕ is 1-Lipschitz, $\phi(0) = 0$, and $s_2\geq 0$, $|\phi(s_2)|\leq s_2$ which concludes $a\geq b$. Also, by (17), it implies $b'\geq s_1+\varepsilon_2\phi(s_2)\geq s_1-\varepsilon_2 s_2 = b$. Furthermore, as $s_2\leq t_2$, we have $\phi(t_2)-\phi(s_2)\leq(t_2-s_2)$ and hence
   $$a-b = \varepsilon_2(s_2-\phi(s_2)) \leq \varepsilon_2(t_2-\phi(t_2)) = a'-b'.$$

Since Ψ is convex and increasing, for any arbitrary fixed $x > 0$, the function $\Psi(\cdot + x) - \Psi(\cdot)$ is increasing. Setting $x = a - b \geq 0$, since $b \leq b'$, we have

$$\Psi(a) - \Psi(b) \leq \Psi(b' + (a - b)) - \Psi(b')$$

(if $a = b$, it is true always). As $b' + a - b \leq a'$, we obtain $\Psi(b' + (a - b)) - \Psi(b') \leq \Psi(a') - \Psi(b')$ and consequently, $\Psi(a) - \Psi(b) \leq \Psi(a') - \Psi(b')$ which is equivalent to $2I \leq \Psi(t_1 + \varepsilon_2 t_2) + \Psi(s_1 - \varepsilon_2 s_2)$.

- If $t_2 \leq s_2$, set $a = t_1 + \varepsilon_2\phi(t_2)$, $b = t_1 - \varepsilon_2 t_2$, $a' = s_1 + \varepsilon_2 s_2$, $b' = s_1 - \varepsilon_2\phi(s_2)$. Since ϕ is 1-Lipschitz, $|\phi(t_2)| \leq t_2$ which implies $a \geq b$. Again, by (17),

$$b' = s_1 - \varepsilon_2\phi(s_2) \geq t_1 - \varepsilon_2\phi(t_2) \geq t_1 - \varepsilon_2 t_2 = b.$$

As $\phi(t_2) - \phi(s_2) \leq s_2 - t_2$,

$$a - b = \varepsilon_2(t_2 + \phi(t_2)) \leq \varepsilon_2(s_2 + \phi(s_2)) = a' - b'.$$

Again, as $\Psi(\cdot + x) - \Psi(\cdot)$ is increasing for each fixed $x > 0$ and $b \leq b'$, by setting $x = a - b \geq 0$, we have

$$\Psi(a) - \Psi(b) \leq \Psi(b' + (a - b)) - \Psi(b')$$

(if $a = b$, it is true always). As $b' + a - b \leq a'$, we obtain $\Psi(b' + (a - b)) - \Psi(b') \leq \Psi(a') - \Psi(b')$ and consequently, $\Psi(a) - \Psi(b) \leq \Psi(a') - \Psi(b')$ which is equivalent to $2I \leq \Psi(s_1 + \varepsilon_2 s_2) + \Psi(t_1 - \varepsilon_2 t_2)$.

2. **Case $t_2, s_2 \leq 0$.** This case is similar to the previous case.

3. **Case $t_2 \geq 0, s_2 \leq 0$.** Since Ψ is increasing and $\phi(t_2) \leq t_2, -\phi(s_2) \leq -s_2$, we have

$$2I \leq \Psi(t_1 + \varepsilon_2 t_2) + \Psi(s_1 - \varepsilon_2 s_2).$$

4. **Case $t_2 \leq 0, s_2 \geq 0$.** This case follows with a similar argument to Case (3). This completes the proof of (16) and consequently the proof of (15).

To conclude the theorem, by convexity of Ψ,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\frac{1}{2}\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left|\sum_{i=1}^m \varrho_i\phi_i(t_i)\right|\right) &\leq \frac{1}{2}\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left[\sum_{i=1}^m \varrho_i\phi_i(t_i)\right]^+\right) \\
&\quad + \frac{1}{2}\mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left[\sum_{i=1}^m \varrho_i\phi_i(t_i)\right]^-\right) \\
&\leq \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left[\sum_{i=1}^m \varrho_i\phi_i(t_i)\right]^+\right) \\
&= \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\left[\sup_{\boldsymbol{t}\in\boldsymbol{T}}\sum_{i=1}^m \varrho_i\phi_i(t_i)\right]^+\right) \\
(*) \quad &\leq \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\left[\sup_{\boldsymbol{t}\in\boldsymbol{T}}\sum_{i=1}^m \varrho_i t_i\right]^+\right) \\
&\leq \mathop{\mathbb{E}}_{\varrho_{1:m}\sim\prod_{i=1}^m\{\pm\varepsilon_i\}} \Psi\left(\sup_{\boldsymbol{t}\in\boldsymbol{T}}\left|\sum_{i=1}^m \varrho_i t_i\right|\right),
\end{aligned}$$

where $(*)$ is true by applying (15) on the convex and increasing function $\Psi(\max(0, \cdot))$. This completes the proof. □

Corollary C.4. *Let $\boldsymbol{T} \subset \mathbb{R}^m$ be a bounded set, and $\varepsilon_1, \ldots, \varepsilon_m > 0$. For any L-Lipschitz functions $\phi_i \colon \mathbb{R} \longrightarrow \mathbb{R}$ with $\phi_i(0) = 0$ for each $i \in [m]$, we have*

$$\mathop{\mathbb{E}}_{\varrho_{1:m} \sim \prod_{i=1}^m \{\pm\varepsilon_i\}} \sup_{\boldsymbol{t} \in \boldsymbol{T}} \left| \sum_{i=1}^m \varrho_i \phi_i(t_i) \right| \leq 2L \mathop{\mathbb{E}}_{\varrho_{1:m} \sim \prod_{i=1}^m \{\pm\varepsilon_i\}} \sup_{\boldsymbol{t} \in \boldsymbol{T}} \left| \sum_{i=1}^m \varrho_i t_i \right|.$$

Proof. Setting Ψ as identity and replacing ϕ_i with $\frac{\phi_i}{L}$, we conclude the result by Lemma C.3. □

In our setting, we only deal with the case that each $\varepsilon_i \in (0, 1]$. This particular case can be directly inferred from Theorem 4.12 in (Ledoux & Talagrand, 1991). Nevertheless, Lemma C.3 and Corollary C.4, in the presented general form, can be of independent interest.

C.3. Proof of Theorem A.3

This subsection primarily focuses on proving Theorem A.3 (and Theorem 1.4), a crucial tool enabling us to derive our main results. We start with bounding the Radamacher complexity of $\mathcal{T}_{\mathcal{F}}$.

Lemma C.5. *For a well-behaved measure P (see Definition A.1), if $s(\cdot) : \mathcal{X} \longrightarrow (1, \infty)$ is an upper sensitivity function for $\mathcal{L}^H_{\phi,k}$ with total sensitivity $S = \int_{\mathcal{X}} s(x) \mathrm{d}P$, then*

$$\mathcal{R}^q_m(\mathcal{T}_{\mathcal{L}^H_{\phi,k}}) \leq C \sqrt{\frac{S}{m}},$$

where $C = (2LE_1 + \phi(0)) \max\left(4E_1 k, \frac{1}{E_2}\right) + 8LkE_1 + 1$. The same statement holds if we replace $\mathcal{L}^H_{\phi,k}$ by $\bar{\mathcal{L}}^H_{\phi,k}$.

Proof. Due to similarity, we only work with $\mathcal{L}_{\phi,k}$. We first show that

$$\mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{\sigma_{1:m} \sim \{-1,1\}} \left\| \sum_{i=1}^m \sigma_i \frac{\mathsf{K}(x_i, \cdot)}{s(x_i)} \right\|_H \leq \sqrt{\frac{m}{S}} E_1 \quad \text{and} \quad \mathop{\mathbb{E}}_{x_{1:m} \sim q} \mathop{\mathbb{E}}_{\sigma_{1:m} \sim \{-1,1\}} \left| \sum_{i=1}^m \sigma_i \frac{1}{s(x_i)} \right| \leq \sqrt{\frac{m}{S}}. \tag{18}$$

Because of the similarity, we only prove the first one. To this end, using Jensen's inequality for the concave function $t \mapsto \sqrt{t}$,

we obtain

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left\|\sum_{i=1}^m \sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H \\
&\qquad\le \sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left\|\sum_{i=1}^m \sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H^2} \\
&\qquad= \sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left\langle\sum_{i=1}^m \sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)},\sum_{i=1}^m \sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\rangle_H} \\
&\qquad= \sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left[\sum_{i=1}^m \sigma_i^2\frac{\mathsf{K}(x_i,x_i)}{s(x_i)^2}+\sum_{i\ne j}\sigma_i\sigma_j\frac{\mathsf{K}(x_i,x_j)}{s(x_i)s(x_j)}\right]} \\
&\qquad= \sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\left[\sum_{i=1}^m \frac{\mathsf{K}(x_i,x_i)}{s^2(x_i)}\right]} \\
\text{(since } q(x)=\frac{s(x)}{S}p(x)\text{)}\quad &= \frac{1}{\sqrt{S}}\sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim p}\left[\sum_{i=1}^m \frac{\mathsf{K}(x_i,x_i)}{s(x_i)}\right]} \\
\text{(since } s(x)\ge 1\text{)}\quad &\le \frac{1}{\sqrt{S}}\sqrt{\mathop{\mathbb{E}}_{x_{1:m}\sim p}\left[\sum_{i=1}^m \mathsf{K}(x_i,x_i)\right]} \\
&= \frac{1}{\sqrt{S}}\sqrt{m\mathop{\mathbb{E}}_{x\sim p}\mathsf{K}(x,x)}\le\sqrt{\frac{m}{S}}E_1.
\end{aligned}$$

For simplicity set, $f_w(x)=\ell^H_{\phi,k}(w,x)$ and $\alpha(w)=\frac{S}{\int_{\mathcal{X}} f_w(x)\mathrm{d}P(x)}$. During the proof, we need some good upper bounds for $\alpha(w)\|w\|_H^2$ and $\alpha(w)$. Note that $\int_{\mathcal{X}} f_w(x)\mathrm{d}P(x)=\frac{1}{k}\|w\|_H^2+\int_{\mathcal{X}}\phi(\mathsf{K}(w,x))\mathrm{d}P(x)\ge\frac{1}{k}\|w\|_H^2$ which implies $\alpha(w)\|w\|_H^2\le Sk$. Also,

$$\begin{aligned}
\int_{\mathcal{X}} f_w(x)\mathrm{d}P(x)&=\frac{1}{k}\|w\|_H^2+\int_{\mathcal{X}}\phi(\mathsf{K}(w,x))\mathrm{d}P(x)\\
&\ge\frac{1}{k}\|w\|_H^2+\int_{\mathcal{X}}\phi(\|w\|_H\|x\|_H)\mathrm{d}P(x)\qquad(\phi\text{ is non-increasing \& Cauchy–Schwarz ineq.})\\
&\ge\begin{cases}\frac{1}{4E_1^2k} & \|w\|_H\ge\frac{1}{2E_1}\\ \int_{\mathcal{X}}\phi(\frac{\|x\|_H}{2E_1})\mathrm{d}P(x) & \|w\|_H\le\frac{1}{2E_1}\end{cases}\\
&\ge\begin{cases}\frac{1}{4E_1k} & \|w\|_H\ge\frac{1}{2E_1}\\ E_2 & \|w\|_H\le\frac{1}{2E_1}\end{cases}\\
&=\min\left(\frac{1}{4E_1k},E_2\right).
\end{aligned}$$

This concludes $\alpha(w)\le\max\left(4SE_1k,\frac{S}{E_2}\right)$ for each w. Note that if $2^n\le\|w\|_H\le 2^{n+1}$, then

$$\int_{\mathcal{X}} f_w(x)\mathrm{d}P(x)=\frac{1}{k}\|w\|_H^2+\int_{\mathcal{X}}\phi(\mathsf{K}(w,x))\mathrm{d}P(x)\ge\frac{2^{2n}}{k}$$

which implies $\alpha(w)\le\frac{kS}{2^{2n}}$. So, we obtained

$$\alpha(w)\le\begin{cases}\max\left(4SE_1k,\frac{S}{E_2}\right) & \|w\|_2\le 1\\ \frac{kS}{2^{2n}} & 2^n\le\|w\|_H\le 2^{n+1}.\end{cases}\tag{19}$$

To bound the supremum within the Radamacher complexity, we utilize these inequalities to partition the parameter space. It is necessary as we want to use Lemma C.4 in which we need a bounded property. Consequently,

$$\begin{aligned}
\mathcal{R}_m^q(\mathcal{T}_{\mathcal{L}_{\phi,k}^H}) &= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{f\in\mathcal{L}_{\phi,k}^H}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i T_f(x_i)\right] \\
&= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\underbrace{\frac{S}{\int_{\mathcal{X}} f_w(x)\mathrm{d}\mu}}_{=\alpha(w)}\left[\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{f_w(x_i)}{s(x_i)}\right] \\
&= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\alpha(w)\left[\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\phi(\mathsf{K}(w,x_i))+\frac{1}{k}\|w\|_H^2}{s(x_i)}\right] \\
&\le \underbrace{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\phi(\mathsf{K}(w,x_i))}{s(x_i)}\right|}_{=M} \\
&\qquad + \frac{1}{k}\underbrace{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\|w\|_H^2}{s(x_i)}\right|}_{=N} \\
(*) \quad &\le \left[2LE_1\max\left(4E_1k,\frac{1}{E_2}\right)+8LkE_1+\max\left(4E_1k,\frac{1}{E_2}\right)\phi(0)+1\right]\sqrt{\frac{S}{m}}.
\end{aligned}$$

To prove $(*)$, we deal with M and N separately. Note that $\bar{\phi}(\cdot)=\phi(\cdot)-\phi(0)$ is an L-Lipschitz functions with $\bar{\phi}(0)=0$ and thus satisfies the condition of Corollary C.4. Note that

$$\begin{aligned}
M \le& \underbrace{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right|}_{=M_1} \\
&+ \underbrace{\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{w\in\mathcal{X}}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\phi(0)}{s(x_i)}\right|}_{=M_2} \\
\le& M_1+\phi(0)\max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left|\frac{1}{m}\sum_{i=1}^m \frac{\sigma_i}{s(x_i)}\right| \\
\le& M_1+\sqrt{\frac{S}{m}}\max\left(4E_1k,\frac{1}{E_2}\right)\phi(0) \qquad \text{by (18).}
\end{aligned}$$

To upper bound M_1, define

$$M_1^0 = \mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{\|w\|_H\le 1}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right|$$

and, for $n\in\mathbb{N}$,

$$M_1^n = \mathop{\mathbb{E}}_{x_{1:m}\sim q}\mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\sup_{2^{n-1}\le\|w\|\le 2^n}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^m \sigma_i \frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right|$$

and note that $M_1 \leq \sum_{n=0}^{\infty} M_1^n$. We can write

$$
\begin{aligned}
M_1^0 &= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{\|w\|_H\leq 1} \alpha(w)\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right| \\
\text{(since } \alpha(w)\leq \max(4SE_1k,\frac{S}{E_2})\text{)} \quad &\leq \max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{\|w\|_H\leq 1}\frac{1}{m}\left|\sum_{i=1}^{m}\sigma_i\frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right| \\
\text{(Cor. C.4 with } \varepsilon_i=\frac{1}{s(x_i)}\text{)} \quad &\leq 2L\max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{\|w\|_H\leq 1}\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(w,x_i)}{s(x_i)}\right| \\
&= \frac{2L}{m}\max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{\|w\|_H\leq 1}\left|\left\langle \mathsf{K}(w,\cdot),\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\rangle_H\right| \\
\text{(Cauchy–Schwarz inequality)} \quad &\leq \frac{2L}{m}\max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{\|w\|_H\leq 1}\|w\|_H\left\|\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H \\
&\leq \frac{2L}{m}\max\left(4SE_1k,\frac{S}{E_2}\right)\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left\|\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H \\
\text{(by (18))} \quad &\leq 2LE_1\max\left(4E_1k,\frac{1}{E_2}\right)\sqrt{\frac{S}{m}}.
\end{aligned}
$$

and similarly,

$$
\begin{aligned}
M_1^n &= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{2^{n-1}\leq\|w\|_H\leq 2^n}\alpha(w)\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right| \\
\text{(Since } \alpha(w)\leq\frac{kS}{2^{2(n-1)}}\text{)} \quad &\leq \frac{kS}{m4^{n-1}}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{2^{n-1}\leq\|w\|_H\leq 2^n}\left|\sum_{i=1}^{m}\sigma_i\frac{\bar{\phi}(\mathsf{K}(w,x_i))}{s(x_i)}\right| \\
\text{(Cor. C.4 with } \varepsilon_i=\frac{1}{s(x_i)}\text{)} \quad &\leq \frac{2LkS}{m4^{n-1}}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{2^{n-1}\leq\|w\|_H\leq 2^n}\left|\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(w,x_i)}{s(x_i)}\right| \\
&= \frac{2LkS}{m4^{n-1}}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{2^{n-1}\leq\|w\|_H\leq 2^n}\left|\left\langle \mathsf{K}(w,\cdot),\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\rangle_H\right| \\
&\leq \frac{2LkS}{m4^{n-1}}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\ \sup_{2^{n-1}\leq\|w\|_H\leq 2^n}\|w\|_H\left\|\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H \\
&\leq \frac{LkS}{m2^{n-3}}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}}\left\|\sum_{i=1}^{m}\sigma_i\frac{\mathsf{K}(x_i,\cdot)}{s(x_i)}\right\|_H \\
\text{(by (18))} \quad &\leq \frac{LkE_1}{2^{n-3}}\sqrt{\frac{S}{m}}.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
M_1 &\leq M_1^0+\sum_{n=1}^{\infty}M_1^n \\
&\leq 2LE_1\max\left(2E_1k,\frac{1}{E_2}\right)\sqrt{\frac{S}{m}}+\sum_{n=1}^{\infty}\frac{LkE_1}{2^{n-3}}\sqrt{\frac{S}{m}} \\
&= 2LE_1\max\left(2E_1k,\frac{1}{E_2}\right)\sqrt{\frac{S}{m}}+8LkE_1\sqrt{\frac{S}{m}}
\end{aligned}
$$

and thus

$$M \leq \left[2LE_1 \max\left(4E_1k, \frac{1}{E_2}\right) + 8LkE_1 + \max\left(4E_1k, \frac{1}{E_2}\right)\phi(0)\right]\sqrt{\frac{S}{m}}.$$

To complete the proof, we need to upper bound N which is done as follows

$$\begin{aligned}
N &= \mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}} \sup_{w\in\mathcal{X}} \alpha(w)\left|\frac{1}{m}\sum_{i=1}^{m}\sigma_i\frac{\|w\|_H^2}{s(x_i)}\right| \\
&= \frac{1}{m}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:m}\sim\{-1,1\}} \sup_{w\in\mathbb{R}^d} \alpha(w)\|w\|_H^2\left[\sum_{i=1}^{m}\sigma_i\frac{1}{s(x_i)}\right] \\
\text{(since } \alpha(w)\|w\|^2 \leq Sk\text{)} \qquad &\leq \frac{Sk}{m}\mathop{\mathbb{E}}_{x_{1:m}\sim q}\ \mathop{\mathbb{E}}_{\sigma_{1:k}}\left|\sum_{i=1}^{m}\frac{\sigma_i}{s(x_i)}\right| \\
\text{(by (18))} \qquad &\leq \sqrt{\frac{S}{m}}k,
\end{aligned}$$

which completes the proof. □

We are now in a position to state the following conclusion.

Theorem C.6. [Theorem A.3, Restated] *For a well behaved probability measure P (see Definition A.1), if $s(\cdot) : \mathcal{X} \longrightarrow (1,\infty)$ is an upper sensitivity function for $\mathcal{L}^H_{\phi,k}$ with total sensitivity S and $m \geq \frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right)$, then, with probability at least $1-\delta$, any s-sensitivity sample $x_1,\ldots,x_m$ from $\mathcal{X}$ with weights $u_i = \frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathcal{X},\ P,\ \mathcal{X},\ \ell^H_{\phi,k})$, where*

$$C = (2LE_1 + \phi(0))\max(4E_1k, \frac{1}{E_2}) + 8LkE_1 + 1.$$

The statement remains true if we replace $\mathcal{L}^H_{\phi,k}$ and $(\mathcal{X},\ P,\ \mathcal{X},\ \ell^H_{\phi,k})$ by $\bar{\mathcal{L}}^H_{\phi,k}$ and $(\mathcal{X},\ P,\ \mathcal{X},\ \bar{\ell}^H_{\phi,k})$.

Proof. The proof immediately follows from Theorem C.2 and Lemma C.5. □

When E_1 is dimension independent, this theorem provides a no-dimensional ε-coreset for $(\mathbb{R}^d,\ P,\ \mathbb{R}^d,\ \ell_{\phi,k})$, that is, whose size is independent of the dimension of the space. In the following subsection, we present some applications of this theorem. This theorem holds significant importance in deriving the dimension-free results outlined in Table 3.

D. Coresets for Monotonic Functions

Let H be a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ and P a probability measure over $\mathcal{X}$. For a non-increasing L-Lipschitz function $\phi\colon \mathbb{R} \longrightarrow (0,\infty)$, we remind that $\ell^H_{\phi,k} : \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ where $\ell^H_{\phi,k}(x,w) = \phi(\mathsf{K}(x,w)) + \frac{1}{k}\|w\|^2_H$. An ε-coreset for $(\mathcal{X}, P, \mathcal{L}^H_{\phi,k})$, called *$\varepsilon$-coreset for a monotonic function ϕ with respect to H*, is a set $X = \{x_1,\ldots,x_m\} \subseteq \mathcal{X}$ accompanied by a measure (or weight function) ν such that

$$\left|\int_{x\in\mathcal{X}} \ell^H_{\phi,k}(x,w)\mathrm{d}P(x) - \sum_{i=1}^{m}\nu(x_i)\ell^H(x_i,w)\right| \leq \epsilon\int_{x\in\mathcal{X}}\ell^H_{\phi,k}(x,w)\mathrm{d}P(x) \quad \forall w\in\mathcal{X}. \tag{20}$$

We start with a straightforward observation.

Lemma D.1. *If $\phi : \mathbb{R} \longrightarrow (0,\infty)$ is a non-increasing function and $0 < K_1 \leq K_2$, then*

$$\frac{\phi(-t) + \frac{t^2}{K_1}}{\phi(t) + \frac{t^2}{K_1}} \leq \frac{\phi(-t) + \frac{t^2}{K_2}}{\phi(t) + \frac{t^2}{K_2}} \qquad \textit{for each } t \geq 0.$$

Proof. For $t \geq 0$, we should check the inequality

$$\left(\phi(-t) + \frac{t^2}{K_1}\right)\left(\phi(t) + \frac{t^2}{K_2}\right) \leq \left(\phi(-t) + \frac{t^2}{K_2}\right)\left(\phi(t) + \frac{t^2}{K_1}\right)$$

which is equivalent to

$$\phi(-t)\frac{t^2}{K_2} + \phi(t)\frac{t^2}{K_1} \leq \phi(-t)\frac{t^2}{K_1} + \phi(t)\frac{t^2}{K_2}$$

which is valid since

$$\phi(-t)t^2\left(\frac{1}{K_1} - \frac{1}{K_2}\right) + \phi(t)t^2\left(\frac{1}{K_2} - \frac{1}{K_1}\right) = t^2\left(\frac{1}{K_1} - \frac{1}{K_2}\right)(\phi(-t) - \phi(t)) \geq 0,$$

completing the proof. □

A version of the lemma described below appears in (Tolochinksy et al., 2022). We carefully revisit the ideas in their proof in order to extend and enhance the lemma.

Lemma D.2. *Let $\phi : \mathbb{R} \longrightarrow (0, \infty)$ be a non-increasing function such that*

$$\frac{\phi(-\alpha z) + \frac{z^2}{k}}{\phi(\alpha z) + \frac{z^2}{k}} \leq \beta(\alpha) \qquad 0 \leq \alpha \leq B_1, 0 \leq z \leq B_2. \tag{21}$$

If we set $M = \phi(-B_1B_2)$, then, for each $x, y, w \in \mathcal{X}$ with $\|x\|_H, \|y\|_H \leq B_1, \|w\|_H \leq B_2$, we have

$$\frac{\phi(0)}{M\beta(\|x\|_H)} \leq \frac{\ell^H_{\phi,k}(x,w)}{\ell^H_{\phi,k}(y,w)} \qquad \textit{and} \qquad \frac{\phi(0)}{M\beta(\|x\|_H)} \leq \frac{\bar{\ell}^H_{\phi,k}(x,w)}{\bar{\ell}^H_{\phi,k}(y,w)}.$$

If ϕ is universally bounded by M, then we do not need upper bounds B_1, B_2 for α, z and consequentially do not need upper bounds for $\|x\|_H, \|y\|_H$, and $\|w\|_H$, and the same statement holds.

Proof. Due to similarity, we only prove the lemma for $\ell^H_{\phi,k}$. First, note that as ϕ is non-increasing, $\beta(\alpha) \geq 1$ for each α. Consider arbitrary $x, y, w \in \mathcal{X}$ with $\|x\|_H, \|y\|_H \leq B_1, \|w\|_H \leq B_2$. In the following, we deal with the two different cases $\mathsf{K}(x,w) \leq 0$ and $\mathsf{K}(x,w) > 0$ separately. If $\mathsf{K}(x,w) \leq 0$, then

$$\begin{aligned}
\phi(\mathsf{K}(y,w)) &= \phi(\langle y, w\rangle_H) \\
&\leq \phi(-\|y\|_H\|w\|_H) && (\phi \text{ is non-increasing along with Cauchy–Schwarz inequality}) \\
&\leq \phi(-B_1B_2) \\
&= M = \frac{M}{\phi(0)}\phi(0) \\
&\leq \frac{M}{\phi(0)}\phi(\mathsf{K}(x,w)) && (\text{since } \mathsf{K}(x,w) \leq 0 \text{ and } \phi \text{ is non-increasing})
\end{aligned}$$

and hence, as $\frac{M}{\phi(0)}, \beta(\|x\|_H) \geq 1$, by adding $\frac{\|w\|_H^2}{k}$ to both sides, we obtain

$$\ell_{\phi,k}(y,w) \leq \frac{M}{\phi(0)}\ell^H_{\phi,k}(x,w) \leq \frac{M}{\phi(0)}\beta(\|x\|_H)\ell^H_{\phi,k}(x,w). \tag{22}$$

Now, assume that $\mathsf{K}(x,w) > 0$. Similarly,

$$\phi(\mathsf{K}(y,w)) \leq M \leq \frac{M}{\phi(0)}\phi(-\mathsf{K}(x,w)) \leq \frac{M}{\phi(0)}\phi(-\|x\|_H\|w\|_H).$$

Again, by adding $\frac{\|w\|_H^2}{k}$ to the both sides, we have

$$\begin{aligned}
\ell_{\phi,k}^H(y,w) = \phi(\mathsf{K}(y,w)) + \frac{\|w\|_H^2}{k} &\leq \frac{M}{\phi(0)}\left(\phi(-\|x\|_H\|w\|_H) + \frac{\|w\|_H^2}{k}\right) \\
\text{(Property 21)} \quad &\leq \frac{M}{\phi(0)}\beta(\|x\|_H)\left(\phi(\|x\|_H\|w\|_H) + \frac{\|w\|_H^2}{k}\right) \\
\text{(Cauchy-Schwarts Inequality and } \phi \text{ is non-increasing)} \quad &\leq \frac{M}{\phi(0)}\beta(\|x\|_H)\underbrace{\left(\phi(\mathsf{K}(x,w)) + \frac{\|w\|_H^2}{k}\right)}_{=\ell_{\phi,k}^H(x,w)},
\end{aligned}$$

which implies

$$\ell_{\phi,k}^H(y,w) \leq \frac{M}{\phi(0)}\beta(\|x\|_H)\ell_{\phi,k}^H(x,w). \tag{23}$$

Combining Equations (22) and (23), we obtain the desired inequality. □

This lemma provides us with a function $\gamma(\alpha) = \frac{\phi(0)}{M\beta(\alpha)} \leq 1$ ensuring $\gamma(\|x\|_H) \leq \frac{\ell_{\phi,k}^H(x,w)}{\ell_{\phi,k}^H(y,w)}$. The subsequent lemma will highlight the usefulness of this function.

Lemma D.3. *Let P be a probability measure over $\mathcal{X}$. Assume that $\mathcal{W} \subseteq \mathcal{X}$, $\ell : \mathcal{X} \times \mathcal{X} \longrightarrow (0,\infty)$, and $\gamma : [0,\infty) \longrightarrow [0,\infty)$ such that $0 < \int_{\mathcal{X}} \gamma(\|x\|_H)\mathrm{d}P(x) < \infty$ and*

$$\gamma(\|x\|_H) \leq \frac{\ell(x,w)}{\ell(y,w)} \qquad \forall x,y \in \mathcal{X}, w \in \mathcal{W}. \tag{24}$$

Then

$$s(y) = \frac{1}{\int_{\mathcal{X}} \gamma(\|x\|_H)\mathrm{d}P(x)}$$

is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}_W)$, where $\mathcal{L}_{\mathcal{W}} = \{l(\cdot,w) \colon w \in \mathcal{W}\}$.

Proof. For a $y \in \mathcal{X}$, we have

$$\begin{aligned}
\sup_{w\in\mathcal{W}} \frac{\ell(y,w)}{\int_{\mathcal{X}} \ell(x,w)\mathrm{d}P(x)} &= \sup_{w\in\mathcal{W}} \frac{1}{\int_X \frac{\ell(x,w)}{\ell(y,w)}\mathrm{d}P(x)} \\
&\leq \sup_{w\in\mathcal{W}} \frac{1}{\int_X \gamma(\|x\|_H)\mathrm{d}P(x)} \\
&= \frac{1}{\int_X \gamma(\|x\|_H)\mathrm{d}P(x)}
\end{aligned}$$

completing the proof. □

It is worth noting that the provided upper sensitivity function $s(y) = \frac{1}{\int_{\mathcal{X}} \gamma(\|x\|)\mathrm{d}P}$ remains constant, irrespective of $y \in \mathcal{X}$. In a specific scenario where γ is lower bounded by L, the function $s(y) = \frac{1}{L}$ serves as an upper sensitivity function, possessing a total value of $S = \frac{1}{L}$. It is important to note that when ϕ and k grantee Property (24) for $\ell = \ell_{\phi,k}^H$ (like what is established in Lemma D.2), then Lemma D.3 provides a methodology to calculate the upper sensitivity function and its total value for $\mathcal{L}_{\phi,k}^H$. This enables us to obtain s-sensitivity samples from $\mathcal{X}$, making Theorem C.6 (and Theorem 1.2) applicable in this context.

D.1. Coreset for Sigmoid function

Tolochinksy et al. (2022) examined Property (24) for specific functions ϕ. They demonstrated for the sigmoid function $\sigma(x) = \frac{1}{1+e^x}$ that, given $\alpha > 0$, there exists a threshold k_α such that for any $k \geq k_\alpha$

$$\frac{\sigma(-\alpha z) + \frac{z^2}{k}}{\sigma(\alpha z) + \frac{z^2}{k}} \leq 66k\alpha \qquad \forall z \geq 0. \tag{25}$$

Using this result, they established the first item of Theorem 1.5 in which, as a drawback, k should be sufficiently large contingent on the given data points $\mathcal{X}$. This dependency comes from the role of k_α in Equation (25). In the subsequent lemma, we eliminate this dependency.

Lemma D.4. *For $\sigma(x) = \frac{1}{1+e^x}$ and fixed $k > 0$, we have*

$$\frac{\sigma(-\alpha z) + \frac{z^2}{k}}{\sigma(\alpha z) + \frac{z^2}{k}} \leq 4\left(1 + \max(e, \alpha^2 k)\right) \qquad \forall \alpha, z \geq 0.$$

Proof. One can verify that $\sigma : \mathbb{R} \longrightarrow (0,1)$ is a positive decreasing function and $\sigma'(t) = -\sigma(t)\sigma(-t)$. For a fixed $\alpha > 0$, if we set $t = \alpha z$ and $K = \alpha^2 k$,

$$\frac{\sigma(-\alpha z) + \frac{z^2}{k}}{\sigma(\alpha z) + \frac{z^2}{k}} = \frac{\sigma(-t) + \frac{t^2}{K}}{\sigma(t) + \frac{t^2}{K}} \leq \frac{1 + \frac{t^2}{K}}{\sigma(t) + \frac{t^2}{K}} = f(t).$$

To find the maximum of $f(t)$ on $(0, \infty)$ in terms of K, we begin by calculating $f'(t)$, the derivative of f, after factoring out $1/(\sigma(t) + \frac{t^2}{K})^2$, as follows:

$$\begin{aligned}
\left(\sigma(t) + \frac{t^2}{K}\right)^2 f'(t) =& \frac{2t}{K}\left(\sigma(t) + \frac{t^2}{K}\right) - \left(\sigma'(t) + \frac{2t}{K}\right)\left(1 + \frac{t^2}{K}\right) \\
=& \sigma(t)\frac{2t}{K} - \sigma'(t) - \frac{2t}{K} - \sigma'(t)\frac{t^2}{K} \\
=& \sigma(t)\frac{2t}{K} + \sigma(t)\sigma(-t) - \frac{2t}{K} + \sigma(t)\sigma(-t)\frac{t^2}{K} \\
=& \frac{2t}{K}\left(\sigma(t) - 1\right) + \sigma(t)\sigma(-t)\left(1 + \frac{t^2}{K}\right) \\
\text{(since } \sigma(t) + \sigma(-t) = 1\text{)} \quad =& -\sigma(-t)\frac{2t}{K} + \sigma(t)\sigma(-t)\left(1 + \frac{t^2}{K}\right) \\
=& \sigma(-t)\left[-\frac{2t}{K} + \sigma(t)\left(1 + \frac{t^2}{K}\right)\right].
\end{aligned}$$

So, as $(\sigma(t) + t^2/K)^2 > 0$, whenever $\sigma(t)\left(1 + \frac{t^2}{K}\right) < \frac{2t}{K}$ or equivalently $\frac{K}{2t} + \frac{t}{2} < 1 + e^t$, f' is negative. For

$$h(t) = 1 + e^t - \frac{t}{2} - \frac{K}{2t},$$

we have

$$h'(t) = e^t - \frac{1}{2} + \frac{K}{2t^2} > 0 \qquad \text{for each } t \geq 0.$$

This indicates that h is increasing, implying it can have at most one zero. Since h is negative for small t and positive for large t, it follows that h possesses a unique zero, denoted as t_0. This implies that f is decreasing for $t \geq t_0$ and increasing for $t \leq t_0$, thus attaining its maximum at t_0. In the following discussion, we distinguish between two cases: $K \geq e$ and $K \leq e$. Initially, we focus on bounding $\sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K}}{\sigma(t)+\frac{t^2}{K}}$ for the scenario where $K \geq e$. Subsequently, we employ this bound, along with a straightforward observation, to derive a bound for $\sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K}}{\sigma(t)+\frac{t^2}{K}}$ when $K \leq e$.

1. **Case 1: $K \geq e$.** We start with $t \geq \log K \geq 1$. As h is increasing

$$h(t) \geq h(\log K) = 1 + K - \frac{\log K}{2} - \frac{K}{2\log K} \geq 1.$$

This means that h is positive for $t \geq \log K$. Now, assume that $t \leq 0.5 \log K$. Again, because h is increasing, we have

$$h(t) \leq h(0.5 \log K) \leq 1 + \sqrt{K} - \frac{\log K}{4} - \frac{K}{\log K} < 0.$$

These observations together yield $t_0 \in [0.5 \log K,\ \log K]$. Consequently,

$$\begin{aligned}\sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K}}{\sigma(t)+\frac{t^2}{K}} &= \frac{1+\frac{t_0^2}{K}}{\sigma(t_0)+\frac{t_0^2}{K}} \\ &\le \frac{1+\frac{\log^2 K}{K}}{\sigma(\log K)+\frac{\log^2 K}{4K}} \\ &\le \frac{1+\frac{\log^2 K}{K}}{\frac{\log^2 K}{4K}} \\ &\le 4\left(1+\frac{K}{\log^2 K}\right) \le 4(1+K).\end{aligned}$$

2. **Case 2: $K \le e$.** Calculation shows that if $0 < K_1 \le K_2$, then

$$\frac{1+\frac{t^2}{K_1}}{\sigma(t)+\frac{t^2}{K_1}} \le \frac{1+\frac{t^2}{K_2}}{\sigma(t)+\frac{t^2}{K_2}} \qquad \forall t \ge 0$$

which concludes

$$\sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K_1}}{\sigma(t)+\frac{t^2}{K_1}} \le \sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K_2}}{\sigma(t)+\frac{t^2}{K_2}}.$$

Consequently, for $K \le e$,

$$\sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{K}}{\sigma(t)+\frac{t^2}{K}} \le \sup_{t\in(0,\infty)} \frac{1+\frac{t^2}{e}}{\sigma(t)+\frac{t^2}{e}} \le 4(1+e),$$

where the right most inequality comes from Case (1).

Putting Cases (1) and (2) together, we obtain

$$\max_{t\in[0,t_0]} f(t) \le \begin{cases} 4(1+K) & K \ge e \\ 4(1+e) & K < e. \end{cases}$$

So, replacing K with $\alpha^2 k$, we obtain

$$\frac{\sigma(-\alpha z)+\frac{z^2}{k}}{\sigma(\alpha z)+\frac{z^2}{k}} \le \beta(\alpha) = \begin{cases} 4(1+\alpha^2 k) & \alpha^2 k \ge e \\ 4(1+e) & \alpha^2 k \le e \end{cases} \qquad \forall \alpha, z > 0.$$

Note that for $\alpha = 0$ or $z = 0$, the inequality is valid as well. □

Lemma D.5. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim P} \|x\|_H^2 \le E_1^2$. Then $s(x) = 60 + 32kE_1^2$ is an upper sensitivity function for $(\mathcal{X},\ P,\ \ell_{\sigma,k}^H)$ and $(\mathcal{X},\ P,\ \bar{\ell}_{\sigma,k}^H)$.*

Proof. Given $\mathbb{E}_{x\sim P}(\|x\|_H^2) \le E_1^2$, Markov's inequality implies that $\mathbb{P}_{x\sim P}(\|x\|_H \ge 2E_1^2) \le \frac{1}{2}$ and consequently $\mathbb{P}_{x\sim P}(\|x\|_H \le 2E_1^2) \ge \frac{1}{2}$ which will be used later to derive inequality marked by in $(*)$. Since σ is universally bounded by 1. Lemma D.4 indicates that Lemma D.2 is applicable with

$$\beta_\sigma(\alpha) = \begin{cases} 4(1+\alpha^2 k) & \alpha^2 k \ge e \\ 4(1+e) & \alpha^2 k \le e \end{cases}$$

and for all $x, y, w \in \mathcal{X}$ and $M = 1$. Using Lemma D.3 with $\gamma(\|x\|_H) = \frac{\sigma(0)}{M\beta_\sigma(\|x\|_H)} = \frac{1}{2\beta_\sigma(\|x\|_H)}$, we conclude that $s(y) = \frac{2}{\int_{\mathbb{R}^d} \frac{1}{\beta_\sigma(\|x\|_H)} \mathrm{d}P}$ is an upper sensitivity function for $(\mathcal{X},\ P,\ \mathcal{X},\ \ell^H_{\sigma,k})$ and $(\mathcal{X},\ P,\ \mathcal{X},\ \bar{\ell}^H_{\sigma,k})$. Note that

$$
\begin{aligned}
\int_{\mathcal{X}} \frac{1}{\beta_\sigma(\|x\|_H)} \mathrm{d}P(x) &\geq \int_{\{x : e \leq k\|x\|_H^2 \leq 2kE_1^2\}} \frac{1}{4(1+\|x\|_H^2 k)} \mathrm{d}P(x) + \int_{\{x : k\|x\|_H^2 < e\}} \frac{1}{4(1+e)} \mathrm{d}P(x) \\
&\geq \frac{1}{4(1+2kE_1^2)} P\left(\left\{x : \frac{e}{k} \leq \|x\|_H^2 \leq 2E_1^2\right\}\right) + \frac{1}{4(1+e)} P\left(\left\{x : \|x\|_H^2 < \frac{e}{k}\right\}\right) \\
&\geq \mathbb{P}\left(\left\{x : \|x\|_H^2 \leq 2E_1^2\right\}\right) \min\left(\frac{1}{4(1+2kE_1^2)}, \frac{1}{4(1+e)}\right) \\
(*) \qquad &\geq \frac{1}{2} \min\left(\frac{1}{4(1+2kE_1^2)}, \frac{1}{4(1+e)}\right) \\
&\geq \frac{1}{8(1+e+2kE_1^2)} \geq \frac{1}{30+16kE_1^2}.
\end{aligned}
$$

This concludes $s(x) = 60 + 32kE_1^2$ is an upper sensitivity function for $(\mathcal{X}, P, \ell^H_{\sigma,k})$ and $(\mathcal{X}, P, \bar{\ell}^H_{\sigma,k})$. □

Now, we are at a point to put Theorem A.3 into action.

Theorem D.6 (No Dimensional Sigmoid Coreset)**.** *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$, P is a probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim p} \|x\|_H^2 \leq E_1^2$, $S = 60+32kE_1^2$, and $C = \left(2E_1 + \frac{1}{2}\right)\max\left(4E_1 k, \frac{2}{5}\right) + 8kE_1 + 1$. For $m \geq \frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$, any iid sample $x_1, \ldots, x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathcal{X}, P, \mathcal{X}, \ell^H_{\sigma,k})$ (respectively for $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\sigma,k})$) with probability at least $1-\delta$.*

Proof. Lemma D.5 indicates that $s(x) = 60 + 32kE_1^2$ is an upper sensitivity function. Given that it is a constant function, s-sensitivity sampling is equivalent to sampling according to P. Notice that σ is a 1-Lipschitz function. As $\log(\cdot)$ is a concave function, using Jensen's inequality, we have $\mathbb{E}\left(\log\sigma\left(\frac{\|x_i\|_H}{2E_1}\right)\right) \leq \log\mathbb{E}\left(\sigma\left(\frac{\|x_i\|_H}{2E_1}\right)\right)$. On the other hand, since $-\log\sigma(t) = \log(1+e^t) \leq 1+t$, we have

$$
\mathbb{E}\left(\log\sigma\left(\frac{\|x_i\|_H}{2E_1}\right)\right) \geq -\mathbb{E}\left(1 + \frac{\|x_i\|_H}{2E_1}\right) = -1 - \frac{\mathbb{E}\|x_i\|_H}{2E_1} \geq -\frac{3}{2},
$$

which concludes

$$
\mathbb{E}\left(\sigma\left(\frac{\|x_i\|_H}{2E_1}\right)\right) \geq e^{\mathbb{E}\left(\log\sigma\left(\frac{\|x_i\|_H}{2E_1}\right)\right)} \geq e^{-\frac{3}{2}} \geq 0.4 = E_2.
$$

Hence, P is well behaved (see Definition A.1), thereby allowing Theorem A.3 to establish the statement. □

While Theorem D.6 holds for any reproducing kernel Hilbert space meeting the specified criteria, the most intriguing instance arises in Hilbert space $\mathbb{R}^d$ equipped with the standard inner product (also known as the dot product) as its kernel. In this scenario, we can also employ Theorem 1.2, which leads to the subsequent theorem.

Theorem D.7. *Let P be a probability measure over $\mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{x\sim p} \|x\|_2^2 \leq E_1$ and $S = 60 + 32kE_1^2$. There is an $m = O\left(\frac{S}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)\right)$ such that any iid sample $x_1, \ldots, x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\sigma,k})$ (respectively for $(\mathbb{R}^d, P, \mathbb{R}^d, \bar{\ell}_{\sigma,k})$) with probability at least $1-\delta$.*

Proof. Because of similarity, we only prove the statement for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\sigma,k})$. Given that Lemma D.5 asserts the constant function $s(x) = 60+32kE_1^2$ as an upper sensitivity function for $\mathcal{L}_{\sigma,k}$, we deduce $\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\sigma,k}}, \succ) = \mathbf{Ranges}(\mathcal{L}_{\sigma,k}, \succ)$. To clarify, every function in $\mathcal{T}_{\mathcal{L}_{\sigma,k}}$ is a function in $\mathcal{L}_{\sigma,k}$ scaled by a positive constant. This implies $\mathrm{VCdim}(\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\sigma,k}}, \succ)) = \mathrm{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\sigma,k}, \succ))$. We would like to remind that

$$
\mathcal{T}_{\mathcal{L}_{\sigma,k}} = \left\{ f_w(\cdot) = \frac{\sigma(\langle w, \cdot\rangle) + \frac{1}{k}\|w\|_2^2}{s(\cdot)} \colon w \in \mathbb{R}^d \right\}.
$$

For an $f_w \in \mathcal{L}_{\sigma,k}$ and $r \geq 0$,

$$\begin{aligned}
\mathbf{range}(f_w, \succ, r) &= \left\{x \in \mathbb{R}^d \colon f_w(x) > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \sigma(\langle x, w\rangle) + \frac{\|w\|^2}{k} > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \frac{1}{1+e^{\langle x,w\rangle}} > \underbrace{r - \frac{\|w\|^2}{k}}_{=t}\right\} \\
&= \begin{cases} \mathbb{R}^d & t \leq 0 \\ \varnothing & t \geq 1 \\ \left\{x \in \mathbb{R}^d \colon \langle x, w\rangle < \log(\frac{1}{t} - 1)\right\} & 0 < t < 1, \end{cases}
\end{aligned}$$

which concludes that $\mathbf{Ranges}(\mathcal{L}_{\sigma,k}, \succ)$ only includes half-spaces, empty set, and the whole space $\mathbb{R}^d$. Therefore, by Radon's theorem, $\mathrm{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\sigma,k}, \succ)) \leq d+1$ (for a proof, see Lemma 10.3.1 in (Matousek, 2002)). Leveraging Theorem 1.2, we derive the statement for $m = O(\frac{S}{\varepsilon^2}\left(d \log S + \log \frac{1}{\delta}\right))$, thereby completing the proof. □

Theorems D.6 and D.7 can be considered as an advancement of the first item of Theorem 1.5. Theorem 1.5 is applicable when $\mathcal{X}$ is a finite set inside the unit ball, and m is function that is linear in terms of $\log(|\mathcal{X}|)$ and quadratic in terms d, whereas in Theorem D.6, $\mathcal{X}$ can be infinite, and moreover, the presented m remains constant concerning those parameters (assuming E_1 is independent of d). Moreover, in Theorem D.7, m linearly depends on d, whereas in Theorem 1.5, it has a quadratic dependence on d.

D.2. Coreset for Logistic Function

In binary logistic regression, minimizing the negative log-likelihood involves working with the logistic function $\mathrm{logistic}(t) = \log(1 + e^{-t})$. Therefore, having small coresets for logistic functions holds significant value. In this section, we establish an upper bound on the size of a coreset for the logistic function.

Lemma D.8. *For* $\mathrm{logistic}(x) = \log(1 + e^{-t})$ *and fixed* $k > 0$*, we have*

$$\frac{\mathrm{logistic}(-\alpha z) + \frac{z^2}{k}}{\mathrm{logistic}(\alpha z) + \frac{z^2}{k}} \leq \begin{cases} \frac{85\alpha^2 k}{\log(\alpha^2 k)} & \alpha^2 k \geq e \\ 85 & \alpha^2 k \leq e. \end{cases} \qquad \forall \alpha, z \geq 0.$$

Proof. Simplifying notation, let $\phi(t) = \mathrm{logistic}(t)$. For a given $\alpha > 0$, if we set $t = \alpha z$ and $K = \alpha^2 k$,

$$\frac{\phi(-\alpha z) + \frac{z^2}{k}}{\phi(\alpha z) + \frac{z^2}{k}} = \frac{\phi(-t) + \frac{t^2}{K}}{\phi(t) + \frac{t^2}{K}} = f_K(t) \qquad t \geq 0.$$

Note that $\phi'(t) = -\sigma(t) = \frac{-1}{1+e^t}$. Upon computing the derivative of f_K, we obtain

$$\begin{aligned}
\left(\phi(t) + \frac{t^2}{K}\right)^2 f_K'(t) &= \left(-\phi'(-t) + \frac{2t}{K}\right)\left(\phi(t) + \frac{t^2}{K}\right) - \left(\phi'(t) + \frac{2t}{K}\right)\left(\phi(-t) + \frac{t^2}{K}\right) \\
&= -\phi'(-t)\phi(t) - \phi'(-t)\frac{t^2}{K} + \phi(t)\frac{2t}{K} - \phi'(t)\phi(-t) - \phi'(t)\frac{t^2}{K} - \phi(-t)\frac{2t}{K} \\
&= -\left(\phi'(-t)\phi(t) + \phi'(t)\phi(-t)\right) + \left(\phi(t) - \phi(-t)\right)\frac{2t}{K} - \left(\phi'(t) + \phi'(-t)\right)\frac{t^2}{K} \\
&= \left(\sigma(-t)\phi(t) + \sigma(t)\phi(-t)\right) + \left(\phi(t) - \phi(-t)\right)\frac{2t}{K} + \underbrace{\left(\sigma(t) + \sigma(-t)\right)}_{=1}\frac{t^2}{K} \\
&= \left((1-\sigma(t))\phi(t) + \sigma(t)\phi(-t)\right) + \left(\phi(t) - \phi(-t)\right)\frac{2t}{K} + \underbrace{\left(\sigma(t) + \sigma(-t)\right)}_{=1}\frac{t^2}{K} \\
&= \phi(t) + \underbrace{\left(\phi(t) - \phi(-t)\right)}_{=-t\leq 0}\left(\frac{2t}{K} - \sigma(t)\right) + \frac{t^2}{K} = h(t).
\end{aligned}$$

We investigate the two cases $K \geq e$ and $K \leq e$ separately. First, assume that $K \geq e$. In the subsequent analysis, our goal is to determine $0 < t_1 < t_2$ such that $h(t) > 0$ (equivalently f_K is increasing) for $t \in [0, t_1]$ and $h(t) < 0$ (equivalently f_K is decreasing) for $t \in [t_1, \infty)$. This concludes that f_K takes its maximum at some $t_0 \in [t_1, t_2]$. Considering the formula of $h(t)$, if $\frac{2t}{K} \leq \sigma(t)$ or equivalently $g_1(t) = 1 + e^t - \frac{K}{2t} \leq 0$, then $h(t)$ is positive. Since the function g_1 is increasing, for $t \in (0,\ 0.4\log K]$,

$$g_1(t) \leq g_1(0.4\log k) = 1 + \sqrt[5]{K^2} - \frac{5K}{4\log K} < 0 \quad \Longrightarrow \quad h(t) > 0.$$

Thus, f_K is increasing for $t \in [0,\ 0.4\log K]$. Note

$$h(t) < 0 \qquad \Longleftrightarrow \qquad \phi(t) + \frac{t^2}{K} < t\left(\frac{2t}{K} - \sigma(t)\right) \qquad \Longleftrightarrow \qquad t\sigma(t) + \phi(t) < \frac{t^2}{K}. \tag{26}$$

As $t\sigma(t) + \phi(t) < te^{-t} + e^{-t} = e^{-t}(1+t)$, having $g_2(t) = e^{-t}(1+t) - \frac{t^2}{K} < 0$ implies $h(t) < 0$. One can see that g_2 is decreasing and $g_2(2\log K) < 0$. Therefore, for $t \geq t_1 = 2\log K$, $g_2(t) < 0$ and hence $h(t) < 0$ which concludes f is decreasing for $t \geq 2\log K$. So far, we have seen that f is increasing for $t \in [0,\ 0.4\log K]$ and decreasing for $t \geq 2\log K$, which yields

$$\begin{aligned}
\sup_{t\in[0,\ \infty)} f_K(t) &= \sup_{t\in[0.4\log K,\ 2\log K]} f_K(t) \\
&= \sup_{t\in[0.4\log K,\ 2\log K]} \frac{\phi(-t) + \frac{t^2}{K}}{\phi(t) + \frac{t^2}{K}} \\
&\leq \frac{\phi(-2\log K) + \frac{(2\log K)^2}{K}}{\frac{4\log^2 K}{25K}} \\
\text{(since } \phi(-t) \leq 1 + t\text{)} \quad &\leq \frac{1 + 2\log K + \frac{4\log^2 K}{K}}{\frac{4\log^2 K}{25K}} \\
&= \frac{25K}{4\log^2 K} + \frac{25K}{2\log K} + 25 \\
&\leq 25 + \frac{75K}{4\log K}.
\end{aligned}$$

Up to this point, we have established $\sup_{t\in[0,\ \infty)} f_K(t) \leq 25 + \frac{75K}{4\log K}$ under the condition that $K \geq e$. To conclude the argument, we need to determine an upper bound for $\sup_{t\in[0,\ \infty)} f_K(t)$ when $0 < K \leq e$. Utilizing Lemma D.1, for $0 < K \leq e$, we deduce

$$\sup_{t\in[0,\ \infty)} f_K(t) \leq \sup_{t\in[0,\ \infty)} f_e(t) \leq 25 + \frac{75e}{4} \leq 85$$

which concludes

$$\sup_{t\in[0,\,\infty)} f_K(t) \leq \begin{cases} 25 + \frac{75K}{4\log K} & K \geq e \\ \\ 85 & K \leq e \end{cases} \leq \begin{cases} \frac{85\alpha^2 k}{\log(\alpha^2 k)} & \alpha^2 k \geq e \\ \\ 85 & \alpha^2 k \leq e. \end{cases}$$

□

Lemma D.9. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $P(\{x\in\mathcal{X}\colon \|x\|_H \geq A\}) = 0$. Then $s(y) = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$ is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}_{\text{logistic}})$ and $(\mathcal{X}, P, \bar{\mathcal{L}}_{\text{logistic}})$.*

Proof. As $P(\{x\in\mathcal{X}\colon \|x\|_H \geq A\}) = 0$, we only need to determine $s(y)$ for y with $\|y\|_H \leq A$. Henceforth, we assume that $\|y\|_H \leq A$. For $w\in\mathcal{W} = \{w\in\mathcal{X}\colon \|w\|_H \leq \frac{1}{A}\sqrt{\max(1,\log(A^2k))}\}$, Lemma D.8 indicates that Lemma D.2 is applicable with $B_1 = A$, $B_2 = \frac{1}{A}\sqrt{\max(1,\log(A^2k))}$, $\phi(-B_1B_2) \leq 1 + \sqrt{\max(1,\log(A^2k))} = M$, and

$$\beta_{\text{logistic}}(\alpha) = \begin{cases} \frac{85\alpha^2 k}{\log(\alpha^2 k)} & \alpha^2 k \geq e \\ \\ 85 & \alpha^2 k \geq e. \end{cases}$$

Using Lemma D.3 with $\gamma(\|x\|_H) = \frac{\text{logistic}(0)}{M\beta_{\text{logistic}}(\|x\|_H)} \geq \frac{1}{2M\beta_{\text{logistic}}(\|x\|_H)}$, we conclude that

$$\sup_{\|w\|_H\leq B_2} \frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}} \ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)} \leq \frac{2M}{\int_{x\in\mathcal{X}} \frac{1}{\beta_{\text{logistic}}(\|x\|_H)}\mathrm{d}P(x)}.$$

Note that

$$\begin{aligned} \int_{x\in\mathcal{X}} \frac{1}{\beta_{\text{logistic}}(\|x\|_H)}\mathrm{d}P(x) &\geq \int_{\{x:e\leq k\|x\|_H^2\}} \frac{\log(k\|x\|_H^2)}{85k\|x\|_H^2}\mathrm{d}P(x) + \int_{\{x:k\|x\|_H^2<e\}} \frac{1}{85}\mathrm{d}P(x) \\ &\geq \frac{\log(kA^2)}{85kA^2} P\left(\left\{x : \frac{e}{k} \leq \|x\|_H^2\right\}\right) + \frac{1}{85} P\left(\left\{x : \|x\|_H^2 < \frac{e}{k}\right\}\right) \\ &\geq \begin{cases} \frac{\log(kA^2)}{85kA^2} & A^2k \geq e \\ \\ \frac{1}{85} & A^2k < e \end{cases} \\ &\geq \frac{\max(1,\log(kA^2))}{85(1+kA^2)}. \end{aligned}$$

This concludes

$$\begin{aligned} \sup_{\|w\|_H\leq B_2} \frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}} \ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)} &\leq \frac{170(1+kA^2)}{\max(1,\log(kA^2))}\left(1 + \sqrt{\max(1,\log(kA^2))}\right) \\ &\leq \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}. \end{aligned}$$

As $\int_{x\in\mathcal{X}} \ell_{\text{logistic},k}(x,w)\mathrm{d}P \geq \frac{1}{k}\|w\|_H^2$, we have

$$\begin{aligned}
\sup_{\|w\|_H\geq B_2} \frac{\ell_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}} \ell_{\text{logistic},k}(x,w)\mathrm{d}P} &\leq \sup_{\|w\|\geq B_2} \frac{\frac{1}{k}\|w\|^2 + \log(1+e^{-\langle y,w\rangle})}{\frac{1}{k}\|w\|^2} \\
&\leq \sup_{\|w\|_H\geq B_2} \left[1 + \frac{k}{\|w\|_H^2}\log(1+e^{-\langle y,w\rangle})\right] \\
&\leq \sup_{\|w\|_H\geq B_2} \left[1 + \frac{k}{\|w\|_H^2}\log(1+e^{\|y\|_H\|w\|_H})\right] \\
\text{(since } \log(1+e^t)\leq 1+t \text{ for } t\geq 0) \qquad &\leq \sup_{\|w\|_H\geq B_2} \left[1 + \frac{k}{\|w\|_H^2}(1+\|y\|_H\|w\|_H)\right] \\
\text{(since } \|w\|_H\geq B_2 \text{ and } \|y\|_H\leq A) \qquad &\leq 1 + \frac{kA^2}{\max(1,\log(A^2k))} + \frac{kA^2}{\sqrt{\max(1,\log(A^2k))}} \\
&\leq 1 + \frac{2kA^2}{\sqrt{\max(1,\log(A^2k))}}.
\end{aligned}$$

So, $s(y) = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$ is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}^H_{\text{logistic},k})$. The proof for $(\mathcal{X}, P, \bar{\mathcal{L}}^H_{\text{logistic},k})$ proceeds in a similar manner. □

Theorem D.10 (No Dimensional Logistic Coreset)**.** *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X}\longrightarrow\mathbb{R}$, P is probability measure over $\mathcal{X}$ such that $P(\{x\in\mathcal{X}\colon \|x\|_H\geq A\}) = 0$, $S = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$, and $C = (2A+1)\max(4Ak, 2.5) + 8Ak + 1$. For $m \geq \frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$, any iid sample $x_1,\ldots,x_m\in\mathcal{X}$ according to P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathcal{X}, P, \mathcal{X}, \ell^H_{\text{logistic},k})$ (respectively for $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\text{logistic},k})$) with probability at least $1-\delta$.*

Proof. Lemma D.9 asserts that $s(y) = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$ serves as an upper sensitivity function for $(\mathcal{X},\ P,\ \mathcal{L}^H_{\text{logistic},k})$ and $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\text{logistic},k})$. Since it is a constant function, s-sensitivity sampling is equivalent to sampling according to P. It is worth noting that logistic is a 1-Lipschitz, convex, and decreasing function. For $E_1 = A$, we have

$$\mathop{\mathbb{E}}_{x\sim p} \text{logistic}\left(\frac{\|x\|_H}{2E_1}\right) \geq \text{logistic}\left(\frac{1}{2}\right) \geq \frac{2}{5} = E_2.$$

This implies that Definition A.1 is satisfied for $\mathcal{L}^H_{\text{logistic},k}$ and thus Theorem A.3 concludes the statement for $m \geq \frac{2S}{\varepsilon^2}\left(8C^2 + S\log\frac{2}{\delta}\right)$. □

In line with the most compelling scenario, Hilbert space $\mathbb{R}^d$, utilizing the standard inner product as its kernel, we present the following theorem.

Theorem D.11. *Let P be a probability measure over $\mathbb{R}^d$ such that $P(\{x\in\mathcal{X}\colon \|x\|_H\geq A\}) = 0$ and $S = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$. There is an $m = O\left(\frac{S}{\varepsilon^2}\left(d\log S + \log\frac{1}{\delta}\right)\right)$ such that any iid sample $x_1,\ldots,x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\text{logistic},k})$ (respectively for $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\text{logistic},k})$) with probability at least $1-\delta$.*

Proof. Because of similarity, we only handle the case $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\text{logistic},k})$. Given that Lemma D.9 asserts the constant function $s(x) = 1 + \frac{340(1+kA^2)}{\sqrt{\max(1,\log(kA^2))}}$ as an upper sensitivity function for $\mathcal{L}_{\text{logistic},k}$, we deduce $\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\text{logistic},k}}, \succ) = \mathbf{Ranges}(\mathcal{L}_{\text{logistic},k}, \succ)$. To clarify, every function in $\mathcal{T}_{\mathcal{L}_{\text{logistic},k}}$ is a function in $\mathcal{L}_{\text{logistic},k}$ scaled by a positive constant. This implies $\text{VCdim}(\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\text{logistic},k}}, \succ)) = \text{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\text{logistic},k}, \succ))$. We would like to remind that

$$\mathcal{T}_{\mathcal{L}_{\text{logistic},k}} = \left\{ f_w(\cdot) = \frac{\text{logistic}(\langle w,\cdot\rangle) + \frac{1}{k}\|w\|_2^2}{s(\cdot)} \colon w\in\mathbb{R}^d \right\}.$$

For an $f_w \in \mathcal{L}_{\text{logistic},k}$ and $r \geq 0$,

$$\begin{aligned}
\mathbf{range}(f_w, \succ, r) &= \left\{x \in \mathbb{R}^d \colon f_w(x) > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \text{logistic}(\langle x, w\rangle) + \frac{\|w\|^2}{k} > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \log\left(1 + e^{-\langle x,w\rangle}\right) > \underbrace{r - \frac{\|w\|^2}{k}}_{=t}\right\} \\
&= \begin{cases} \mathbb{R}^d & t \leq 0 \\ \left\{x \in \mathbb{R}^d \colon \langle x, w\rangle > \log(e^t - 1)\right\} & t > 0, \end{cases}
\end{aligned}$$

which concludes that $\mathbf{Ranges}(\mathcal{L}_{\text{logistic},k}, \succ)$ only includes half-spaces, empty set, and the whole space $\mathbb{R}^d$. Therefore, by Radon's theorem, $\text{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\text{logistic},k}, \succ)) \leq d+1$ (for a proof, see Lemma 10.3.1 in Matousek 2002). Leveraging Theorem 1.2, we derive the statement for $m = O(\frac{S}{\varepsilon^2}\left(d \log S + \log \frac{1}{\delta}\right))$, thereby completing the proof. □

In the previous theorem, we assumed that the data is hard-bounded. In the following results, we try to relax this assumption to a kind of soft bounding similar to what we have in Theorem D.6.

Lemma D.12. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K} \colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim p} \|x\|_H^2 \leq E_1^2$. Then*

$$s(y) = 2 + \left(2 + \frac{\|y\|_H}{E_1}\right) \frac{680(1 + kE_1^2)}{\sqrt{\max(1, \log(2E_1^2 k)}} + \frac{2kE_1^2 + 2kE_1\|y\|_H}{\sqrt{\max(1, \log(2E_1^2 k)}}$$

is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}^H_{\text{logistic},k})$ and $(\mathcal{X}, P, \bar{\mathcal{L}}^H_{\text{logistic},k})$ with total sensitivity $S = O\left(\frac{E_1^2 k}{\sqrt{\log(2E_1^2 k)}}\right)$.

Proof. As $\mathbb{E}_{x\sim P}(\|x\|_H^2) \leq E_1^2$, Markov's inequality implies $\mathbb{P}_{x\sim P}(\|x\|_H^2 \geq 2E_1^2) \leq \frac{1}{2}$. For an arbitrary $y \in \mathcal{X}$, define $\mathcal{X}_y = \left\{x \in \mathcal{X} \colon \|x\|_H \leq \sqrt{2}E_1 \max(1, \frac{\|y\|_H}{E_1})\right\}$. Additionally, define

$$\mathcal{W} = \left\{w \in R^d \colon \|w\|_H \leq \frac{1}{\sqrt{2}E_1}\sqrt{\max(1, \log(2E_1^2 k)}\right\}.$$

Lemma D.8 indicates that Lemma D.2 is applicable with

$$B_1 = \sqrt{2}E_1 \max(1, \frac{\|y\|_H}{E_1}), B_2 = \frac{1}{\sqrt{2}E_1}\sqrt{\max(1, \log(2E_1^2 k)},$$

$$\begin{aligned}
M_y = \phi\left(-B_1 B_2\right) &\leq 1 + \max(1, \frac{\|y\|_H}{E_1})\sqrt{\max(1, \log(2E_1^2 k)} \\
&\leq \left(2 + \frac{\|y\|_H}{E_1}\right)\sqrt{\max(1, \log(2E_1^2 k)},
\end{aligned}$$

and

$$\beta(\alpha) = \beta_{\text{logistic}}(\alpha) = \begin{cases} \frac{85\alpha^2 k}{\log(\alpha^2 k)} & \alpha^2 k \geq e \\ 84 & \alpha^2 k \leq e. \end{cases}$$

Consequently,

$$\begin{aligned}
\sup_{\|w\|_H\le B_2}\frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)} &= \sup_{\|w\|_H\le B_2}\frac{1}{\int_{x\in\mathcal{X}}\frac{\ell^H_{\text{logistic},k}(x,w)}{\ell^H_{\text{logistic},k}(y,w)}\mathrm{d}P(x)}\\
&\le \sup_{\|w\|_H\le B_2}\frac{1}{\int_{x\in\mathcal{X}_y}\frac{\ell^H_{\text{logistic},k}(x,w)}{\ell^H_{\text{logistic},k}(y,w)}\mathrm{d}P(x)}\\
\text{(Lemma D.2)}\quad &\le \sup_{\|w\|_H\le B_2}\frac{1}{\int_{x\in\mathcal{X}_y}\frac{\text{logistic}(0)}{M_y\beta(\|x\|_H)}\mathrm{d}P(x)}\\
&= \frac{M_y}{\text{logistic}(0)}\frac{1}{\int_{x\in\mathcal{X}_y}\frac{1}{\beta(\|x\|_H)}\mathrm{d}P(x)}\\
(*)\quad &\le 2\left(2+\frac{\|y\|_H}{E_1}\right)\sqrt{\max(1,\log(2E_1^2k)}\frac{340(1+kE_1^2)}{\max(1,\log(2kE_1^2))}.\\
&= \left(2+\frac{\|y\|_H}{E_1}\right)\frac{680(1+kE_1^2)}{\sqrt{\max(1,\log(2E_1^2k)}}.
\end{aligned}$$

To see $(*)$, note that $M_y\le\left(2+\frac{\|y\|_H}{E_1}\right)\sqrt{\max(1,\log(2E_1^2k)}$, $\text{logistic}(2)\ge 0.5$, and

$$\begin{aligned}
\int_{x\in\mathcal{X}_y}\frac{1}{\beta(\|x\|_H)}\mathrm{d}P(x) &\ge \int_{\{x:e\le k\|x\|_H^2\le 2kE_1^2\}}\frac{\log(\|x\|_H^2k)}{85\|x\|^2k}\mathrm{d}P(x)+\int_{\{x:k\|x\|_H^2<e\}}\frac{1}{85}\mathrm{d}P(x)\\
&\ge \frac{\log(2kE_1^2)}{170kE_1^2}P\left(\{x:e\le k\|x\|_H^2\le 2kE_1^2\}\right)+\frac{1}{85}P\left(\{x:k\|x\|_H^2<e\}\right)\\
&\ge \frac{\max(1,\log(2kE_1^2))}{170(1+kE_1^2)}P\left(\{x:\|x\|_H^2\le 2E_1^2\}\right)\\
&\ge \frac{\max(1,\log(2kE_1^2))}{340(1+kE_1^2)}.
\end{aligned}$$

Up to this point, our analysis was concentrated on the scenario where $\|w\|_H\le B_2$. Subsequently, we address the case that $\|w\|_H\ge B_2=\frac{1}{\sqrt{2}E_1}\sqrt{\max(1,\log(2E_1^2k)}$. Since $\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)\ge\frac{1}{k}\|w\|_H^2$, we have

$$\begin{aligned}
\sup_{\|w\|_H\ge B_2}\frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)} &\le \sup_{\|w\|_H\ge B_2}\frac{\frac{1}{k}\|w\|_H^2+\log(1+e^{-\mathsf{K}(y,w)})}{\frac{1}{k}\|w\|_H^2}\\
&\le \sup_{\|w\|_H\ge B_2}\left[1+\frac{k}{\|w\|_H^2}\log(1+e^{-\mathsf{K}(y,w)})\right]\\
&\le \sup_{\|w\|_H\ge B_2}\left[1+\frac{k}{\|w\|_H^2}\log(1+e^{\|y\|_H\|w\|_H})\right]\\
&\le \sup_{\|w\|_H\ge B_2}\left[1+\frac{k}{\|w\|_H^2}(1+\|y\|_H\|w\|_H)\right]\\
&\le 2+\frac{k}{B_2^2}+\frac{k\|y\|_H}{B_2}\\
&= 2+\frac{2kE_1^2}{\max(1,\log(2E_1^2k)}+\frac{\sqrt{2}kE_1\|y\|_H}{\sqrt{\max(1,\log(2E_1^2k)}}\\
&\le 2+\frac{2kE_1^2+2kE_1\|y\|_H}{\sqrt{\max(1,\log(2E_1^2k)}}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{w\in\mathcal{X}}\frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)} &\le \max\left\{\sup_{\|w\|\le B_2}\frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)},\right.\\ &\qquad\left.\sup_{\|w\|\ge B_2}\frac{\ell^H_{\text{logistic},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\text{logistic},k}(x,w)\mathrm{d}P(x)}\right\}\\ &\le \max\left\{\left(2+\frac{\|y\|_H}{E_1}\right)\frac{680(1+kE_1^2)}{\sqrt{\max(1,\log(2E_1^2k)}},\ 2+\frac{2kE_1^2+2kE_1\|y\|_H}{\sqrt{\max(1,\log(2E_1^2k)}}\right\}\\ &\le 2+\left(2+\frac{\|y\|_H}{E_1}\right)\frac{680(1+kE_1^2)}{\sqrt{\max(1,\log(2E_1^2k)}}+\frac{2kE_1^2+2kE_1\|y\|_H}{\sqrt{\max(1,\log(2E_1^2k)}}=s(y).\end{aligned}$$

Consequently, $s(y)$ is an upper sensitivity for $(\mathcal{X},\ P,\ \mathcal{L}^H_{\text{logistic},k})$. Since $\mathbb{E}_{y\sim P}(\|y\|_H)\le\sqrt{\mathbb{E}_{y\sim P}(\|y\|_H^2)}\le E_1$, for the total sensitivity, we have

$$S=\int_{y\in\mathcal{X}}s(y)\mathrm{d}P(y)=O\left(\frac{E_1^2k}{\sqrt{\max(1,\log(2E_1^2k)}}\right).$$

A similar argument concludes the statement for $(\mathcal{X},P,\bar{\mathcal{L}}^H_{\text{logistic},k})$. □

Theorem D.13. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon\mathcal{X}\times\mathcal{X}\longrightarrow\mathbb{R}$, P is probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim p}\|x\|_H^2\le E_1$,*

$$s(x)=2+\left(2+\frac{\|y\|_H}{E_1}\right)\frac{680(1+kE_1^2)}{\sqrt{\max(1,\log(2E_1^2k)}}+\frac{2kE_1^2+2kE_1\|y\|_H}{\sqrt{\max(1,\log(2E_1^2k)}},$$

and $S=O\left(\frac{E_1^2k}{\sqrt{\max(1,\log(2E_1^2k)}}\right)$. For $m\ge\frac{2S}{\varepsilon^2}\left(8C^2+S\log\frac{2}{\delta}\right)$, any s-sensitivity sample $x_1,\ldots,x_m$ from $\mathcal{X}$ with weights $u_i=\frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathcal{X},P,\mathcal{X},\ell_{\text{logistic},k})$ (respectively for $(\mathcal{X},P,\mathcal{X},\bar{\ell}_{\text{logistic},k})$) with probability at least $1-\delta$, where $C=(2E_1+1)\max(4E_1k,2.5)+8E_1k+1$.

Proof. Lemma D.12 indicates that $s(\cdot)$ and S are an upper sensitivity function its corresponding total sensitivity for both $(\mathcal{X},P,\mathcal{L}^H_{\text{logistic},k})$ and $(\mathcal{X},P,\bar{\mathcal{L}}^H_{\text{logistic},k})$. Note that $\text{logistic}(\cdot)$ is decreasing and convex. using Equation (6), we have

$$\mathop{\mathbb{E}}_{x\sim p}\text{logistic}\left(\frac{\|x_i\|_H}{2E_1}\right)\ge\text{logistic}\left(\frac{1}{2}\right)\ge\frac{2}{5}=E_2.$$

Now, the proof follows by applying Theorem C.6. □

Returning to the scenario where our Hilbert space is $\mathbb{R}^d$ equipped with the standard inner product as its kernel, to utilize Theorem 1.2, the computation of $\text{VCdim}(\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\text{logistic},k}},\succ))$ is necessary. The following theorem serves as a valuable instrument for handling the VC-dimension of linked-range spaces.

For a positive integer l, a function $f:\mathbb{R}^d\longrightarrow[0,\infty)$ is called *l-simply computable*, if the inequality $f(x)>r$ can be verified using $O(d^{l-1})$ steps via simple arithmetic operations $+,-,\times,/$, jumps conditioned on $>,\ \ge,\ <,\ \le,\ =,\ \neq$ comparisons on real numbers, and $O(1)$ evaluations of the exponential function $t\to e^t$ on real number t.

Theorem D.14 (Anthony & Bartlett 2009, Theorem 8.14)**.** *For a family of functions $\mathcal{F}$, if each $f\in\mathcal{F}$ is l-simply computable, then*

$$\text{VCdim}(\mathbf{Ranges}(\mathcal{F},\succ))=O(d^l).$$

For each $r\ge0$,

$$\mathbf{range}(w,\succ,r)=\left\{x\in\mathbb{R}^d\colon f_w(x)=\frac{\text{logistic}(\langle w,\cdot\rangle)+\frac{1}{k}\|w\|^2}{s(\cdot)}>r\right\}.$$

The inequality $f_w(x) > r$ is not simply computable since it contains the evaluation $\|x\|$ (to compute $s(x)$), which cannot be done by the allowed operations. To circumvent this issue, we propose a modification to the upper sensitivity function s outlined in Lemma D.12 as follows:

$$\begin{aligned} s(x) &= 2 + \left(2 + \frac{\|y\|_H}{E_1}\right) \frac{680(1+kE_1^2)}{\sqrt{\max(1, \log(2E_1^2 k)}} + \frac{2kE_1^2 + 2kE_1\|y\|_H}{\sqrt{\max(1, \log(2E_1^2 k)}} \\ &= 2 + \left(3 + \frac{\|y\|_H^2}{E_1^2}\right) \frac{680(1+kE_1^2)}{\sqrt{\max(1, \log(2E_1^2 k)}} + \frac{2kE_1^2 + 2k(E_1^2 + \|y\|_H^2)}{\sqrt{\max(1, \log(2E_1^2 k)}} \\ &= s_1(x). \end{aligned}$$

Note that s_1 is also an upper sensitivity function for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\text{logistic},k})$ with total sensitivity $S_1 = O\left(\frac{E_1^2 k}{\sqrt{\log(E_1^2 k)}}\right)$.

Now, if we work with s_1 as the upper sensitivity function for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\text{logistic},k})$, then $f_w(x) = \frac{\text{logistic}(\langle w,x\rangle) + \frac{1}{k}\|w\|_2^2}{s_1(\cdot)} > r$ is 2-simply computable. This is true since

$$f_w(x) > r \qquad \iff \qquad 1 + e^{-\langle w,x\rangle} > \exp\left\{s_1(x)r - \frac{1}{k}\|w\|^2\right\}.$$

So, using Theorems 1.2 and D.14, we have the next statement.

Theorem D.15. *Let P be a probability measure over $\mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{x\sim p} \|x\|_2^2 \le E_1^2$,*

$$s(x) = 2 + \left(3 + \frac{\|y\|_H^2}{E_1^2}\right) \frac{680(1+kE_1^2)}{\sqrt{\max(1, \log(2E_1^2 k)}} + \frac{2kE_1^2 + 2k(E_1^2 + \|y\|_H^2)}{\sqrt{\max(1, \log(2E_1^2 k)}},$$

and $S = O\left(\frac{E_1^2 k}{\sqrt{\log(E_1^2 k)}}\right)$. There is $m = O(\frac{2S}{\varepsilon^2}\left(d^2 \log S + \log 1/\delta\right))$ such that any s-sensitivity sample $x_1, \ldots, x_m$ from $\mathcal{X}$ with weights $u_i = \frac{S}{m s(x_i)}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\text{logistic},k})$ (respectively for $(\mathbb{R}^d, P, \mathbb{R}^d, \bar{\ell}_{\text{logistic},k})$) with probability at least $1 - \delta$.

One might question the necessity of having Theorems D.10 and D.7 while Theorems D.13 and D.15 appear stronger and broader. It is crucial to note that in Theorems D.13 and D.15, the process demands re-sampling using s-sensitivity samples, which relies on evaluating the function s. In contrast, Theorems D.10 and D.7 permit the direct (via uniform sampling) use of given data points since they were originally sampled from P.

D.3. Coreset for svm Function

We would like to remind the reader that $\text{svm}(t) = \max(0, 1 - t)$. The SVM loss, also referred to as Hinge Loss, serves as a loss function in training classifiers, such as in Support Vector Machines. Following a similar approach to the one outlined above, we initiate the process by identifying a function $\beta(\cdot)$ that satisfies the properties stated in Lemma D.2 for $\phi = \text{svm}$.

Lemma D.16. *For $\text{svm}(x) = \max(0, 1 - t)$ and $k > 0$, we have*

$$\frac{\text{svm}(-\alpha z) + \frac{z^2}{k}}{\text{svm}(\alpha z) + \frac{z^2}{k}} \le 1 + 2\max(1, \alpha^2 k) \qquad \forall \alpha, z \ge 0.$$

Proof. For a fixed $\alpha > 0$, set $t = \alpha z$, $K = \alpha^2 k$, and

$$\frac{\text{svm}(-\alpha z) + \frac{z^2}{k}}{\text{svm}(\alpha z) + \frac{z^2}{k}} = \frac{\text{svm}(-t) + \frac{t^2}{K}}{\text{svm}(t) + \frac{t^2}{K}} = f_K(t) \qquad t \ge 0.$$

We can rewrite f_K as

$$f_K(t) = \begin{cases} \frac{1+t+\frac{t^2}{K}}{\frac{t^2}{K}} & 1 \le t \\ \\ \frac{1+t+\frac{t^2}{K}}{1-t+\frac{t^2}{K}} & 0 \le t \le 1 \end{cases} \tag{27}$$

$$= \begin{cases} \frac{K}{t^2} + \frac{K}{t} + 1 & 1 \le t \\ \\ 1 + \frac{2t}{1-t+\frac{t^2}{K}} & 0 \le t \le 1 \end{cases} \tag{28}$$

One can simply check that $\max_{t\ge 1} f_K(t) = 1 + 2K$. Also, if we assume that $K \ge 1$, then, by simple calculation, we can derive that $\max_{0\le t\le 1} f_K(t) = f_K(1) = 1 + 2K$. So, using Lemma D.1, we obtain

$$\sup_{t\ge 0} f_K(t) \le 1 + 2\max(1, K) \qquad \forall K > 0.$$

Replacing K with $\alpha^2 z$ concludes the result. □

Lemma D.17. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $P(\{x \in \mathcal{X}\colon \|x\|_H \ge A\}) = 0$. Then $s(y) = 6 + 4kA^2$ is an upper sensitivity function for $(\mathcal{X},\ P,\ \mathcal{L}^H_{\mathrm{svm},k})$ and $(\mathcal{X},\ P,\ \bar{\mathcal{L}}^H_{\mathrm{svm},k})$.*

Proof. Since $P(\{x \in \mathcal{X}\colon \|x\|_H \ge A\}) = 0$, our computation of $s(y)$ is limited to $\|y\|_H \le A$. For $w \in \mathcal{W} = \{w \in \mathcal{X}\colon \|w\|_H \le \frac{1}{A}\}$, Lemma D.16 indicates that Lemma D.2 is applicable with $B_1 = A, B_2 = \frac{1}{A}, M = \phi(-1) = 2$, and $\beta_{\mathrm{svm}}(\alpha) = 1 + 2\max(1, \alpha^2 k)$. Using Lemma D.3 with $\gamma(\|x\|_H) = \frac{\mathrm{svm}(0)}{M\beta_{\mathrm{svm}}(\|x\|_H)} = \frac{1}{2\beta_{\mathrm{svm}}(\|x\|_H)}$, we conclude that

$$\sup_{\|w\|_H\le\frac{1}{A}} \frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathcal{X}} \ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P(x)} \le \frac{2}{\int_{\mathcal{X}} \frac{1}{\beta_{\mathrm{svm}}(\|x\|_H)}\mathrm{d}P(x)}.$$

Note that

$$\begin{aligned}
\int_{x\in\mathcal{X}} \frac{1}{\beta_{\mathrm{svm}}(\|x\|_H)}\mathrm{d}P(x) &\ge \int_{\{x:1\le k\|x\|_H^2\}} \frac{1}{1+2k\|x\|_H^2}\mathrm{d}P(x) + \int_{\{x:k\|x\|^2<1\}} \frac{1}{3}\mathrm{d}P(x) \\
&\ge \frac{1}{1+2kA^2} P\left(\left\{x : \frac{1}{k} \le \|x\|_H^2\right\}\right) + \frac{1}{3}P\left(\left\{x : \|x\|_H^2 < \frac{1}{k}\right\}\right) \\
&\ge \frac{1}{3+2kA^2}.
\end{aligned}$$

This concludes

$$\sup_{\|w\|_H\le\sqrt{k}} \frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathcal{X}} \ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P(x)} \le 6 + 4kA^2.$$

Since $\int_{x\in\mathbb{R}^d} \ell_{\mathrm{svm},k}(x,w)\mathrm{d}P \ge \frac{1}{k}\|w\|^2$, we have

$$\begin{aligned}
\sup_{\|w\|_H\ge\frac{1}{A}} \frac{\ell_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathcal{X}} \ell_{\mathrm{svm},k}(x,w)\mathrm{d}P} &\le \sup_{\|w\|_H\ge\frac{1}{A}} \frac{\frac{1}{k}\|w\|_H^2 + \max(0, 1-\langle y,w\rangle_H)}{\frac{1}{k}\|w\|_H^2} \\
&= \sup_{\|w\|_H\ge\frac{1}{A}} \left[1 + \frac{k}{\|w\|_H^2}\max(0, 1-\langle y,w\rangle_H)\right] \\
&\le \sup_{\|w\|_H\ge\frac{1}{A}} \left[1 + \frac{k}{\|w\|_H^2}\max(0, 1+\|y\|_H\|w\|_H)\right] \\
&\le \sup_{\|w\|_H\ge\frac{1}{A}} \left[1 + \frac{k}{\|w\|_H^2}(1+\|y\|_H\|w\|_H)\right] \\
&\le 1 + 2A^2k.
\end{aligned}$$

So, $s(y) = 6 + 4kA^2$ is an upper sensitivity function for $(\mathcal{X},\ P,\ \mathcal{L}^H_{\mathrm{svm},k})$. □

Theorem D.18. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $P(\{x \in \mathcal{X}\colon \|x\|_H \geq A\}) = 0$. For $S = 6 + 4kA^2$, $C = (4A+2)\max(2Ak, 1) + 8Ak + 1$, and $m = \frac{2S}{\varepsilon^2}(8C^2 + S\log\frac{2}{\delta})$, any iid sample $x_1, \ldots, x_m$ from $\mathcal{X}$ with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathcal{X}, P, \mathcal{X}, \mathcal{L}^H_{\mathrm{svm},k})$ (respectively for $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\mathrm{svm},k})$) with probability at least $1-\delta$.*

Proof. Lemma D.21 indicates that $s(y) = 6 + 4kA^2$ is an upper sensitivity function. As it is a constant function, s-sensitivity sampling is the same as sampling according to P. Notice that svm is convex and non-increasing. For $E_1 = A$, using Equation 6, we have

$$\mathop{\mathbb{E}}_{x\sim p} \mathrm{svm}\left(\frac{\|x_i\|_2}{2E_1}\right) \geq \mathrm{svm}\left(\frac{1}{2}\right) \geq \frac{1}{2} = E_2.$$

This implies that Definition A.1 is satisfied for $\mathcal{L}^H_{\mathrm{svm},k}$ and thus Theorem A.3 concludes the statement for $m \geq \frac{2S}{\varepsilon^2}\left(8C^2 + \log\frac{1}{\delta}\right)$. □

Following directly from this theorem, we observe the following corollary.

Corollary D.19 (Theorem A.5, Restated)**.** *Let $\mathsf{K} : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow (0,1]$ be a reproducing kernel, i.e., a kernel associated with an RKHS, and P be a probability measure over $\mathbb{R}^d$. For $\varepsilon, \delta \in (0,1)$, there exists a universal constant C (independent of d and K) such that if $m \geq \frac{C}{\varepsilon^2}\log\frac{1}{\delta}$, then, with probability at least $1-\delta$, for any random sample $X = \{x_1, \ldots, x_m\}$ based on P, we have*

$$\sup_{w\in\mathbb{R}^d} |\mathrm{KDE}_P(w) - \mathrm{KDE}_X(w)| \leq \varepsilon.$$

Proof. Note that K and $\phi(t) = \max(0, 1-t)$ satisfy Definition A.1 with $E_1 = 1$, $E_2 = \frac{1}{2}$, and

$$\ell^H_{\phi,k}(x,w) = \phi(\mathsf{K}(x,w)) + \frac{1}{k}\|w\|^2_H = 1 - \mathsf{K}(x,w) + \frac{1}{k}\|w\|^2_H \leq 1 - \mathsf{K}(x,w) + \frac{1}{k}.$$

If set $k = 1$, then, with probability at least $1-\delta$, for each $w \in \mathbb{R}^d$, we have

$$\begin{aligned}
\left|\int_{x\in\mathbb{R}^d} \mathsf{K}(x,w)\mathrm{d}P(x) - \frac{1}{m}\sum_{i=1}^m \mathsf{K}(x_i,w)\right| &= \left|\int_{x\in\mathbb{R}^d} \ell^H_{\phi,1}(x,w)\mathrm{d}P(x) - \frac{1}{m}\sum_{i=1}^m \ell^H_{\phi,1}(x_i,w)\right| \\
\text{(by Theorem D.18)} \qquad &\leq \frac{\varepsilon}{2}\int_{x\in\mathbb{R}^d} \ell^H_{\phi,1}(x,w)\mathrm{d}P(x) \\
&\leq \frac{\varepsilon}{2}\int_{x\in\mathbb{R}^d} (2 - \mathsf{K}(x,w))\mathrm{d}P(x) \\
\text{(since } \mathsf{K}(x,w) \leq 1) \qquad &\leq \varepsilon.
\end{aligned}$$

□

Theorem D.20. *Let P be a probability measure over $\mathbb{R}^d$ such that $P(\{x \in \mathcal{X}\colon \|x\|_H \geq A\}) = 0$ and $S = 6 + 4kA^2$. There exists $m = O\left(\frac{S}{\varepsilon^2}(d\log S + \log\frac{1}{\delta})\right)$ such that any iid sample $x_1, \ldots, x_m$ from P with weights $u_i = \frac{1}{m}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\mathrm{svm},k})$ (respectively for $(\mathcal{X}, P, \mathcal{X}, \bar{\ell}^H_{\mathrm{svm},k})$with probability at least $1-\delta$.*

Proof. Given that Lemma D.21 asserts the constant function $s(x) = 3 + 2kA^2 + \sqrt{k}A$ as an upper sensitivity function for $\mathcal{L}_{\mathrm{svm},k}$, we deduce $\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\mathrm{svm},k}}, \succ) = \mathbf{Ranges}(\mathcal{L}_{\mathrm{svm},k}, \succ)$. Notice $\mathbf{Ranges}(\mathcal{T}_{\mathcal{L}_{\mathrm{svm},k}}, \succ) = \mathbf{Ranges}(\mathcal{L}_{\mathrm{svm},k}, \succ)$

since each function in $\mathcal{T}_{\mathcal{L}_{\text{svm},k}}$ is a function in $\mathcal{L}_{\text{svm},k}$ scaled by a positive constant. For an $f_w \in \mathcal{L}_{\text{svm},k}$ and $r \geq 0$,

$$\begin{aligned}
\mathbf{range}(f_w, \succ, r) &= \left\{x \in \mathbb{R}^d \colon f_w(x) > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \max(0, 1 - \langle x, w\rangle) + \frac{\|w\|^2}{K} > r\right\} \\
&= \left\{x \in \mathbb{R}^d \colon \max(0, 1 - \langle x, w\rangle) > \underbrace{r - \frac{\|w\|^2}{K}}_{=t}\right\} \\
&= \begin{cases} \mathbb{R}^d & t < 0 \\ \left\{x \in \mathbb{R}^d \colon \langle x, w\rangle < 1 - t\right\} & t \geq 0, \end{cases}
\end{aligned}$$

which concludes that $\mathbf{Ranges}(\mathcal{L}_{\text{svm},k}, \succ)$ only includes half-spaces and the whole space $\mathbb{R}^d$. Therefore, by Radon's theorem, $\text{VCdim}(\mathbf{Ranges}(\mathcal{L}_{\text{svm},k}, \succ)) \leq d+1$ (for a proof, see Lemma 10.3.1 in (Matousek, 2002)). Using Theorem 1.2, we also have the statement of Theorem D.10 for $m \geq O(\frac{S}{\varepsilon^2}\left(d \log S + \log \frac{1}{\delta}\right))$, completing the proof. A similar argument establishes the result for $(\mathcal{X}, P, \bar{\mathcal{L}}^H_{\text{svm},k})$. □

Continuing in line with prior sections, we aim to establish a theorem that addresses-soft bounding on data instead of employing a hard boundary. In the subsequent discussion, we substantiate such a result.

Lemma D.21. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K} \colon \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $\mathbb{E}_{x\sim p} \|x\|_H \leq E_1^2$. Then $s(x) = (6 + 8kE_1^2)\left(2 + \frac{\|y\|_H}{E_1}\right)$ is an upper sensitivity function for $(\mathcal{X}, P, \mathcal{L}^H_{\text{svm},k})$ and $(\mathcal{X}, P, \bar{\mathcal{L}}^H_{\text{svm},k})$ with total sensitivity at most $18 + 24kE_1^2$.*

Proof. As $\mathbb{E}_{x\sim P}(\|x\|_H^2) \leq E_1^2$, Markov's inequality implies $\mathbb{P}_{x\sim P}(\|x\|_H^2 \geq 2E_1^2) \leq \frac{1}{2}$. For an arbitrary $y \in \mathcal{X}$, define $\mathcal{X}_y = \left\{x \in \mathcal{X} \colon \|x\|_H \leq \sqrt{2}\max(E_1, \|y\|_H)\right\}$. Additionally, define $\mathcal{W} = \{w \in \mathcal{X} \colon \|w\|_H \leq \frac{1}{\sqrt{2}E_1}\}$. Lemma D.8 indicates that Lemma D.2 is applicable with $B_1 = \sqrt{2}\max(E_1, \|y\|_H)$, $B_2 = \frac{1}{\sqrt{2}E_1}$, $\text{svm}(-B_1B_2) = 1 + \max\left(1, \frac{\|y\|_H}{E_1}\right) \leq 2 + \frac{\|y\|_H}{E_1} = M_y$, and

$$\beta(\alpha) = \beta_{\text{svm}}(\alpha) = 1 + 2\max(1, \alpha^2 k).$$

Consequently,

$$\begin{aligned}
\sup_{\|w\|_H \leq B_2} \frac{\ell^H_{\text{svm},k}(y, w)}{\int_{x\in\mathcal{X}} \ell^H_{\text{svm},k}(x, w)\mathrm{d}P(x)} &= \sup_{\|w\|_H \leq B_2} \frac{1}{\int_{x\in\mathcal{X}} \frac{\ell^H_{\text{svm},k}(x,w)}{\ell^H_{\text{svm},k}(y,w)}\mathrm{d}P(x)} \\
&\leq \sup_{\|w\|_H \leq B_2} \frac{1}{\int_{x\in\mathcal{X}_y} \frac{\ell^H_{\text{svm},k}(x,w)}{\ell^H_{\text{svm},k}(y,w)}\mathrm{d}P(x)} \\
\text{(Lemma D.2)} \quad &\leq \sup_{\|w\|_H \leq B_2} \frac{1}{\int_{x\in\mathcal{X}} \frac{\text{svm}(0)}{M_y\beta(\|x\|_H)}\mathrm{d}P(x)} \\
&= \frac{M_y}{\int_{x\in\mathcal{X}_y} \frac{1}{\beta(\|x\|_H)}\mathrm{d}P(x)} \\
&\leq \frac{2 + \frac{\|y\|_H}{E_1}}{\int_{x\in\mathcal{X}_y} \frac{1}{\beta(\|x\|_H)}\mathrm{d}P(x)} \\
(*) \quad &\leq (6 + 8kE_1^2)\left(2 + \frac{\|y\|_H}{E_1}\right).
\end{aligned}$$

To see $(*)$, note that

$$\begin{aligned}\int_{x\in\mathcal{X}_y}\frac{1}{\beta(\|x\|_H)}\mathrm{d}P &\geq \int_{\{x:1\leq k\|x\|_H^2\leq 2kE_1^2\}}\frac{1}{1+2k\|x\|_H^2}\mathrm{d}P+\int_{\{x:k\|x\|_H^2<1\}}\frac{1}{3}\mathrm{d}P\\ &\geq \frac{1}{1+4kE_1^2}P\left(\{x:1\leq k\|x\|_H^2\leq 2kE_1^2\}\right)+\frac{1}{3}P\left(\{x:k\|x\|_H^2<1\}\right)\\ &\geq \min\left(\frac{1}{1+4kE_1^2},\frac{1}{3}\right)P\left(\{x:\|x\|_H^2\leq 2E_1^2\}\right)\\ &\geq \frac{1}{6+8kE_1^2}.\end{aligned}$$

As we have only considered the case that $\|w\|_H\leq B_2$, to complete the proof, we should work with $\|w\|_H\geq B_2$ as well. Since $\int_{x\in\mathcal{X}}\ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P\geq\frac{1}{k}\|w\|_H^2$, we have

$$\begin{aligned}\sup_{\|w\|_H\geq B_2}\frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathcal{X}}\ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P} &\leq \sup_{\|w\|_H\geq B_2}\frac{\frac{1}{k}\|w\|_H^2+\max(0,1-\langle y,w\rangle_H)}{\frac{1}{k}\|w\|_H^2}\\ &\leq \sup_{\|w\|_H\geq B_2}\left[1+\frac{k}{\|w\|_H^2}\max(0,1-\langle y,w\rangle_H)\right]\\ &\leq \sup_{\|w\|_H\geq B_2}\left[1+\frac{k}{\|w\|_H^2}\max(0,1+\|y\|_H\|w\|_H)\right]\\ &= \sup_{\|w\|_H\geq B_2}\left[1+\frac{k}{\|w\|_H^2}(1+\|y\|_H\|w\|_H)\right]\\ &\leq 1+2kE_1^2+\sqrt{2}kE_1\|y\|_H.\end{aligned}$$

Therefore,

$$\begin{aligned}\sup_{w\in\mathcal{X}}\frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathbb{R}^d}\ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P} \leq &\max\left\{\sup_{\|w\|\leq\sqrt{k}}\frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathbb{R}^d}\ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P},\right.\\ &\left.\sup_{\|w\|\geq\sqrt{k}}\frac{\ell^H_{\mathrm{svm},k}(y,w)}{\int_{x\in\mathbb{R}^d}\ell^H_{\mathrm{svm},k}(x,w)\mathrm{d}P}\right\}\\ \leq &\max\left\{(6+8kE_1^2)\left(2+\frac{\|y\|_H}{E_1}\right),\ 1+2kE_1^2+\sqrt{2}kE_1\|y\|_H\right\}\\ = &(6+8kE_1^2)\left(2+\frac{\|y\|_H}{E_1}\right)=s(y).\end{aligned}$$

Consequently, $s(y)$ is an upper sensitivity for $(\mathbb{R}^d,P,\mathbb{R}^d,\ell_{\mathrm{svm},k})$. Since $\mathbb{E}_{y\sim P}(\|y\|_H)\leq\sqrt{\mathbb{E}_{y\sim P}(\|y\|_H^2)}\leq E_1$, for the total sensitivity, we have

$$S=\int s(y)\mathrm{d}P\leq(18+24kE_1^2)=O(E_1^2k).$$

□

With a similar proof, we can restate the following versions of Theorems D.13 and D.15 replacing logistic by svm.

Theorem D.22. *Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon\mathcal{X}\times\mathcal{X}\longrightarrow\mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim p}\|x\|_H^2\leq E_1$,*

$$s(x)=(6+8kE_1^2)\left(2+\frac{\|y\|_H}{E_1}\right),$$

and $S=O(E_1^2k)$. For $m\geq\frac{2S}{\varepsilon^2}\left(8C^2+S\log\frac{2}{\delta}\right)$, any s-sensitivity sample $x_1,\ldots,x_m$ from $\mathcal{X}$ with weights $u_i=\frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathcal{X},P,\mathcal{X},\ell^H_{\mathrm{svm},k})$) respectively for $(\mathcal{X},P,\mathcal{X},\bar{\ell}^H_{\mathrm{svm},k}))$ with probability at least $1-\delta$, where $C=(4E_1+2)\max(E_1k,1)+8E_1k+1$.

Theorem D.23. *Let P be a probability measure over $\mathbb{R}^d$ such that $\mathop{\mathbb{E}}_{x\sim p}\|x\|_2^2 \leq E_1^2$, $s(x) = (6+8kE_1^2)\left(2+\frac{\|y\|_H}{E_1}\right)$, and $S = O\left(E_1^2 k\right)$. For $m \geq O\left(\frac{S}{\varepsilon^2}\left(d^2\log S + \log\frac{1}{\delta}\right)\right)$, any s-sensitivity sample $x_1,\ldots,x_m$ from $\mathcal{X}$ with weights $u_i = \frac{S}{ms(x_i)}$ provides an ε-coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\mathrm{svm},k})$) respectively for $(\mathbb{R}^d, P, \mathbb{R}^d, \bar{\ell}^H_{\mathrm{svm},k})$) with probability at least $1-\delta$.*

D.4. Coreset for ReLU Function

For $\mathrm{ReLU}(t) = \max(0,t)$, we derive our results from those in previous section using a simple trick. Assume that H is a reproducing kernel Hilbert space of real-valued functions on $\mathcal{X}$ with kernel $\mathsf{K}\colon \mathcal{X}\times\mathcal{X} \longrightarrow \mathbb{R}$ and P is a probability measure over $\mathcal{X}$ such that $\mathop{\mathbb{E}}_{x\sim p}\|x\|_H^2 \leq E_1^2$ or $P\left(\{x\in\mathcal{X}\colon \|x\|_H \geq A\}\right) = 0$. It is easy to verify that $\mathsf{K}'(\cdot) = 1+\mathsf{K}(\cdot)$ is also a reproducing kernel associated with H' with $\mathop{\mathbb{E}}_{x\sim p}\|x\|_{H'}^2 \leq (1+E_1)^2$ or $P\left(\{x\in\mathcal{X}\colon \|x\|_{H'} \geq 1+A\}\right) = 0$. As $\mathrm{svm}(1+t) = \max(0,-t)$, if we plug K' in the results given in the previous section, we can replicate them for $\phi(t) = \max(0,-t)$ by replacing E_1 and A with $(1+E_1)$ and $(1+A)$. While ϕ is not precisely the ReLU function, for the case that the kernel is the standard linear kernel in Euclidean space,

$$\ell_{\phi,k}(x,w) = \max(0,-\langle x,w\rangle) + \frac{1}{k}\|w\|_2^2 = \max(0,\langle x,-w\rangle) + \frac{1}{k}\|w\|_2^2 = \ell_{\mathrm{ReLU},k}(x,-w).$$

This implies that if we have a coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\phi,k})$, it automatically serves as a coreset for $(\mathbb{R}^d, P, \mathbb{R}^d, \ell_{\mathrm{ReLU},k})$ as well. It is noteworthy that this conclusion cannot be extended to general kernels as $-\mathsf{K}(x,w) = \mathsf{K}(x,-w)$ is not generally true.