Compressing data for generalized linear regression

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Technischen Universität Dortmund

an der Fakultät Statistik

Simon Omlor

Dortmund

2022

Referees:

Dr. Alexander Munteanu

Prof. Dr. Katja Ickstadt

Date of submission: 03.03.2023

Acknowledgments

I would like to thank to Dr. Alexander Munteanu for guiding my research and advising me. I would like to thank Prof. Dr. Katja Ickstadt for helpful discussions. Also, I would like to thank my family for always supporting me and my friends who read over the manuscript and gave helpful hints.

Abstract

In this thesis we work on algorithmic data and dimension reduction techniques to solve scalability issues and to allow better analysis of massive data. For our algorithms we use the sketch and solve paradigm as well as some initialization tricks. We will analyze a tradeoff between accuracy, running time and storage. We also show some lower bounds on the best possible data reduction factors. While we are focusing on generalized linear regression mostly, logistic and p-probit regression to be precise, we are also dealing with two layer Rectified Linear Unit (ReLU) networks with logistic loss which can be seen as an extension of logistic regression, i.e. logistic regression on the neural tangent kernel. We present coresets via sampling, sketches via random projections and several algorithmic techniques and prove that our algorithms are guaranteed to work with high probability.

First, we consider the problem of logistic regression where the aim is to find the parameter β maximizing the likelihood. We are constructing a sketch in a single pass over a turnstile data stream. Depending on some parameters we can tweak size, running time and approximation guarantee of the sketch. We also show that our sketch works for other target functions as well.

Second, we construct an ε-coreset for p-probit regression, which is a generalized version of probit regression. Therefore we first compute the QR decomposition of a sketched version of our dataset in a first pass. We then use the matrix R to compute an approximation of the ℓ_p-leverage scores of our data points which we use to compute sampling probabilities to construct the coreset. We then analyze the negative log likelihood of the p-generalized normal distribution to prove that this results in an ε-coreset.

Finally, we look at two layer ReLU networks with logistic loss. Here we show that using a coupled initialization we can reduce the width of the networks to get a good approximation down from γ^{-8} (Ji and Telgarsky, 2020) to γ^{-2} where γ is the so called separation margin. We further give an example where we prove that a width of γ^{-1} is necessary to get less than constant error.

Contents

1 Introduction and motivation **1**

1.1 Data compression . . . 1

1.2 Neural networks . . . 2

1.3 Problems considered in this manuscript . . . 2

1.4 Outline and results . . . 2

1.5 Publications . . . 3

2 Preliminaries **5**

2.1 General notation . . . 5

2.2 Input formats . . . 5

2.3 Basics on linear algebra . . . 6

2.4 Probability distributions and common inequalities . . . 8

2.5 Data reduction methods . . . 11

2.6 Linear regression, logistic regression and p-probit regression . . . 16

2.7 Artificial neural networks . . . 18

2.7.1 Two-layer ReLU networks . . . 19

2.8 Convex optimization . . . 21

2.8.1 Gradient and Hessian Matrix for p-probit regression . . . 22

2.9 Related work . . . 24

3 Sketching for logistic regression **31**

3.1 Setting and notations . . . 31

3.2 The algorithm . . . 32

3.2.1 Motivation . . . 32

3.2.2 Parameters . . . 32

3.2.3 Pseudo code . . . 33

3.2.4 Description of the algorithm . . . 34

3.2.5 Idea of the analysis . . . 34

3.2.6 Outline of the analysis . . . 35

3.3 High level description of the analysis . . . 35

3.4 Analysis . . . 38

3.4.1 Assumptions . . . 38

3.4.2 Estimating the small parts of f . . . 39

3.4.3 Estimating $\|z^+\|_1$ . . . 41

3.4.4 Analysis for a single level . . . 42

3.4.5 Heavy hitters . . . 47
3.4.6 Contraction bounds for a single point . . . 49
3.4.7 Net argument . . . 50
3.4.8 Dilation bounds . . . 51
3.5 Main result . . . 55
3.6 Extension to linear ℓ_1-regression . . . 56
3.6.1 Dilation bounds for ℓ_1 . . . 56
3.6.2 Net argument . . . 58
3.7 Extension to logistic regression with variance-based regularization . . . 59
3.8 Lower bound . . . 64

4 ℓ_p-leverage score sampling for p-probit regression 67
4.1 Setting and notations . . . 67
4.2 The algorithm . . . 67
4.2.1 High level description . . . 67
4.2.2 Pseudo code . . . 68
4.3 Analysis . . . 68
4.3.1 Outline of the analysis . . . 68
4.3.2 Tails of the p-generalized normal distribution . . . 69
4.3.3 Properties of g . . . 73
4.3.4 Bounding the VC-Dimension . . . 78
4.3.5 Bounding the Sensitivities . . . 81
4.3.6 Well Conditioned Bases and Approximate Leverage Scores . . . 82
4.4 Main Results . . . 85

5 Reducing the width of two layer ReLU networks 87
5.1 Setting and notations . . . 87
5.2 The initialization and its motivation . . . 88
5.3 Outline of the analysis . . . 89
5.4 Main assumption and examples . . . 89
5.4.1 Main assumption . . . 89
5.4.2 Example 1: orthonormal unit vectors . . . 94
5.4.3 Example 2: Two differently labeled points at distance b . . . 94
5.4.4 Example 3: Constant labels . . . 95
5.4.5 Example 4: The hypercube . . . 96
5.5 Lower bounds for log width . . . 98
5.5.1 Example 5: Alternating points on a circle . . . 98

5.5.2 Lower Bounds 100
5.6 Upper bound 102
5.7 On the construction of U 105
5.7.1 Tightness of the construction of U 105
5.7.2 The two dimensional case (upper bound) 106

6 Conclusion and open problems **108**
6.1 Sketching for logistic regression 108
6.2 ℓ_p-leverage score sampling for probit regression 109
6.3 Reducing the width of two layer ReLU networks 109

7 Bibliography **111**

1 Introduction and motivation

With improving hardware and increasing amounts of computation power as well as increasing storage sizes the amount of available data often is massive also known as the phenomenon of Big Data.

Analyzing data has become an important job for computers in the twenty-first century, be it for spotting diseases in a pandemic, efficient finances, machine learning task, such as autonomous driving or to determine the best ways to save energies. Thus always improving algorithms have been developed to deal with data more accurately and more efficiently. However with the increasing amounts of data, which are helpful at first glance as the laws of large numbers and central limit theorems tell us the more data we have the more accurate are our predictions, even the best algorithms are unable to deal with the large amounts of data. Either they are too slow or the memory used is not sufficient. Hence an efficient preprocessing is needed.

Efficient preprocessing is not only helpful from a practical view point it also helps us to understand the mathematics behind the possible distributions of the parameters with respect to the optimized target function.

This thesis provides new methods to compress data without losing important information with respect to different target functions and analyzes the amount of compression as well as the running time and the approximation guarantee.

1.1 Data compression

In the last century a lot of algorithms have been developed to analyze datasets. While these work well on small and medium size datasets, they become inefficient as the size of the dataset starts to grow. To deal with this data compression is used. More precisely before running the algorithm to get an accurate analysis we first run a fast algorithm that drops unimportant data and merges similar data points.

Sketch and solve paradigm (Woodruff (2014); Munteanu (2023))

The following idea which will be used in Section 3 as well as Section 4 is the sketch and solve paradigm: We consider a data matrix $X \in \mathbb{R}^{n\times d}$. The task is to find a parameter $\beta \in \mathbb{R}^d$ such that $f(X\beta)$ is minimized for some function f. Thus our goal is to find a mapping Π which maps X to much smaller dataset $\Pi(X)$, also called coreset. More precisely $\Pi(X)$ will usually be in $\mathbb{R}^{n'\times d}$ where $n' \ll n$ with some additional weights $w \in \mathbb{R}^{n'}$ such that for all $\beta \in B$ we have that $f(X\beta) \approx f_w(X'\beta)$ where f_w is a weighted version of f.

We are using the following three techniques to compress the data:

- **Linear sketching:** The idea of sketching is that we apply some random projection to our data. More precisely given a data matrix $X \in \mathbb{R}^{n\times d}$ we multiply X with a random matrix $S \in \mathbb{R}^{m\times n}$ to get a new smaller dataset $X' = SX \in \mathbb{R}^{m\times d}$ with similar properties, i.e $f(X\beta) \approx f(X'\beta)$ holds for some loss function f.

- **Subsampling:** Sampling is a well known method for reducing the size of datasets. The most commonly used sampling method is uniform sampling where each point is sampled with the same probability. However we will focus on sensitivity sampling where important points have a higher chance of getting picked while unimportant points get dropped and frequent points, which individually are less likely to get picked, get a high weight if picked to indicate their frequency.
- **Dimension reduction:** When considering two layer ReLU-networks a clever initialization allows us to use a smaller inner layer for good convergence which can be seen as a compressed version of the so called (infinite dimensional) neural tangent kernel.

1.2 Neural networks

Neural networks have been a popular topic in recent research. While they perform well in practice little is known in theory. They are usually trained given some training data and then they can give predictions for new input points using matrix multiplication and activation functions. In this manuscript we will have a look at two layer ReLU networks in particular with rectified linear unit (ReLU) function. In Section 2 we will give a precise definition for those.

1.3 Problems considered in this manuscript

In everything what follows we are considering a fixed dataset $X \in \mathbb{R}^{n \times d}$. We also have a vector of target values $y \in \mathbb{R}^n$ corresponding to the dataset. In some cases there is an equivalent instance where all labels are equal and we will omit y in this case.

For the first set of problems we are given a target function f and our goal is to find a significantly smaller dataset X' such that $f(X'\beta) \approx f(X\beta)$. Here following functions f are considered:

- logistic regression;
- ℓ_1-regression;
- variance-based regularized logistic regression;
- p-probit regression

Second we consider neural networks. Here we do not shrink the dataset (X, y) itself but rather the number of neurons needed in the middle layer.

1.4 Outline and results

The remaining manuscript is structured into Chapters 2 – 6 dealing with the following content summarized below.

- **Section 2** In this section we introduce general notations. We continue with some basic definitions and results from linear algebra. Then we state the used probability distributions as well as some well known inequalities. Afterwards we give a brief introduction into sketching and sensitivity sampling. We then continue by giving the definitions of some regression models and two layer ReLU networks. We then state the problems we consider and mention some of the related work.

- **Section 3** Here we introduce our first algorithm to construct a sketch for logistic regression. The algorithm works in a single pass over a turnstile datastream. We show that there is a tradeoff between size, running time and approximation guarantee. More precisely we show that with linear running time in the number of non zero entries of our dataset we can get a weak 1-sketch (see Definition 2.18) in expectation, which can be used to compute a 2-approximation. If we allow the sketch size to be exponential in ε^{-1} we can get down to a weak ε-sketch, which can be used to compute a $(1+\varepsilon)$-approximation. Last with some increased running time we can get a sketch of very small size with constant approximation guarantee. More precisely we can get arbitrarily close to size linear in d. For the details see Theorem 1. We conclude the section by showing that the algorithm also works for ℓ_1-regression as well as variance-based regularized logistic regression (see Theorem 2 and Theorem 3).

- **Section 4** In this section we focus on p-generalized probit regression. We show that we can approximate the so called ℓ_p-leverage scores which can be used as sampling probabilities up to some scalar to construct an ε-coreset if the dataset is μ-complex for some $\mu \geq 1$. To show this we analyze the negative log-likelihood of p-generalized normal distribution and determine its non-asymptotic tail behavior.

- **Section 5** Here we consider two layer ReLU networks with logistic loss on the output layer. By using coupled initialization we can improve the upper bound on the width of the networks to get an arbitrarily small error down from $\tilde{O}(\gamma^{-8})$ (Ji and Telgarsky, 2020) to $\tilde{O}(\gamma^{-2})$ where γ is a parameter introduced in (Ji and Telgarsky, 2020) We also improve the lower bounds to $\Omega(\gamma^{-1})$ reducing the gaps between the bounds even further.

- **Section 6** Last we recap our results and state directions for future research.

1.5 Publications

The present manuscript is based on the following publications:

- Section 3 is based on Munteanu et al. (2021),

  Alexander Munteanu, Simon Omlor, and David P. Woodruff. Oblivious sketching for logistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7861–7871, 2021

  and on Munteanu et al. (2023),

  Alexander Munteanu, Simon Omlor, and David P. Woodruff. Almost linear constant-factor sketch-

ing for ℓ_1 and logistic regression. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. to appear

- Section 4 is based on Munteanu et al. (2022a),

  Alexander Munteanu, Simon Omlor, and Christian Peters. p-Generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 2073–2100, 2022a

- Section 5 is based on Munteanu et al. (2022b),

  Alexander Munteanu, Simon Omlor, Zhao Song, and David Woodruff. Bounding the width of neural networks via coupled initialization - a worst case analysis. In *International Conference on Machine Learning*, pages 16083–16122, 2022b

All authors contributed equally and are stated in alphabetical order. My main focus was on the technical parts. Experiments in the stated papers were done by research assistants and guided by us. In (Munteanu et al., 2022b) two loss functions are considered, the logistic loss and the squared loss. We (Munteanu, Omlor, Woodruff) focused on the logistic loss while the squared loss was analyzed by our coauthors and is only mentioned in the related work section in this manuscript.

2 Preliminaries

In this section we will describe the notations we use in this thesis. We will also recall some basic definitions, important theorems, as well as some motivation and literature overview. Readers that are familiar with those might skip this chapter entirely.

2.1 General notation

We are using the following notations:

- We use $\mathbb{N}$ to denote the natural numbers including 0.
- For $x \in \mathbb{R}$ we use the notation $\mathbb{R}_{\geq x} := \{y \in \mathbb{R} \mid y \geq x\}$. Similarly we define $\mathbb{R}_{>x}, \mathbb{R}_{\leq x}$ and $\mathbb{R}_{<x}$.
- For a natural number $n \geq 1$ we set $[n] := \{1, 2, \dots n\}$.
- We use $I_n \in \mathbb{R}^{n \times n}$ to indicate the identity matrix of dimension n.
- We use $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$ for our data matrix and observations. By $x_i \in \mathbb{R}^d$ we denote the i-th row of X.
- Our goal is to minimize a target function which is of the form $f(X, y, \beta) = \sum_{i=1}^{n} g(x_i, y, \beta)$, where g is the loss function of individual points. If we have a weight vector $w \in \mathbb{R}^n$ then the target function is changed to $f_w(X, y, \beta) = \sum_{i=1}^{n} w_i g(x_i, y, \beta)$.
- If $z \in \mathbb{R}^n$ then $z^+ \in \mathbb{R}^n$ is the vector with $z_i^+ = z_i$ if $z_i \geq 0$ and $z_i^+ = 0$ otherwise.
- For a matrix $X \in \mathbb{R}^{n \times d}$ we denote by $\text{nnz}(X)$ the number of non zero entries of X.
- $T = \text{poly}(d)$ means that there exists a constant $c \geq 1$ such that $T = \Theta(d^c)$.

2.2 Input formats

Streaming [see, e.g., Muthukrishnan (2005) for a survey]

In the streaming model we receive our data row by row in a fixed order. We can go through our data multiple times and the approximation guarantee has to be achieved only at the end algorithm. However the data might be saved externally and going over it can take some time thus one our goal is to reduce the memory and the number of passes we need.

Online algorithms In the online setting each data point can only be accessed once and we need to achieve our approximation guarantee at any time. Thus we need to do computations continuously during accessing the points to be able to make final decisions at any time. For more details on online algorithms we refer to (Borodin and El-Yaniv, 2005).

Turnstile datastreams We follow the description of turnstile datastreams from (Munteanu, 2018): In this model we initialize a matrix A to the all-zero matrix. The stream consists of $(key, value)$ updates of the form (i, j, v), meaning that A_{ij} will be updated to $A_{ij} + v$. A single entry can be defined by a single update or by a subsequence of not necessarily consecutive updates. For instance, a sequence $\ldots, (i, j, 27), \ldots, (i, j, -5), \ldots$ will result in $A_{ij} = 22$. Deletions are possible in this setting by using negative updates matching previous insertions. At first glance this model might seem technical or unnatural but we stress that for dealing with unstructured data, the design of algorithms working in the turnstile model is of high importance.

2.3 Basics on linear algebra

ℓ_p-Norm

Definition 2.1. *(Golub and Van Loan (2013)) Given a vector $v \in \mathbb{R}^n$ and $p \in [1, \infty)$ the ℓ_p-norm of v is given by*

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$

and $\|v\|_\infty = \lim_{p\to\infty} \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}$

Note that for any two vectors $x, y \in \mathbb{R}^n$ Hölder's inequality states that $x^T y \leq \|x\|_p \|y\|_q$ for any $p, q \in [1, \infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$. A special case is the is the Cauchy Schwarz inequality (see Golub and Van Loan (2013)) which states that

$$x^T y \leq \|x\|_2 \|y\|_2.$$

Ky-Fan norm The following norm is useful when we are only interested in the large values of a vector. It was used before in (Clarkson and Woodruff, 2015) and is an adaptation of the k-th Ky-Fan-norm for matrixes which is given by the sum of the k largest eigenvalues of the matrix.

Definition 2.2. *Given a vector $v \in \mathbb{R}^n$ let T be the set of the k entries of v with the largest absolute value. Then the k-th Ky-Fan norm of v is defined by*

$$\|v\|_{Kf(k)} = \sum_{v_i \in T} |v_i|.$$

Note that the k-th Ky-Fan norm of v is equal to k-th Ky-Fan-norm of the diagonal matrix D_v with $(D_v)_{ii} = v_i$ and $(D_v)_{ij} = 0$ if $i \neq j$.

Orthonormal matrixes and the QR-decomposition

Definition 2.3. *(Golub and Van Loan (2013)) A Matrix $U \in \mathbb{R}^{n\times d}$ with columns $u^{(1)}, \dots, u^{(d)}$ is called orthonormal if*

- *For any $i \in [d]$ it holds that $\|u^{(i)}\|_2 = 1$;*
- *For any $i, j \in [d]$ with $i \neq j$ it holds that $\langle u^{(i)}, u^{(j)}\rangle = 0$.*

A matrix is orthonormal if and only if $U^T U = I_d$.

Proposition 2.4 (QR-decomposition)**.** *Golub and Van Loan (2013) For any matrix $X \in \mathbb{R}^{n\times d}$ there exist matrices $R \in \mathbb{R}^{d\times d}$ and $Q \in \mathbb{R}^{n\times d}$ such that $X = QR$, Q is orthonormal. If X is of full rank then R is invertible.*

Leverage scores

Definition 2.5. *(Dasgupta et al. (2009)) Given a matrix $X \in \mathbb{R}^{n\times d}$ with rows $x_1 \dots, x_n$ we define the i-th ℓ_p-leverage score of X by*

$$u_i := \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{|x_i\beta|^p}{\sum_{j=1}^n |x_j\beta|^p} = \sup_{\beta\in\mathbb{R}^d, \|\beta\|_2=1} \frac{|x_i\beta|^p}{\sum_{j=1}^n |x_j\beta|^p} = \max_{\beta\in\mathbb{R}^d, \|\beta\|_p=1} \frac{|x_i\beta|^p}{\sum_{j=1}^n |x_i\beta|^p}.$$

Here the last equality follows as $\{\beta \in \mathbb{R}^d, \|\beta\|_2 = 1\}$ is a compact set.

Leverage scores are in some sense the importance scores for the ℓ_p-norm for the rows of X. The roots of ℓ_2 leverage scores go back to (Cook, 1977). They are useful for sampling algorithms when looking at target functions close to the p-th power of the ℓ_p-norm. One way of getting an exact bound for the ℓ_2-leverage scores which are the most commonly used leverage score is to look at the QR-decomposition of $X = QR$ and determine the squared ℓ_2-norm of the corresponding rows of Q. We will later prove this.

Well conditioned basis One can approximate the ℓ_p leverage scores using an orthonormal basis for the column space of X. Unfortunately this gives only an n^c-approximation for $p \neq 2$. We thus work with a generalization to so called *well-conditioned bases*. An (α, β, p)-well-conditioned basis V is a basis that preserves the norm of each vector well, as detailed in the following definition.

Definition 2.6 (Dasgupta et al. (2009))**.** *Let X be an $n \times m$ matrix of rank d, let $p \in [1, \infty)$, and let q be its dual norm, i.e., $q \in (1, \infty]$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Then an $n \times d$ matrix V is an* (α, β, p)-well-conditioned basis *for the column space of X if*
(1) $\|V\|_p := \left(\sum_{i\le n, j\le d} |V_{ij}|^p\right)^{1/p} \le \alpha$, and
(2) for all $z \in \mathbb{R}^d$, $\|z\|_q \le \beta\|Vz\|_p$.

We say that V is a p-well-conditioned basis *for the column space of X if α and β are $d^{O(1)}$, independent of m and n.*

A prominent example of a well-conditioned basis is the aforementioned orthonormal basis for ℓ_2, which can be obtained by QR-decomposition (or SVD) in $O(nd^2)$ time. Such a basis Q is $(\sqrt{d}, 1, 2)$-well-conditioned, since $\|Q\|_F = \sqrt{d}$ and $\|Qz\|_2 = \|z\|_2$ due to rotational invariance of the ℓ_2-norm. For general p there exist so called Auerbach bases (Auerbach, 1930) with $\alpha = d$ and $\beta = 1$ (for a proof see (Woodruff and Yasuda, 2023)), and approximations thereof can be computed in time $O(nd^5 \log n)$ via Löwner–John ellipsoids (Clarkson, 2005; Dasgupta et al., 2009).

We can bound the leverage scores in terms of the row-wise p-norms of such a basis.

Lemma 2.7. *Let V be an (α, β, p)*-well-conditioned basis *for the column space of X and let u_i be the ℓ_p-leverage score of row i. Then it holds for all $i \in [n]$ that $u_i \leq \beta^p \|v_i\|_p^p$. As a direct consequence we have $\sum_{i=1}^n u_i \leq \beta^p \|V\|_p^p \leq (\alpha\beta)^p$.*

Proof. We have by a change of basis

$$u_i = \sup_{z \in \mathbb{R}^d \setminus \{0\}} \frac{|(Xz)_i|^p}{\|Xz\|_p^p} = \sup_{z \in \mathbb{R}^d \setminus \{0\}} \frac{|(Vz)_i|^p}{\|Vz\|_p^p}.$$

Now assume that z attains the value $\sup_{z \in \mathbb{R}^d \setminus \{0\}} \frac{|(Vz)_i|^p}{\|Vz\|_p^p}$. Then we get by using Hölder's inequality and the properties of V that

$$u_i = \frac{|(Vz)_i|^p}{\|Vz\|_p^p} \leq \frac{\beta^p |(Vz)_i|^p}{\|z\|_q^p} \leq \frac{\beta^p \|v_i\|_p^p \|z\|_q^p}{\|z\|_q^p} = \beta^p \|v_i\|_p^p.$$

□

2.4 Probability distributions and common inequalities

Normal distribution The normal distribution (or Gaussian distribution) with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_{\geq 0}$ is given via a density function (see for instance Johnson et al. (1994))

$$\varphi(r) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{|r-\mu|^2}{2\sigma^2}\right)$$

for $r \in \mathbb{R}$. The cdf of the normal distribution then is given by

$$\Phi(r) = \int_{-\infty}^{r} \varphi(t)\, dt$$

For higher dimensions the normal distribution can be extended to the multivariate normal distribution with mean $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$ using the density function (see for instance Kotz et al. (2004))

$$\tilde{\varphi}(y) = \frac{\exp\left(-\frac{1}{2} \cdot (y-\mu)^T \Sigma^{-1} (y-\mu)\right)}{\sqrt{(2\pi)^d \det(\Sigma)}}$$

for $y \in \mathbb{R}^d$. For us the only important case will be when Σ is a diagonal matrix, i.e. all coordinates are independent of each other.

p-generalized normal distribution (see for instance Kleiber and Kotz (2003); Subbotin (1923)) The p-generalized normal distribution is a generalization of the normal distribution so that the exponent in the the density function behaves like a polynomial of degree $p \in \mathbb{R}_{\geq 0}$ rather than 2. The cdf of the p-generalized normal distribution (Kalke and Richter (2013))is given by

$$\Phi_p(r) = \frac{p^{1-1/p}}{2\Gamma(1/p)} \int_{-\infty}^{r} \exp(-|t|^p/p)\, dt$$

where $\Gamma(r) = \int_0^\infty t^{r-1}e^{-t}\, dt$ is the gamma function.

Bernoulli distribution (see for instance Johnson et al. (1994))

The Bernoulli distribution Bern is the distribution with

$$P(y = 1) = p \text{ and } P(y = 0) = 1 - p$$

for some parameter $p \in [0, 1]$.

Binomial distribution (see for instance Johnson et al. (1994))

The binomial distribution is the distribution we get from multiple added Bernoulli random variables with parameter p. We have that

$$P(y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k \in \{0, 1 \dots n\}$.

Concentration inequalities

Lemma 2.8 (Markov's inequality Pishro-Nik (2014))**.** *Let X be a positive random variable. Then for any $a \in \mathbb{R}_{>0}$ it holds that*

$$P(X \geq a) \leq \frac{\mathbb{E}(X)}{a}.$$

Lemma 2.9, 2.10 and 2.11 concern tail bounds for random scalar variables. Lemma 2.13 is helpful for proving lower bounds.

Lemma 2.9 (Chernoff bound Chernoff (1952))**.** *Let $X = \sum_{i=1}^n X_i$, where $X_i = 1$ with probability p_i and*

$X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = \mathbb{E}[X] = \sum_{i=1}^n p_i$. Then

$$\Pr[X \geq (1+\delta)\mu] \leq \exp(-\delta^2\mu/3), \ \forall 1 > \delta > 0; \tag{1}$$

$$\Pr[X \leq (1-\delta)\mu] \leq \exp(-\delta^2\mu/2), \ \forall 0 < \delta; \tag{2}$$

$$\Pr[X \geq (1+\delta)\mu] \leq \exp(-\delta\mu/3), \ \forall \delta \geq 1; \tag{3}$$

Lemma 2.10 (Hoeffding bound Hoeffding (1963))**.** *Let $X_1, \cdots, X_n$ denote n independent bounded random variables in $[a_i, b_i]$. Let $X = \sum_{i=1}^n X_i$. Then we have*

$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma 2.11 (Bernstein's inequality Bernstein (1924))**.** *Let $X_1, \cdots, X_n$ be independent zero-mean random variables. Suppose that $|X_i| \leq M$ almost surely, for all i. Then, for all positive t,*

$$\Pr\left[\sum_{i=1}^n X_i > t\right] \leq \exp\left(-\frac{t^2/2}{\sum_{j=1}^n \mathbb{E}[X_j^2] + Mt/3}\right).$$

Lemma 2.12 (Anti-concentration of the Gaussian distribution)**.** *Let $X \sim \mathcal{N}(0, \sigma^2)$, that is, the probability density function of X is given by $\varphi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}$. Then*

$$\Pr[|X| \leq t] \leq \frac{4}{5} \cdot \frac{t}{\sigma}.$$

Proof. It holds that

$$\Pr[|X| \leq t] = 2 \cdot \int_{x=0}^{t} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} dx \leq \frac{2}{\sqrt{2\pi\sigma^2}} \cdot t \leq \frac{4t}{5\sigma}.$$

□

Lemma 2.13 (Feller (1943))**.** *Let Z be a sum of independent random variables, each attaining values in $[0, 1]$, and let $\sigma = \sqrt{\mathrm{Var}(Z)} \geq 200$. Then for all $t \in [0, \frac{\sigma^2}{100}]$ we have*

$$\Pr[X \geq \mathbb{E}[X] + t] \geq c \cdot \exp(-t^2/(3\sigma^2))$$

where $c > 0$ is some fixed constant.

The following is useful when dealing with binomially distributed random variables.

Lemma 2.14. *Let y be a binomially distributed random variable with parameters n, p. Let $n' \in \mathbb{N}$. Then if*

$n' \geq pn$ we have that

$$P(|y - pn| > n') \leq 2\exp\left(-n'/3\right)$$

Else if $n' = \varepsilon pn$ we have that

$$P(|y - pn| > n') \leq 2\exp\left(-\varepsilon n'/3\right)$$

Proof. Note that $\mathbb{E}(y) = pn$. Using the Chernoff bound we get that if $n' \geq pn$

$$P(|y - pn| > n') \leq 2\exp(-(n'/np)np/3) \leq 2\exp(-n'/3).$$

If $n' = \varepsilon pn$ the Chernoff bound implies that

$$P(|y - pn| > n') \leq 2\exp(-\varepsilon^2 np/3) = 2\exp\left(-\varepsilon n'/3\right).$$

□

2.5 Data reduction methods

Coresets and Subspace embeddings We give a short introduction into data reduction methods. Our data reduction methods are used to construct coresets and skechtes which are smaller data sets with similar properties:

Definition 2.15. *(cf. Munteanu and Schwiegelshohn (2018)) A weighted ε-coreset $C = (X', w)$ for a function f is a matrix $X' \in \mathbb{R}^{k \times d}$ together with a weight vector $w \in \mathbb{R}^k_{>0}$ such that for all $\beta \in \mathbb{R}^d$ it holds that*

$$|f_w(X'\beta) - f(X\beta)| \leq \varepsilon \cdot f(X\beta).$$

Coresets have been a popular topic in research in the last years. They are useful in practice when it comes to big data and they are also interesting from a theoretic perspective giving us information how large sets can grow without redundant information for a given optimization function. For a detailed survey of coresets, see (Munteanu and Schwiegelshohn, 2018).

Using a coreset we can find an approximate solution to the problem of finding β minimizing $f(X\beta)$.

Corollary 2.16. *Let (X', w) be a weighted ε-coreset for f. Let $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} f_w(X'\beta)$. Then it holds that $f(X\tilde{\beta}) \leq (1 + 3\varepsilon) \min_{\beta \in \mathbb{R}^d} f(X\beta)$.*

Proof. Let $\beta^* \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} f(X\beta)$. Since (X', w) is a coreset we have by Definition 2.15 and using the

optimality of $\tilde{\beta}$ for the coreset that

$$\begin{aligned} f(X\tilde{\beta}) &\leq f_w(X'\tilde{\beta})/(1-\varepsilon) \leq f_w(X'\beta^*)/(1-\varepsilon) \\ &\leq f(X\beta^*)(1+\varepsilon)/(1-\varepsilon) \leq f(X\beta^*)(1+3\varepsilon). \end{aligned}$$

□

One of the first problems where coresets where considered is the squared loss. One way to construct coresets for the ℓ_2 norm are subspace embeddings:

Definition 2.17. *Given a matrix $X \in \mathbb{R}^{n\times d}$, an integer $k < n$ and a parameter $\varepsilon \in (0, 1/2]$ an ε-subspace embedding is a matrix $\Pi \in \mathbb{R}^{k\times n}$ such that for any $\beta \in \mathbb{R}^d$ it holds that*

$$(1-\varepsilon)\|X\beta\|_2^2 \leq \|\Pi X\beta\|_2^2 \leq (1+\varepsilon)\|X\beta\|_2^2.$$

Coresets have the advantage that they have a good approximation guarantee for any parameter vector β. If we just aim to approximate the optimal β given our data it suffices if the target value of the optimal β is preserved or well approximated and no parameter vector gets significantly better than before. Thus we will also consider weak (weighted) sketches which have only weaker guarantees than coresets but are sufficient for approximating the optimal β and can be constructed via random linear maps.

Definition 2.18. *Given a dataset (X, w), a subset $V \subset \mathbb{R}^d$, $a > 1$ and $\varepsilon, \delta > 0$, a weak weighted (V, a, ε)-sketch $C = (X', w')$ for f is a matrix $X' \in \mathbb{R}^{k\times d}$ of the form $X' = SX$ for some $S \in \mathbb{R}^{k\times n}$ together with a weight vector $w' \in \mathbb{R}^k_{>0}$ such that it holds simultaneously that: For all $\beta \in V$ we have*

$$f_{w'}(X'\beta) \geq (1-\varepsilon) f_w(X\beta)$$

and for $\beta^* \in V$ minimizing $f_w(X\beta)$ it holds that

$$f_{w'}(X'\beta^*) \leq a f_w(X\beta^*).$$

Further for any $\beta \in \mathbb{R}^d \setminus V$ it holds that

$$f_{w'}(X'\beta) > \min_{\beta\in V} f_{w'}(X'\beta).$$

For us the following two methods are of particular interest to construct coresets.

Random projections To get a subspace embedding one often multiplies X with a random matrix. We present some of these random matrices.

JL-Transformation The first random matrix we consider consists of $O(\ln(n))$ Gaussians. It can be used to approximate the ℓ_2-norm.

Lemma 2.19 (Johnson Lindenstrauss transformation)**.** *(Johnson and Lindenstrauss (1984)) Let $0 < \varepsilon, \delta \leq 1/2$. Let $G \in \mathbb{R}^{d \times k}$ be a matrix consisting of $k \in O(\varepsilon^{-2} \ln(n/\delta))$ random Gaussian vectors. Let $X \in \mathbb{R}^{n \times d}$. Then, with probability $1 - \delta$, for all $i \in [n]$ it holds that*

$$\|x_i G\|_2 = (1 \pm \varepsilon) \|x_i\|_2.$$

Linear Sketching The idea of (linear) sketching is to multiply our data matrix $X \in \mathbb{R}^{n \times d}$ with a sparse random matrix $S \in \mathbb{R}^{n' \times n}$. Linear sketching uses much sparser matrices than the JL transformation, therefore the number of rows is also larger than for JL transformations. This way it is faster but requires more space, or in other words there is a tradeoff between the size of the random matrix and thus the running time and the number n' of rows of the sketch X'. For a detailed work on sketching see for instance (Woodruff, 2014).

The Count sketch (Charikar et al. (2004))

Given a data set X the idea of the count sketch is to go over all datapoints x_i once and for each $i \in [n]$ to choose one of N buckets as well as a random sign in $\{-1, 1\}$, both uniformly at random. In the end or rather in the process of adding each element to its bucket all elements are multiplied with their random sign and added up to the remaining elements in the same bucket. This can also be depicted as multiplying with a matrix $\Pi \in \mathbb{R}^{k \times n}$ which has a single non-zero entry at each column which is either 1 or -1 each with probability 1/2. Note that this can be done via a random hashing map and thus can be done on a turnstile stream. The count sketch is faster than most other random matrices like the so called Rademacher Matrix (cf. Verbin and Zhang (2012)) or the so called Randomized Hadamard Transform (cf. Ailon and Liberty (2009)) but the dimension k in order to achieve its guarantees is also slightly larger.

The following calculations give us an idea why the count sketch and other methods using random signs preserve the ℓ_2-norm:

Consider any two vectors $u, v \in \mathbb{R}^d$ and let σ be a random sign. Then it holds that

$$\begin{aligned}
\mathbb{E}(\|u + \sigma v\|_2^2) &= \frac{1}{2}\langle u + v, u + v\rangle + \frac{1}{2}\langle u - v, u - v\rangle \\
&= \frac{1}{2}\left(\langle u + v, u + v\rangle + \langle u - v, u - v\rangle\right) \\
&= \frac{1}{2}\left(\langle u, u\rangle + 2\langle u, v\rangle + \langle v, v\rangle + \langle u, u\rangle - 2\langle u, v\rangle + \langle v, v\rangle\right) \\
&= \langle u, u\rangle + \langle v, v\rangle = \|u\|^2 + \|v\|^2.
\end{aligned}$$

The Count-Min sketch The Count-Min Sketch (Cormode and Muthukrishnan (2005)) works similarly to the count sketch but there is no random sign. This way directions are preserved but the ℓ_2 norm is no longer guaranteed to be preserved. The Count-Min Sketch was previously used for frequency analysis, as it

preserves the ℓ_0-norm, i.e the number of distinct elements. We will use it for problems where directions are important as those are also preserved to some extent when using the Count-Min Sketch which is not the case for the Count sketch.

Matrix representation of the Count and Count-Min sketch For both of the mentioned sketches there exists a matrix representation: The sketching matrix Π can be constructed as follows: First let $D \in \mathbb{R}^{n \times n}$ be the diagonal matrix with $D_{ii} = 1$ or $D_{ii} = -1$ each with probability $1/2$ for the count sketch and $D = I_n$ for the Count-Min sketch. Further let $h : [n] \to [n']$ be a random map where h hashes each entry of $[n]$ to one of n' buckets uniformly at random. Set $\Psi \in \mathbb{R}^{n' \times n}$ to be the matrix where $\Psi_{h(i)i} = 1$ and $\Psi_{ji} = 0$ if $j \neq h(i)$. Then the matrix representation is given by $\Pi = \Psi D$ (Clarkson and Woodruff, 2017).

Sketching for M-estimators Clarkson and Woodruff (2015) developed another sketching algorithm that works for a class of the so called M-estimators. The idea is that elements get mapped to different levels (with exponentially increasing probabilities) and then the Count sketch is applied to elements at the same level. In Section 3 we will discuss a modification of their algorithm in detail.

Subsampling Sampling algorithms pick some of the elements of the dataset to create a smaller dataset. We consider uniform sampling where each element is picked with the same probability as well as sampling algorithms where different elements can have different sampling probabilities, which is also known under the name of importance sampling. In the second case we also have weights to compensate for different probabilities. More precisely if each element i is picked with probability p_i the weight of x_i is $w_i = \frac{1}{p_i}$ if i is picked.

Uniform sampling Picking k elements uniformly at random is called uniform sampling. Uniform sampling works well if the dataset is of bounded complexity. More precisely if there are no elements that are much more important than most of the other elements then uniform sampling works well.

Leverage score sampling For leverage score sampling the sampling probabilities are proportional (or close to proportional) to the leverage scores. ℓ_p-leverage score sampling preserves the ℓ_p-norm up to a small error. Note that it is also possible to use the square root of the ℓ_2 leverage scores to get a coreset that preserves the ℓ_1-norm. For more details see (Munteanu et al., 2018).

Sensitivity sampling framework Another more general sampling approach is sensitivity sampling. We first give the definition of sensitivities and the VC-dimension and afterwards we state one of the main results from sensitivity sampling.

Definition 2.20. *(Langberg and Schulman (2010)) Consider a family of functions $\mathcal{F} = \{g_1, \ldots, g_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and weighted by $w \in \mathbb{R}^n_{>0}$. The sensitivity of g_i for $f_w(x) = \sum_{i=1}^n w_i g_i(x)$ is*

$$\varsigma_i = \sup \frac{w_i g_i(x)}{f_w(x)} \tag{4}$$

where sup *is over all $x \in \mathbb{R}^d$ with $f_w(x) > 0$. If this set is empty then $\varsigma_i = 0$. The total sensitivity is $\mathfrak{S} = \sum_{i=1}^n \varsigma_i$.*

The sensitivity of a point bounds the maximal relative contribution to the target function the point can have. Computing the sensitivities is often intractable and necessitates approximating the original optimization problem close to optimality. However, this is the problem that we want to solve, see Braverman et al. (2021). Fortunately, for our applications it suffices to obtain a reasonable upper bound for the sensitivities.

Definition 2.21. *A range space is a pair $\mathfrak{R} = (\mathcal{F}, \text{ranges})$ where $\mathcal{F}$ is a set and* ranges *is a family of subsets of $\mathcal{F}$. The VC dimension $\Delta(\mathfrak{R})$ of $\mathfrak{R}$ is the size $|G|$ of the largest subset $G \subseteq \mathcal{F}$ such that G is shattered by* ranges*, i.e.,*

$$|\{G \cap R \mid R \in \text{ranges}\}| = 2^{|G|}.$$

Definition 2.22. *Let $\mathcal{F}$ be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. For every $x \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$, let*

$$\text{range}_{\mathcal{F}}(x, r) = \{f \in \mathcal{F} \mid f(x) \geq r\},$$

and

$$\text{ranges}(\mathcal{F}) = \{\text{range}_{\mathcal{F}}(x, r) \mid x \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\},$$

and

$$\mathfrak{R}_{\mathcal{F}} = (\mathcal{F}, \text{ranges}(\mathcal{F}))$$

be the range space induced by $\mathcal{F}$.

The VC-dimension can be thought of something similar to the dimension of our problem. For example the VC-dimension of the set of hyperplane classifiers in $\mathbb{R}^d$ is $d + 1$ (Kearns and Vazirani, 1994). The sensitivity scores were combined with a theory on the VC-dimension of range spaces in (Feldman and Langberg, 2011; Braverman et al., 2021). We use a more recent version of Feldman et al. (2020).

Proposition 2.23. *(Feldman et al., 2020) Consider a family of functions $\mathcal{F} = \{f_1, \ldots, f_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and a vector of weights $w \in \mathbb{R}^n_{>0}$. Let $\varepsilon, \delta \in (0, 1/2)$. Let $s_i \geq \varsigma_i$. Let $S = \sum_{i=1}^n s_i \geq \mathfrak{S}$. Given s_i one can compute in time $O(|\mathcal{F}|)$ a set $R \subset \mathcal{F}$ of*

$$O\left(\frac{S}{\varepsilon^2}\left(\Delta \ln S + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

weighted functions such that with probability $1-\delta$*, we have for all* $x \in \mathbb{R}^d$ *simultaneously*

$$\left| \sum_{f_i \in \mathcal{F}} w_i f_i(x) - \sum_{f_i \in R} u_i f_i(x) \right| \leq \varepsilon \sum_{f_i \in \mathcal{F}} w_i f_i(x),$$

where each element of R *is sampled i.i.d. with probability* $p_j = \frac{s_j}{S}$ *from* $\mathcal{F}$*,* $u_i = \frac{Sw_j}{s_j|R|}$ *denotes the weight of a function* $f_i \in R$ *that corresponds to* $f_j \in \mathcal{F}$*, and where* Δ *is an upper bound on the VC dimension of the range space* $\mathfrak{R}_{\mathcal{F}^*}$ *induced by* $\mathcal{F}^*$ *obtained by defining* $\mathcal{F}^*$ *to be the set of functions* $f_j \in \mathcal{F}$*, where each function is scaled by* $\frac{Sw_j}{s_j|R|}$*.*

Reservoir samplers Assume we want to sample k elements of $[n]$ such that each element $i \in [n]$ is sampled with probability $p_i = \max\{k \cdot \frac{s_i}{S}\}$ where $S = \sum_{i=1}^{n} s_i$. However we want to do this in the online setting or in a single pass and we can only compute s_i and do not know S in advance. Here one can use a weighted reservoir sampler Chao (1982). The idea of the reservoir sampler is to sample the first k elements and then each further to sample each element i sample i with probability $\frac{s_i}{S_i}$ where $S_i = \sum_{j=1}^{i} S_i$ replacing a random other element.

2.6 Linear regression, logistic regression and p-probit regression

Linear models/regression Groß (2003)

Let $X \in \mathbb{R}^{n \times d}$ be some data set together with observations $y \in \mathbb{R}^n$. A linear model assumes that Y depends on X in the following way:

$$Y = X\beta + \xi$$

where ξ is noise variable whose entries usually follow a normal distribution and $\beta \in \mathbb{R}^d$. More precisely we have that $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$ for some $\sigma \in \mathbb{R}_{\geq 0}$. Then we have that Y is distributed via the following density function:

$$f(Y) = \tilde{\varphi}(X\beta - Y).$$

Generalized linear models/regression McCullagh and Nelder (1989)

Again we have some data set $X \in \mathbb{R}^{n \times d}$ with observations $Y \in \mathbb{R}^n$. A general linear model assumes that

$$h(\mathbb{E}(Y)) = X\beta$$

for some link function $h : \mathbb{R} \to \mathbb{R}$. If h is the identity then we have a linear model. However generalized linear models can also be used to model functions with more complex behaviors or those which are taking only values in a limited range like Bernoulli or Binomial variables.

Logistic regression For logistic regression we have that the link function is given by $h(r) = \ln(\frac{r}{1+r})$ McCullagh and Nelder (1989); Hilbe (2009). The logistic model is useful to learn to predict the probability of an event to happen based on independent observations. For instance what is the likelihood of someone passing a test given some features such as time spent learning etc. The assumption then is that $Y_i \sim \text{Bern}(r_i)$ where $r_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)}$ for some β. We are using labels $y_i \in \{-1, 1\}$ rather than $y_i \in \{0, 1\}$ to simplify notations. Then the likelihood of any $\beta \in \mathbb{R}^d$ for our dataset (X, y) is given by

$$\prod_{i=1}^{n} \frac{\exp(y_i x_i \beta)}{1 + \exp(y_i x_i \beta)}.$$

Taking the negative logarithm we get that the negative log likelihood, the logistic loss, is given by

$$\begin{aligned}
\mathcal{L}(\beta|X, y) &= -\ln\left(\prod_{i=1}^{n} \frac{\exp(y_i x_i \beta)}{1 + \exp(y_i x_i \beta)}\right) \\
&= \sum_{i=1}^{n} -\ln\left(\frac{\exp(y_i x_i \beta)}{1 + \exp(y_i x_i \beta)}\right) \\
&= \sum_{i=1}^{n} \ln\left(1 + \exp(-y_i x_i \beta)\right).
\end{aligned}$$

Variance regularized logistic regression Variance-based regularization was proposed in (Maurer and Pontil, 2009; Duchi and Namkoong, 2019; Yan et al., 2020) to decrease the generalization error. We view our data set as n realizations of a random variable (Z, Y), where each (x_i, y_i) is drawn i.i.d. from an unknown distribution $\mathcal{D}$. Then the expected value of the negative log-likelihood (on the empirical sample) for any fixed β equals $\mathbb{E}(\ell(-YZ\beta)) = \frac{1}{n}\mathcal{L}(\beta|X, Y)$. The variance is given by

$$\text{Var}(\ell(-YZ\beta)) = \mathbb{E}(\ell(-Yz\beta)^2) - \mathbb{E}(\ell(-YZ\beta))^2 = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \ell(-y_i x_i \beta)^2 - \left(\tfrac{1}{n}\sum_{i=1}^{n} \ell(-y_i x_i \beta)\right)^2$$

We also introduce a regularization parameter $\lambda \in \mathbb{R}_{\geq 0}$. Then our objective is to minimize

$$\mathbb{E}(\ell(-YZ\beta)) + \frac{\lambda}{2}\text{Var}(\ell(-YZ\beta)).$$

***p*-probit regression** In the probit model we have that $h(r) = \Phi_2^{-1}(r)$(McCullagh and Nelder, 1989). We generalized the probit model to the p-probit regression where we have that $h(r) = \Phi_p^{-1}(r)$ for $p \in [1, \infty)$ Munteanu et al. (2022a) and applications are similar to logistic regression. In fact 1-probit regression has similar tail behavior as logistic regression but differs in the region close to 0.

The likelihood of any $\beta \in \mathbb{R}^d$ for our dataset (X, y) is given by

$$\prod_{i=1}^{n} \Phi_p(y_i x_i \beta).$$

Taking the negative logarithm we get that the negative log likelihood, the p-probit loss, is given by

$$-\ln\left(\prod_{i=1}^{n}\Phi_p(y_i x_i\beta)\right)=\sum_{i=1}^{n}-\ln\left(\Phi_p(y_i x_i\beta)\right).$$

Labels Note that for both problems, logistic regression as well as p-probit regression the labels are always appearing as a scalar of the datapoints themselves. Thus it is convenient to substitute $x_i'=-y_i x_i$. This way we have that the negative log likelihood is given by $\sum_{i=1}^{n} g(x_i')$ where either $g(r)=\ln(1+e^r)$ or $g=-\ln(\Phi_p(-r))$.

μ-complexity In contrast to the ℓ_p loss ($\sum_{i=1}^{n}\|x_i\beta-y\|_p^p$) both, the logistic and the p-probit loss, are asymmetric functions. As a consequence some of the well known data reduction methods like the count sketch do not work for them. Further small coresets do not exist in general (Munteanu et al., 2018), we will use the following parameter which has been in introduced in (Munteanu et al., 2018) and generalized in (Munteanu et al., 2022a).

Definition 2.24. *Let $X\in\mathbb{R}^{n\times d}$ be any matrix and let $p\in[1,\infty)$. We define*

$$\mu_p(X)=\sup_{\beta\in\mathbb{R}^d\setminus\{0\}}\frac{\sum_{x_i\beta>0}|x_i\beta|^p}{\sum_{x_i\beta<0}|x_i\beta|^p}.$$

We say that X is μ-complex if $\max\{\mu_1(X),\mu_2(X)\}\leq\mu$ if considering (variance regularized) logistic regression or if $\mu_p(X)\leq\mu$ if considering p-probit regression.

2.7 Artificial neural networks

Artificial neural networks are another popular method for predictions. For a more detailed description we refer to (Blum et al., 2020). An artificial network consists of several layers each consisting of one or multiple nodes. Two adjacent layers are connected via edges, each equipped with a weight. In addition each layer except the first layer is equipped with an activation function. The first layer is called input layer and the last layer output layer.

More formally an r layer neural network is given by a vector of matrices $(W_1,\dots W_r)$ and a vector of activation function $(f_1,\dots,f_r)$ where $W_i\in\mathbb{R}^{n_i\times n_{i-1}}$ with n_i being the number of nodes in layer i and $f_i:\mathbb{R}\to\mathbb{R}$. Given an input $x\in\mathbb{R}^{n_0}$ we iteratively define the value of the nodes. The values of the nodes in the input layer are equal to the coordinates of x_i. The values of the nodes of any other layer $i>0$ are computed by multiplying the vector we get by the previous layer $i-1$ with W_i and applying the activation function to all coordinates.

Even though neural networks are known to perform well in practice little is known in theory due to their complexity. As our goal is to analyze two layer networks with the ReLU activation function for binary classification we will in the following restrict to those.

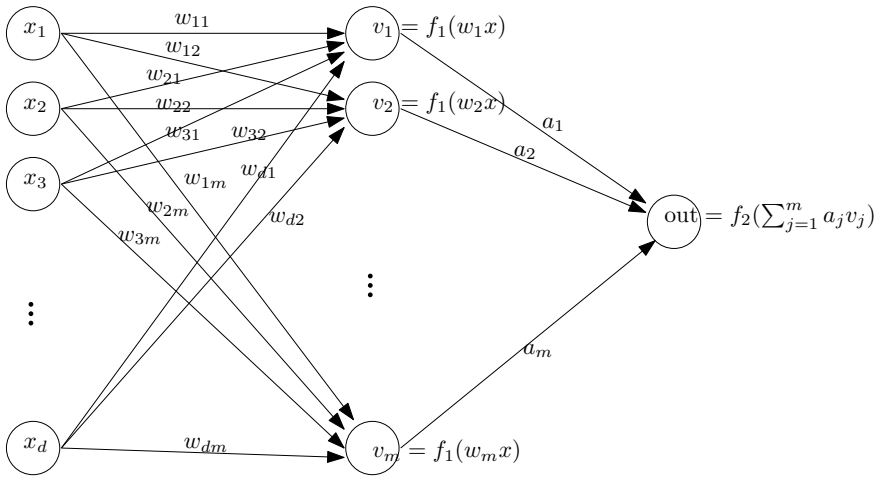

Figure 1: A two layer network with d input nodes and one output node. The rows of the matrix W_1 are given by w_i and W_2 is the vector with the entries a_j.

2.7.1 Two-layer ReLU networks

As the name already suggests a two-layer ReLU network consists of two layers and the input layer. The activation function of the first layer is the ReLU function $\phi : \mathbb{R} \to \mathbb{R}$ with $\phi(r) = \max\{r, 0\}$ and the activation function of the output layer is the identity function. Our inputs are given as d dimensional vectors. We use m to denote the number of nodes in the inner layer, also called the width of the network and we just have a single node at the output layer. We denote the the first sets of weights between the input layer and the first layer by $W \in \mathbb{R}^{n \times d}$ and the weight between first and second layer by $a \in \mathbb{R}^m$.

We follow a standard problem formulation Du et al. (2019c); Song and Yang (2019); Ji and Telgarsky (2020). The output function of our network is given by

$$f(x, W, a) = \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \phi\left(\langle w_s, x \rangle\right), \tag{5}$$

where $x \in \mathbb{R}^d$ is an input point, $w_1, \ldots, w_m \in \mathbb{R}^d$ are weight vectors in the first (hidden) layer, i.e. the rows of W and $a_1, \ldots, a_m \in \{-1, +1\}$ are weights in the second layer.

Loss function In order to train a neural network we need a loss function. Our general goal is to reduce the loss of the network to below an arbirtrarily small $\varepsilon > 0$. The loss of the networks also gives an upper bound of the number of misclassified points as any misclassified point has a bounded minimum contribution to the loss function. In this work, we mainly focus the binary *cross-entropy (logistic) loss* which is arguably the most well-studied for binary classification. We note that for regression squared loss is the most considered loss function which is also considered in (Munteanu et al., 2022b).

As before we are given a set of n input data points and corresponding labels, denoted by

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^d \times \{-1, 1\}.$$

As in Du et al. (2019c); Song and Yang (2019); Ji and Telgarsky (2020), we make a standard normalization assumption. The labels are restricted to $y_i \in \{-1, +1\}$. For simplicity, we assume that $\|x_i\|_2 = 1$[1], $\forall i \in [n]$. We also define the output function on input x_i to be $f_i(W) = f(x_i, W, a)$.

We consider the objective function R:

$$R(W) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i f_i(W))$$

where the individual logistic loss is defined as $\ell(r) = \ln(1 + \exp(-r))$.

For logistic loss, we can compute the gradient of R in terms of $w_r \in \mathbb{R}^d$

$$\frac{\partial R(W)}{\partial w_r} = \sum_{i=1}^{n} \frac{-\exp(-y_i f(W, x_i, a))}{1 + \exp(-y_i f(W, x_i, a))} y_i a_r x_i \mathbf{1}_{w_r^\top x_i \geq 0} \tag{6}$$

We apply gradient descent to optimize the weight matrix W with the following standard update rule,

$$W_{t+1} = W_t - \eta \frac{\partial R(W_t)}{\partial W_t}, \tag{7}$$

where $0 < \eta \leq 1$ determines the step size.

The neural tangent kernel (NTK) The NTK was introduced by Jacot et al. (2018). The NTK is a useful tool for analyzing convergence of the training of neural networks. Given a weight matrix W and a sign vector a the finite NTK is defined by $K_W : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with

$$K_W(x, x') = \cdot \sum_{j=1}^{m} \left\langle \frac{\partial f(x, W, a)}{\partial w_j}, \frac{\partial f(x', W, a)}{\partial w_j} \right\rangle.$$

Given a Gaussian random initialization for a sufficiently large width m it can be shown that the NTK is almost constant during the gradient descent, i.e. it stay close to the infinite NTK which is defined by

$$K(x, x') = \mathbb{E}\left(\sum_{j=1}^{m} \left\langle \frac{\partial f(x, W, a)}{\partial w_j}, \frac{\partial f(x', W, a)}{\partial w_j} \right\rangle \right).$$

[1]We adopt the assumption for a concise presentation, but we note it can be resolved by weaker constant bounds $0 < \text{lb} \leq \|x_i\| \leq \text{ub}$, introducing a constant ub/lb factor, cf. Du et al. (2019c), or otherwise the data can be rescaled and padded with an additional coordinate to ensure $\|x_i\| = 1$, cf. Allen-Zhu et al. (2019a).

In our case the infinite NTK is given by

$$K(x, x') = \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}[x^\top x' \mathbf{1}[\langle x, w\rangle > 0, \langle x', w\rangle > 0]].$$

In previous works it has been shown that if m is large enough, then the finite NTK stays close to the infinite NTK during the gradient descent. We will show how a clever initialization allows us to achieve the same with a much smaller width, allowing us in some sense a notable dimension reduction of the finite NTK.

Separation margin The separation margin γ was introduced by Ji and Telgarsky (2020). It is used as a parameter to get upper bounds on the width necessary to guarantee that the gradient descent algorithm converges to a network of arbitrarily small error.

Definition 2.25. *Given a data set $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ and a map $\bar{v} \in \mathcal{F}_B$ we set*

$$\gamma_{\bar{v}} = \gamma_{\bar{v}}(X, Y) := \min_{i\in[n]} y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z).$$

We say that $\bar{v}$ is optimal if $\gamma_{\bar{v}} = \gamma(X, Y) := \max_{\bar{v}'\in\mathcal{F}_B} \gamma_{\bar{v}'}$.

The authors of (Ji and Telgarsky, 2020) were able to show that by introducing γ it is possible to get an upper bound on the width of the network that is polynomial in γ but polylogarithmic in n, ε and δ, i.e. the width has no polynomial factor in those in it.

2.8 Convex optimization

Most of the functions considered in this manuscript are convex: A function $f : X \to \mathbb{R}$ is convex if for all $x, x' \in X$ and $t \in [0, 1]$ it holds that $f(tx + (1 - t)x') \leq tf(x) + (1 - t)f(x')$. A local minimum of a convex function is also a global minimum thus for convex functions it suffices to find a local minimum. For a book on convex optimization we refer to (Nesterov, 2003) and (Bubeck, 2015).

Gradient descent Given any differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ (in most cases it suffices if the function is differentiable almost everywhere) and a step size $\eta_i \in (0, \infty)$ for each $i \in \mathbb{N}$ the gradient descent is a method to determine a local minimum of f. Let ∇f be the gradient of f. Given a starting point $x_0 \in \mathbb{R}^d$ the i-th output of the gradient descent is given via

$$x_{i+1} = x_i - \eta_i \nabla f.$$

If the step size is not too large the gradient descent converges to a local minimum under certain conditions. Note that if f is convex then any local minimum is a global minimum.

Gradient of the logistic loss function Recall that the logistic loss is given by $f(X\beta) = \sum_{i=1}^{n} \ln(1 + \exp(x_i\beta))$. First note that the derivative of $\ell(r) = \ln(1 + \exp(r))$ is given by

$$\frac{d}{dr}\ell(r) = \frac{\exp(r)}{1+\exp(r)} = \frac{1}{1+\exp(-r)}.$$

Thus the gradient of $f(X\beta)$ is given by

$$\begin{aligned}\nabla f &= \frac{\partial}{\partial\beta} f(X\beta) \\ &= \sum_{i=1}^{n} \ell'(x_i\beta) \cdot x_i \\ &= \sum_{i=1}^{n} \frac{1}{1+\exp(-x_i\beta)} \cdot x_i.\end{aligned}$$

Newton-Raphson method The optimization of f can also be done by applying the Newton-Raphson method (Bubeck, 2015), an iterative procedure that starts at an initial guess $\beta^{(0)}$ and successively applies the following update rule:

$$\beta^{(t)} = \beta^{(t-1)} - \left(\frac{\partial^2 f(\beta^{(t-1)})}{\partial\beta\partial\beta^T}\right)^{-1} \cdot \frac{\partial f(\beta^{(t-1)})}{\partial\beta},$$

where $\left(\frac{\partial^2 f(\beta^{(t-1)})}{\partial\beta\partial\beta^T}\right)^{-1}$ refers to the inverse of the Hessian matrix of f, evaluated at $\beta^{(t-1)}$, and $\frac{\partial f(\beta^{(t-1)})}{\partial\beta}$ refers to the gradient of f, evaluated at $\beta^{(t-1)}$. The idea behind this procedure is, broadly speaking, to approximate f locally around $\beta^{(t)}$ by its second degree Taylor-polynomial and then analytically find the minimum of this polynomial. The minimum of this local polynomial approximation of f is then used iteratively as a basis for the next step of the Newton-Raphson algorithm.

2.8.1 Gradient and Hessian Matrix for p-probit regression

In the following we derive the gradient and the Hessian matrix of f. Since f is a sum of the function g evaluated at different points, it makes sense to first determine the derivative of g. To this end $\varphi_p(r)$ is the density function of the (standardized) p-generalized normal distribution function, cf. (Dytso et al., 2018; Kalke and Richter, 2013):

$$\varphi_p(r) = \frac{p^{1-1/p}}{2\Gamma(1/p)} \exp(-|r|^p/p).$$

We proceed by using the chain rule as follows:

$$\begin{aligned}
\frac{d}{dr}g(r) = \frac{d}{dr} - \ln\left(\Phi_p(-r)\right) &= \frac{d}{dr}\ln\left(\frac{1}{1-\Phi_p(r)}\right) \\
&= (1-\Phi_p(r)) \cdot \frac{d}{dr}\left(\frac{1}{1-\Phi_p(r)}\right) \\
&= (1-\Phi_p(r)) \cdot \frac{(-1)}{(1-\Phi_p(r))^2} \cdot \frac{d}{dr}(1-\Phi_p(r)) \\
&= \frac{(-1)}{1-\Phi_p(r)} \cdot (-1) \cdot \varphi_p(r) \\
&= \frac{\varphi_p(r)}{1-\Phi_p(r)},
\end{aligned}$$

We can use this result to calculate the gradient of f:

$$\begin{aligned}
\frac{\partial}{\partial\beta}f_w(X\beta) &= \frac{\partial}{\partial\beta}\sum_{i=1}^{n} w_i g(x_i\beta) \\
&= \sum_{i=1}^{n} w_i x_i g'(x_i\beta) \\
&= \sum_{i=1}^{n} w_i x_i \frac{\varphi_p(x_i\beta)}{1-\Phi_p(x_i\beta)}
\end{aligned}$$

Next, we need to determine the Hessian matrix of f. To this end, we again start by finding the second derivative of g, this time using the quotient rule.

$$\begin{aligned}
\frac{d^2}{dr^2}g(r) &= \frac{d}{dr}\frac{\varphi_p(r)}{1-\Phi_p(r)} \\
&= \frac{\varphi_p'(r)(1-\Phi_p(r)) - \varphi_p(r)\cdot(-1)\cdot\varphi_p(r)}{(1-\Phi_p(r))^2} \\
&= \frac{(-1)\cdot\operatorname{sgn}(r)\cdot|r|^{p-1}\cdot\varphi_p(r)(1-\Phi_p(r)) - \varphi_p(r)\cdot(-1)\cdot\varphi_p(r)}{(1-\Phi_p(r))^2} \\
&= \frac{[\varphi_p(r)]^2 - \operatorname{sgn}(r)|r|^{p-1}\cdot\varphi_p(r)\cdot(1-\Phi_p(r))}{(1-\Phi_p(r))^2} \\
&= \left(\frac{\varphi_p(r)}{1-\Phi_p(r)}\right)^2 - \operatorname{sgn}(r)|r|^{p-1}\cdot\frac{\varphi_p(r)}{1-\Phi_p(r)} \\
&= \frac{\varphi_p(r)}{1-\Phi_p(r)}\left(\frac{\varphi_p(r)}{1-\Phi_p(r)} - \operatorname{sgn}(r)|r|^{p-1}\right) \\
&= g'(r)\cdot(g'(r) - \operatorname{sgn}(r)|r|^{p-1})
\end{aligned}$$

We can now use this result to find the Hessian matrix of f. Recall that x_i are *row vectors* in our paper and

thus each $x_i^T x_i$ is a $d \times d$-matrix.

$$\begin{aligned}\frac{\partial^2}{\partial\beta\partial\beta^T} f_w(X\beta) &= \sum_{i=1}^{n} \frac{\partial^2}{\partial\beta\partial\beta^T} w_i g(x_i\beta) \\ &= \sum_{i=1}^{n} w_i x_i^T x_i g'(x_i\beta)(g'(x_i\beta) - \operatorname{sgn}(x_i\beta)|x_i\beta|^{p-1}) \\ &= \sum_{i=1}^{n} w_i x_i^T x_i \frac{\varphi_p(x_i\beta)}{1-\Phi_p(x_i\beta)} \left(\frac{\varphi_p(x_i\beta)}{1-\Phi_p(x_i\beta)} - \operatorname{sgn}(x_i\beta)|x_i\beta|^{p-1} \right).\end{aligned}$$

It can be shown, that $f_w(X\beta)$ is a convex function of β, and that the Newton-Raphson algorithm converges to a global optimum when applied to a convex function (Bubeck, 2015). The optimization procedure thus converges to the maximum likelihood estimate $\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} f_w(X\beta)$ provided it exists.

Jensens inequality

Lemma 2.26 (Jensens inequality)**.** *Jensen (1906) Let $f : \mathbb{R} \to \mathbb{R}$ be any convex function. Further let $x_1, \dots x_n \in \mathbb{R}$ be arbitrary points and $\lambda_1, \dots, \lambda_n \in [0, 1]$ with $\sum_{i=1}^n \lambda_i = 1$. Then it holds that*

$$\sum_{i=1}^{n} \lambda_i f(x_i) \geq f(\sum_{i=1}^{n} \lambda_i x_i)$$

2.9 Related work

Coresets Coresets as a data reduction method have been studied for logistic regression before (Huggins et al. (2016); Tolochinsky et al. (2022); Munteanu et al. (2018); Tukan et al. (2020); Samadian et al. (2020)). Those results often rely on regularization as a means to obtain small coresets. This changes the sampling distribution such that they do not generally apply to the unregulated setting that we study. The above coreset constructions usually require random access to the data and are thus not directly suitable for streaming computations. Even where row-order processing is permissible, at least two passes are required, one for calculating or approximating the probabilities and another for subsampling and collecting the data, since the importance sampling distributions usually depend on the data. A widely cited general scheme for making static (or multi-pass) constructions streamable in one pass is the Merge & Reduce framework (Bentley and Saxe (1980)). However, this comes at the cost of additional polylogarithmic overhead in the space requirements and also in the update time. The latter is a severe limitation when it comes to high velocity streams that occur for instance in large scale physical experiments such as the large hadron collider, where up to 100 GB/s need to be processed and data rates are anticipated to grow quickly to several TB/s in the near future Rohr (2018). While the amortized insertion time of Merge & Reduce is constant for some problems, in the worst case $\Theta(\log n)$ repeated coreset constructions are necessary for the standard construction to propagate through the tree structure; see e.g. Feldman et al. (2020). This poses a prohibitive bottleneck in high velocity applications. Any data that passes and cannot be processed in real time will be lost forever.

Another limitation of coresets and the Merge & Reduce scheme is that they work only in insertion streams, where the data is presented row-by-row. However it is unclear how to construct coresets when the data comes in column-wise order, e.g., when we first obtain the incomes of all individuals, then receive their heights and weights, etc. A similar setting arises when the data is distributed *vertically* on numerous sites Stolpe et al. (2013). Sensor networks are another example where each sensor is recording only a single or a small subset of features (columns), e.g., each at one of many different production stages in a factory. Also the usual form of storing data in a table either row- or column-wise is not appropriate or efficient for extremely massive databases. The data is rather stored as a sequence of $(key, value)$ pairs in an arbitrary order in big unstructured databases (Gessert et al. (2017); Siddiqa et al. (2017)).

The only work that can be simulated in a turnstile stream to tackle the extreme settings described above, is arguably Samadian et al. (2020) via uniform subsampling. Their coreset size is roughly $\Theta(d\sqrt{n})$ and works only when the problem is regularized very strongly such that the loss function is within constant factors to the regularizer, and thus widely independent of the input data. Consequently, the contribution of each point becomes roughly equal and thus makes uniform sampling work. However, those arguments do not work for unconstrained logistic regression, where each single point can dominate the cost and thus no sublinear compression below $\Omega(n)$ is possible in the worst case, as was shown in (Munteanu et al. (2018)). To cope with this situation, the authors of Munteanu et al. (2018) introduced a complexity parameter μ that is related to the statistical modeling of logistic regression, and is a useful measure for capturing the complexity of compressing the dataset A for logistic regression. They developed a coreset construction of size $\tilde{O}(\mu d^{3/2}\sqrt{n})$. The coreset size was reduced to $\text{poly}(\mu d \log n)$ but only at the cost of even more row-order passes to compute repeatedly a coreset from a coreset, $O(\log \log n)$ times. Although calculating their sampling distribution can be simulated in a row-order stream, the aforementioned limitation to two passes is an unsolved open problem, which will be dealt with in this manuscript.

Data oblivious sketching Data oblivious sketches have been developed for many problems in computer science, see (Phillips, 2017) for an extensive survey. The seminal work of Sarlós (2006) opened up the toolbox of sketching for numerical linear algebra and machine learning problems, such as linear regression and low rank approximation, cf. (Woodruff, 2014). We note that oblivious sketching is very important to obtain data stream algorithms in the turnstile model (Muthukrishnan, 2005) and there is evidence that linear sketches are optimal for such algorithms under certain conditions (Li et al., 2014; Ai et al., 2016). The classic works on ℓ_2 regression have been generalized to other ℓ_p norms (Sohler and Woodruff, 2011; Woodruff and Zhang, 2013) by combining sketching as a fast but inaccurate preconditioner and subsequent sampling to achieve the desired $(1+\varepsilon)$-approximation bounds. Those works have been generalized further to so-called M-estimators, i.e., Huber (Clarkson and Woodruff, 2015) or Tukey regression loss (Clarkson et al., 2019), that share nice properties such as symmetry and homogeneity leveraged in previous works on ℓ_p norms. In a recent work Mai et al. (2023) present oblivious sketching algorithms for sparse regression and show that subsampling does

not work for sparse regression. They present upper and lower bounds for the sketching dimension various loss functions such as sparse ℓ_p-regression, sparse ReLU and hinge-like functions and LASSO regression with ℓ_1 regularization.

ℓ_1 regression Specifically for ℓ_1, the first sketching algorithms used random variables drawn from 1-stable (Cauchy) distributions to estimate the norm (Indyk, 2006). It is possible to get concentration and a $(1 \pm \varepsilon)$-approximation in near-linear space by using a median estimator. However, in a regression setting this estimator leads to a non-convex optimization problem in the sketch space. Since we want to preserve convexity to facilitate efficient optimization in the sketch space, we focus on sketches that work with an ℓ_1 estimator for solving the ℓ_1 regression problem in the sketch space in order to obtain a constant approximation for the original ℓ_1 problem. With this restriction, it is possible to obtain a contraction bound with high probability so as to union bound over a net, but it is impossible to obtain high probability for the dilation. Indeed, *subspace embeddings* for the ℓ_1 norm have $\tilde{\Theta}(d)$ dilation (Woodruff and Zhang, 2013; Li et al., 2021b; Wang and Woodruff, 2022) hiding polylogarithmic terms. Further, $1 + \varepsilon$ dilation is only known to be possible when mapping to $\exp(O(1/\varepsilon))$ dimensions (Brinkman and Charikar, 2005), even for single vectors as in (Indyk, 2006). We thus focus on obtaining an $O(1)$ approximation in this manuscript. Previous work had either larger $O(\log(d))$ distortion (by an argument in the proof of Lemma 7 of Sohler and Woodruff (2011), see Problem 1 in (Woodruff, 2021)) or larger $\text{poly}(d)$ factors (Indyk, 2006; Sohler and Woodruff, 2011). We note that there is a $(1 + \varepsilon)$-approximation algorithm by Sohler and Woodruff (2011) that works in a turnstile data stream by running two sketches in parallel, one for preconditioning and another that performs ℓ_1-row-sampling from the *sketch* (Andoni et al., 2009). However, it has a worse $\text{poly}(d \log(n)/\varepsilon)$ update time and sketching dimension, see Theorem 13 of (Sohler and Woodruff, 2011). An advantage of our sketch is that it uses only random $\{0, 1\}$-entries, which have better computational and implicit storage properties (Alon et al., 1986, 1999; Rusu and Dobra, 2007). More importantly, our approach works for both, ℓ_1 *and* logistic regression simultaneously, where for the latter no near-linear sketching dimension was known to be possible due to the fact that sketches for ℓ_1 cannot track the sign information of coordinates, which is crucial for preserving any multiplicative error for the asymmetric logistic regression loss function.

Generalized linear models (GLMs) It is important to extend the works on linear regression to more sophisticated and expressive statistical learning problems, such as *generalized linear models* (McCullagh and Nelder, 1989). Unfortunately, taking this step led to impossibility results. Namely, approximating the regression problems on a succinct sketch for strictly monotonic functions such as logistic loss (Munteanu et al., 2018) or heavily imbalanced asymmetric functions such as Poisson regression loss (Molina et al., 2018) allows one to design a low-communication protocol for the INDEXING problem that contradicts its $\Omega(n)$ bit randomized one-way communication complexity (Kremer et al., 1999). This implies an $\tilde{\Omega}(n)$ sketching dimension for these problems. To circumvent this worst-case limitation for logistic regression, Munteanu et al. (2018) introduced a natural data dependent parameter μ that can be used to bound the complexity of

compressing data for logistic regression and related probit regression (Munteanu et al., 2022a). This was used for developing our very first oblivious sketching algorithm for GLMs, specifically for logistic regression (Munteanu et al., 2021), with a large polylogarithmic number of rows for mild μ-complex data. Munteanu et al. (2023) improved this sketching algorithm by giving, the *only* near-linear sketching dimension in d and μ for logistic regression. This sketching dimension was previously only known to be possible by Lewis weight sampling Mai et al. (2021) where in exchange the dependence on μ was quadratic, and crucially their sketch is not oblivious so cannot be implemented in a turnstile data stream, with positive and negative updates to the entries of the input point set. For lower bounds, an $\Omega(d)$ dependence is immediate since mapping to fewer than d dimensions contracts non-zero vectors in the null-space of the sketching matrix to zero. An $\Omega(\mu)$ lower bound is immediate from the streaming lower bound of Munteanu et al. (2018) where $\mu = n$ and was recently generalized by Woodruff and Yasuda (2023) to more natural examples with smaller μ.

Variance-based regularization Regularization techniques have been proposed in the literature for many purposes. Most such techniques are used to reduce to the effective dimension of statistical problems or limit their expressivity to avoid overfitting. Regularization was also proposed to relax the problem of sketching logistic regression problem. In an extreme setting where the regularizer dominates the objective function, the contributions of input data points do not differ significantly. To address the bias-variance tradeoff in machine learning problems in a more meaningful way and to provably reduce the generalization error of models, Maurer and Pontil (2009) proposed to integrate a *data-dependent* variance-based regularization into the objective function. Since this results in a non-convex optimization problem even for convex objectives, Duchi and Namkoong (2019); Yan et al. (2020) used optimization tricks to reformulate a convex variant with additional parameters that can be integrated into standard hyperparameter tuning. Interestingly, this *data-dependent* regularization – in contrast to standard regularization – does not relax the sketching problem but makes it more complicated, requiring in the case of logistic regression a combination of ℓ_1 and ℓ_2 geometries to be preserved. We show that our sketching approach is capable of dealing with both simultaneously.

p-generalized normal distribution and p-probit regression The p-generalized normal distribution was introduced by Subbotin (1923) and became widely popular in the late twentieth century (Goodman and Kotz, 1973; Osiewalski and Steel, 1993; Johnson et al., 1994). We refer to (Dytso et al., 2018) for an extensive survey on applications and analytical properties of the generalized normal distribution. Among other results, this reference provides an asymptotic characterization of the tails of generalized normal distributions, which we concretize in a non-asymptotic way, similar to the classic work of Gordon (1941) on the standard normal distribution. Another nice property is the decomposability into independent marginals, which characterizes the class of multivariate generalized normal distributions (Sinz et al., 2009; Dytso et al., 2018). In summary, the class of p-generalized normal distributions naturally extends the standard normal distribution and retains several of its useful and desirable analytical properties. Hereby, it offers more parametric flexibility allowing for tails that are either heavier ($p < 2$) or lighter ($p > 2$) than normal ($p = 2$) which makes it an excellent

choice in many modeling scenarios (Dytso et al., 2018).

Most related to our work (Munteanu et al. (2022a)) are coreset and sketching algorithms for *linear* ℓ_p regression (Clarkson, 2005; Dasgupta et al., 2009; Sohler and Woodruff, 2011; Meng and Mahoney, 2013; Woodruff and Zhang, 2013; Clarkson et al., 2016), which aims at minimizing $\|Z\beta - Y\|_p$, and can be seen as a standard linear model $Y = Z\beta + \xi$, where the error term η follows a p-generalized normal distribution[1]. The earlier works relied on subsampling according to ℓ_p norms derived from a well-conditioned basis, whose approximation posed the computational bottleneck. Subsequent works improved the previous results significantly by approximating those bases via fast linear sketching techniques. To our knowledge, we are the first to study coresets and sketching for ℓ_p regression in the setting of *generalized* linear models. The authors of (Woodruff and Yasuda, 2023) also, subsequently to our work, consider p-probit regression. They are using ℓ_p-Lewis weights to construct a coreset. Note that the computation of ℓ_p-Lewis weights consumes a significant amount of time. The ℓ_p-Lewis weights can also be computed using an online algorithm but the size of the resulting coreset depends on the condition number of the data matrix.

Two layer ReLU networks The theory of neural networks is a huge and quickly growing field. Here we only give a brief summary of the work most closely related to ours.

Convergence results for neural networks with random inputs. Assuming the input data points are sampled from a Gaussian distribution is often done for proving convergence results (Zhong et al. (2017b); Li and Yuan (2017); Zhong et al. (2017a); Ge et al. (2018); Bakshi et al. (2019); Chen et al. (2022)). A more closely related work is the work of Daniely (2020) who introduced the coupled initialization technique, and showed that $\tilde{O}(n/d)$ hidden neurons can memorize all but an ε fraction of n random binary labels of points uniformly distributed on the sphere. Similar results were obtained for random vertices of a unit hypercube and for random orthonormal basis vectors. In contrast to our work, this reference uses *stochastic* gradient descent, where the nice assumption on the input distribution gives rise to the $1/d$ factor; however, this reference achieves only an approximate memorization. We note that full memorization of *all* input points is needed to achieve an error arbitrarily close to zero, and $\Omega(n)$ neurons are needed for worst case inputs. Similarly, though not necessarily relying on random inputs, Bubeck et al. (2020) shows that for *well-dispersed* inputs, the neural tangent kernel (with ReLU network) can memorize the input data with $\tilde{O}(n/d)$ neurons. However, their training algorithm is neither a gradient descent nor a stochastic gradient descent algorithm, and also their network consists of complex weights rather than real weights. One motivation of our work is to analyze standard algorithms such as gradient descent. In this work, we do not make any input distribution assumptions; therefore, these works are incomparable to ours. In particular, random data sets are often well-dispersed inputs that allow smaller width and tighter concentration, but are hardly realistic. In contrast, we conduct worst case analyses to cover all possible inputs, which might not be well-dispersed.

Convergence results of neural networks for binary classification with logistic loss. When

[1]The connection is not explicitly elaborated in those references.

considering classification with cross-entropy (logistic) loss, the maximum separation margin γ (see Definition 2.25 for a formal definition) is one of the main parameters determining the necessary size of the width. Previous separability assumptions on an infinite-width two-layer ReLU network in Cao and Gu (2019, 2020) and on smooth target functions in Allen-Zhu et al. (2019a) led to polynomial dependencies between the width m and the number n of input points. The work of Nitanda et al. (2019) relies on the NTK separation mentioned above and improved the dependence, but was still polynomial.

A recent work of Ji and Telgarsky (2020) gives the first convergence result based on an NTK analysis where the *direct* dependence on n, i.e., the number of points, is only poly-logarithmic. Specifically, they show that as long as the width of the neural network is polynomially larger than $1/\gamma$ and $\log n$, then gradient descent can achieve zero training loss.

Convergence results for neural networks with squared loss. There is a body of work studying convergence results of over-parameterized neural networks with squared loss Li and Liang (2018); Du et al. (2019c); Allen-Zhu et al. (2019c,b); Du et al. (2019b); Allen-Zhu et al. (2019a); Song and Yang (2019); Arora et al. (2019b,a); Cao and Gu (2019); Zou and Gu (2019); Du et al. (2019a); Lee et al. (2020); Huang and Yau (2020); Chen and Xu (2020); Brand et al. (2021); Li et al. (2021a); Song et al. (2021); Ailon and Shit (2022). One line of work explicitly works on the neural tangent kernel Jacot et al. (2018) with kernel matrix K. Note that most of these papers consider the squared loss function rather than the logistic loss. However the analysis often is very similar. This line of work shows that as long as the width of the neural network is polynomially larger than $n/\lambda_{\min}(K)$, then one can achieve zero training error. Another line of work instead assumes that the input data points are not too "collinear", where this is formalized by the parameter $\delta = \min_{i \neq j}\{\|x_i - x_j\|_2, \|x_i + x_j\|_2\}$[2] Li and Liang (2018); Oymak and Soltanolkotabi (2020). These works show that as long as the width of the neural network is polynomially larger than $1/\delta$ and n, then one can train the neural network to achieve zero training error. The work of Song and Yang (2019) shows that the over-parameterization $m = \Omega(\lambda^{-4} n^4)$ suffices for the same regime we consider[3]. Additional work claims that even a linear dependence is possible, though it is in a different setting. E.g., Kawaguchi and Huang (2019) show that for any neural network with nearly linear width, there exists a trainable data set. Although their width is small, this work does not provide a general convergence result. Similarly, Zhang et al. (2021) use a coupled LeCun initialization scheme that also forces the output at initialization to be 0. This is shown to improve the width bounds for shallow networks below n neurons. However, their convergence analysis is local and restricted to cases where it remains unclear how to find globally optimal or even approximate solutions. Munteanu et al. (2022b) instead focus on cases where gradient descent provably optimizes up to arbitrary small error, for which we give a lower bound of $\Omega(n)$. Note that the part of (Munteanu et al., 2022b) which focuses on squared loss has been done by coauthors independently using the same initialization.

Other than considering over-parameterization in first-order optimization algorithms, such as gradient

[2]This is also sometimes called the separability of data points.

[3]Although the title of Song and Yang (2019) is quadratic, n^2 is only achieved when the finite sample kernel matrix deviates from its limit in norm only by a constant α w.h.p., and the inputs are *well-dispersed* with constant θ, i.e., $|\langle x_i, x_j \rangle| \leq \theta/\sqrt{n}$ for all $i \neq j$. In general, Song and Yang (2019) only achieve a bound of n^4.

References	**Width m**	**Iterations T**	**Loss function**
Ji and Telgarsky (2020)	$O(\gamma^{-8}\log n)$	$O(\varepsilon^{-1}\gamma^{-2}(\sqrt{\log n}+\log(1/\varepsilon))^2)$	logistic loss
Our work (Theorem 5)	$O(\gamma^{-2}\log n)$	$O(\varepsilon^{-1}\gamma^{-2}\log^2(1/\varepsilon))$	logistic loss
Ji and Telgarsky (2020)	$\Omega(\gamma^{-1/2})$	N/A	logistic loss
Our work (Lemma 5.11)	$\Omega(\gamma^{-1}\log n)$	N/A	logistic loss
Du et al. (2019c)	$O(\lambda^{-4}n^6)$	$O(\lambda^{-2}n^2\log(1/\varepsilon))$	squared loss
Song and Yang (2019)	$O(\lambda^{-4}n^4)$	$O(\lambda^{-2}n^2\log(1/\varepsilon))$	squared loss
Munteanu et al. (2022b)	$O(\lambda^{-2}n^2)$	$O(\lambda^{-2}n^2\log(1/\varepsilon))$	squared loss

Table 1: Summary of our results and comparison to previous work. The improvements are mainly in the dependence on the parameters λ, γ, n affecting the width m. None of the results depend on the dimension d, except the lower bounds, which require $d \geq 2$. We note that the difference between regimes comes from different properties of the loss functions that affect the convergence rate, cf. Nitanda et al. (2019).

descent, Brand et al. (2021) show convergence results via second-order optimization, such as Newton's method. Their running time also relies on $m = \Omega(\lambda^{-4}n^4)$, which is the state-of-the-art width for first-order methods Song and Yang (2019), and it was noted that any improvement to m would yield an improved running time bound.

Munteanu et al. (2022b) presented in this thesis continues and improves those lines of research on understanding two-layer ReLU networks. A comparison of our results to the most closely related work is given in Table 1.

3 Sketching for logistic regression

In this section we focus on our sketching algorithm for logistic regression. We first state the setting, the algorithm and its motivation. Then we go into the analysis. Last we finish with some similar target functions for which one can also construct a sketch using our algorithm and a worst case example for variance-based regularized logistic regression.

3.1 Setting and notations

Recall that we are given a data matrix $X \in \mathbb{R}^{n\times d}$ with rows $x_i \in \mathbb{R}^d$ for $i \in [n]$ and a label vector $y \in \mathbb{R}^n$ for ℓ_1-regression. For logistic regression and variance-based regularized logistic regression labels y_i are folded in x_i as described in Subsection 2.6. Our goal is to find a weak weighted (V, a, ε)-sketch (X', w) (see Definition 2.18) for the following target functions:

$$f_1(X\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i\beta) \qquad \text{logistic regression;}$$

$$\|X\beta - y\|_1 \qquad \ell_1\text{-regression;}$$

$$f(X\beta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i\beta) + \frac{\lambda}{2n}\sum_{i=1}^{n} \ell(x_i\beta)^2 - \frac{\lambda}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(x_i\beta)\right)^2 \qquad \text{vbrlr}$$

where $\ell(r) = \ln(1 + e^r) = \ln(e^r(e^{-r} + 1)) = r + \ell(-r)$ and vbrlr stands for variance-based regularized logistic regression.

Note that the weighted versions are given by

$$f_{1w}(X\beta) = \frac{1}{n}\sum_{i=1}^{n} w_i\ell(x_i\beta) \qquad \text{logistic regression;}$$

$$\|X\beta - y\|_{1w} = \sum_{i=1}^{n} w_i|x_i\beta - y_i| = \sum_{i=1}^{n} |w_ix_i\beta - w_iy_i| \qquad \ell_1\text{-regression;}$$

$$f_w(X\beta) = \frac{1}{n}\sum_{i=1}^{n} w_i\ell(x_i\beta) + \frac{\lambda}{2n}\sum_{i=1}^{n} w_i\ell(x_i\beta)^2 - \frac{\lambda}{2}\left(\frac{1}{n}\sum_{i=1}^{n} w_i\ell(x_i\beta)\right)^2 \qquad \text{vbrlr.}$$

Further note that we can split f into three functions $f_1(X\beta)$, $f_2(X\beta) = \frac{\lambda}{2n}\sum_{i=1}^{n} \ell(x_i\beta)^2$, and $f_3(X\beta) = \frac{\lambda}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \ell(x_i\beta)\right)^2 = \frac{\lambda}{2}f_1(X\beta)^2$. As described in Section 2.6 the term $f_2(X\beta) + f_3(X\beta)$ can also be interpreted as a variance regularization.

3.2 The algorithm

3.2.1 Motivation

Our sketching algorithm was first published in (Munteanu et al., 2021) inspired by the algorithm presented in (Clarkson and Woodruff, 2015) and it is to our knowledge the first oblivious sketching algorithm for generalized linear regression. The idea is that we take multiple uniform samples of different sizes and apply the Count-Min sketch to those which is the main difference to the algorithm of (Clarkson and Woodruff, 2015) where the Count sketch is applied. This can also be interpreted by multiplying with a sketching matrix of the form

$$S = \begin{bmatrix} S_0 \\ S_1 \\ \vdots \\ S_{h_m} \end{bmatrix}.$$

Here each $S_i \in \mathbb{R}^{N \times n}$ has at most one single entry 1 in each column and all other entries are 0. More precisely each row of S_i corresponds to a subsample of $[n]$ of size roughly $\frac{n}{b^i}$ for some $b > 2$ where $j \in [n]$ is part of the sample if there is a 1 in the j-th column of S_i and the row with the 1 is picked uniformly at random which corresponds to applying the Count-Min sketch to the sample. Entries in the same rows are added up, meaning the rows of X are in the same bucket. As our first bounds were suboptimal we viewed the analysis from a different point of view. In this manuscript we present an updated version (Munteanu et al. (2023)) of the algorithm we first presented in (Munteanu et al., 2021) which improves both the approximation guarantee as well as the sketch size. More precisely with minor modifications in the algorithm but major modifications in the analysis we show that the size of the sketch can be reduced from roughly $\tilde{O}(\mu^7 d^5)$ to $\tilde{O}(\mu^2 d^3)$, while improving the $O(1)$ approximation guarantee in expectation from at least 8 to roughly 2 for either the logistic or ℓ_1 loss. The most important modifications are:

- Instead of having an uniform sampling separately we use the last level of our sketch as uniform sampling;
- We are modifying level 0 as well as its analysis;
- We give a detailed analysis of each level to get a better idea of what elements are preserved and which elements have a contribution at each level allowing as to get better bounds on the parameters.

3.2.2 Parameters

Practical parameters Our sketch consists of $h_m + 1 = O(\log n)$ levels. In each level we take a subsample of the rows $i \in [n]$ at a different rate and hash the sampled items uniformly to a number of buckets. All items that are mapped to the same bucket are added up. This corresponds to a `CountMin` sketch (Cormode and Muthukrishnan, 2005) applied to the subsample taken at each level.

More specifically, we will use the following parameters:

- h_m: the number of levels,
- N_h: the number of buckets at level h,
- p_h: the probability that any element x_i is sampled at level h.

Our goal it to build a sketch $X' \in \mathbb{R}^{n' \times d}$ where $n' = \sum_{i=0}^{h_m} N_h$. If all levels have the same number N of buckets we have that $n' = (h_m + 1)N$.

Theoretical parameters Besides the mentioned parameters that are crucial for describing the algorithm there are a bunch of theoretical parameters for the analysis determining the values of the practical parameters:

- ε: the relative error for the contraction bounds,
- δ: the allowed failure probability,
- d: the number of features,
- n: the number of rows,
- b: the relative difference between the sample sizes of two adjacent levels, i.e. $p_h = bp_{h+1}$,
- m_1: negative logarithm of the failure probability for the contraction bound.

3.2.3 Pseudo code

Algorithm 1 Oblivious sketching algorithm for logistic regression.

Input: Data $X \in \mathbb{R}^{n \times d}$, number of rows $k = N \cdot h_m + N_u$, parameters $b > 1, s \geq 1$ where $N = s \cdot N'$ for some $N' \in \mathbb{N}$;
Output: weighted Sketch $C = (X', w) \in \mathbb{R}^{k \times d}$ with k rows.;

for $h = 0 \dots h_m$ **do** ▷ construct levels $1, \dots h_m$ of the sketch
  initialize sketch $X'_h = \mathbf{0} \in \mathbb{R}^{N \times d}$ at level h;
  initialize weights $w_h = b^h \cdot \mathbf{1} \in \mathbb{R}^N$ at level h;
set $w_0 = \frac{w_0}{s}$; ▷ adapt weights on level 0 to sparsity s
for $i = 1 \dots n$ **do** ▷ sketch the data
  for $l = 1 \dots s$ **do** ▷ densify level 0
    draw a random number $B_i \in [N']$;
    add x_i to the $((l-1) \cdot N' + B_i)$-th row of X'_0;
  assign x_i to level $h \in [1, h_m - 1]$ with probability $p_h = \frac{1}{b^h}$;
  draw a random number $B_i \in [N]$;
  add x_i to the B_i-th row of X'_h;
  add x_i to uniform sampling level h_m with probability $p_{h_m} = \frac{1}{b^{h_m}}$;
Set $X' = (X'_0, X'_1, \dots X'_{h_m})$;
Set $w = (w_0, w_1, \dots w_{h_m})$;
return $C = (X', w)$;

3.2.4 Description of the algorithm

As we read the input, we sample each element x_i for each level $h \leq h_m$ with probability p_h. The sampling probabilities are exponentially decreasing, i.e., $p_{h+1} = p_h/b$ for some $b \in \mathbb{R}$ with $b > 1$. The weight of any bucket at level h is set to $1/p_h$. At level h_m, we have $p_{h_m} = \frac{N_{h_m}}{n}$. It thus corresponds to a uniform subsample and the number of buckets is equal to the number of rows that are sampled, i.e., $N_u := N_{h_m} \approx np_{h_m} =: np_u$. At level 0 we sample all rows, i.e., $p_0 = 1$ and the number of buckets is either the same as for the levels $h \in (0, h_m)$ or less. Consequently level 0 is a standard `CountMin` sketch of the entire data. Note that here we will also show an alternative version, where at level 0 rows are sampled multiple times. All levels $h \in (0, h_m)$, or in other words all levels but level 0 and level h_m, have the same number of buckets $N_h = N$. For the exact details we refer to the analysis. The idea of our algorithm is that for each fixed $\beta \in \mathbb{R}^d$ we can partition the rows of $X\beta$ into weight classes depending on their contribution to the objective function. Each level catches a certain range of weight classes if their total contribution is large enough. For example level h_m will represent all small weight classes and level 0 will represent the so-called *heavy hitters*, i.e. rows that can have a large contribution to the target function.

In order to get an arbitrarily good approximation we will also present the possibility setting the number of buckets at level 0 at random.

3.2.5 Idea of the analysis

The logistic loss function can be split into two parts:

$$\ell(r) = \ln(1 + e^r) = \ln(e^r(e^{-r} + 1)) = r + \ell(-r).$$

We will use this split as follows: Let $z = X\beta \in \mathbb{R}^n$. Then it holds that

$$f_1(z) = \sum_{z_i > 0} \ell(z_i) \; + \; \sum_{z_i \leq 0} \ell(z_i) = \sum_{z_i > 0} z_i \; + \; \sum_{i=1}^{n} \ell(-|z_i|). =: \|z^+\|_1 + f_s(z)$$

where $z^+ \in \mathbb{R}^n_{\geq 0}$ is the vector that we get by setting all negative coordinates of z to 0.

We will show that our sketch approximates both parts well. For the first part will split the values of z^+ into different weight classes $W_0, W_1 \ldots$ such that all elements in the same weight class roughly have the same contribution to $\|z^+\|_1$. We will show that for each weight class W_q that has a non negligible contribution to $\|z^+\|_1$ there is one level of our sketch which has the same (weighted) contribution to $f_{1w}(X'\beta)$ up to a factor of $1 \pm \varepsilon$. Further there are only few (depending on the size of the sketch) other levels where W_q has a non negligible contribution. More precisely weight classes with small entries will be well represented in higher levels but their contribution at small level will be small and weight classes with large entries will be well

represented in lower levels. The largest entries, the heavy hitters, for instance will be well represented in level 0 and will not be present at higher levels.

The second part f_s will be captured by the uniform sample at level h_m. The remaining levels will only have a negligible contribution to the second part.

3.2.6 Outline of the analysis

The pattern of our analysis is as follows:

1) We fix a point $z = X\beta$ as described before.

2) We prove that the contraction bounds, i.e. $f_{1w}(X'\beta) \geq (1-\varepsilon)f_1(z)$, hold with exponentially small failure probability.

3) We show that there is a net such that if the contraction bound holds for any point of the net, then it holds for any parameter vector β.

4) Lastly, using Markov's inequality (Lemma 2.8) again fixing z we show that we have that $\mathbb{E}(f_{1w}(X'\beta)) \leq kf_1(z)$ for some value k implying we have dilation bounds with constant probability.

5) For ℓ_1 regression and variance based regularized logistic regression the ideas and the proofs are similar but need mostly technical adaptations. We will thus outline the changes in the analysis.

We first start by analyzing f_s here and then turn our attention to $\|z^+\|_1$. For the later before proving contraction and dilation bounds we explain some theory of the weight classes and each level individually:

1) There exists a subset Q^* such that the contribution of all weight classes not in Q^* is negligible.

2) For each level there exists an interval Q_h such that for all $q \in Q_h \cap Q^*$ the contribution of W_q is preserved up to an relative error of ε with exponentially small failure probability.

3) For each level there exists an interval Q'_h such that for all $q \notin Q'_h$ the contribution of W_q is 0 with failure probability bounded by δ/h_m.

3.3 High level description of the analysis

We start by splitting the functions f_1 and f_2 into multiple parts:

Lemma 3.1. *It holds that*

$$nf_1(X\beta) = \sum_{x_i\beta>0} |x_i\beta| \;+\; \sum_{i=1}^{n} \ell(-|x_i\beta|)$$

and similarly we have that

$$nf_2(X\beta) = \textstyle\sum_{x_i\beta>0} |x_i\beta|^2 \;+\; 2\sum_{x_i\beta>0} \ell(-|x_i\beta|)\cdot|x_i\beta| \;+\; \sum_{i=1}^{n} \ell(-|x_i\beta|)^2.$$

Proof. Note that for $r \in \mathbb{R}$ it holds that

$$\ell(r) = \ln(1 + e^r) = \ln\left(\left(e^{-r} + 1\right) e^r\right) = \ln\left(e^{-r} + 1\right) + \ln(e^r) = \ell(-r) + r.$$

Now the first equation follows immediately by

$$\ell(x_i\beta) = x_i\beta + \ell(-x_i\beta) = |x_i\beta| + \ell(-|x_i\beta|) \text{ for } x_i\beta > 0 \text{ and}$$
$$\ell(-|x_i\beta|) = \ell(x_i\beta) \text{ for } x_i\beta \leq 0.$$

Further we have that

$$(x_i\beta + \ell(-x_i\beta))^2 = (x_i\beta)^2 + 2\ell(-x_i\beta)x_i\beta + \ell(-x_i\beta)^2.$$

Thus the second equality follows by substituting $\ell(x_i\beta)^2$ with

$$|x_i\beta|^2 + 2\ell(-|x_i\beta|)|x_i\beta| + \ell(-|x_i\beta|)^2 = (x_i\beta)^2 + 2\ell(-x_i\beta)x_i\beta + \ell(-x_i\beta)^2$$

for $x_i\beta > 0$ and $\ell(-|x_i\beta|)^2$ for $x_i\beta \leq 0$. □

This can be used in the following way: if all $x_i\beta$ make only small contributions then uniform sampling performs well. This is not the case for all parts of f but it holds for some 'small' parts of f that appear in the splitting introduced in Lemma 3.1.

Next we deal with the remaining 'large' parts of f. We will first analyze the approximation for a single β. To this end fix $\beta \in \mathbb{R}^d$ and set $z = X\beta$. Our goal is to approximate $\|z^+\|_1 := \sum_{i:z_i>0} z_i$. We assume w.l.o.g. that $\|z\|_1 = 1$. We can do this since we are only interested in the relative values of the coordinates, i.e. $\frac{|z_i|}{\|z\|_1}$ In order to prove that $\|(Sz)^+\|_1$ approximates $\|z^+\|_1$ well, we define weight classes: given $q \in \mathbb{N}$ we set

$$W_q^+ = \{i \in [n] \mid z_i \in (2^{-q-1}, 2^{-q}]\}.$$

Our analysis applies with slight adaptations to ℓ_1 regression preserving $\|z\|_1$ for the residual vector $z = X\beta - y$.

Next we give a high level description for preserving $\|z^+\|_1$ needed for logistic loss:

Contraction bounds Our first goal is to show that it holds that $f_{1w}(X'\beta) \geq (1-\varepsilon)f_1(z)$ with failure probability bounded by roughly $\exp(-m_1)$, i.e. exponentially small allowing us to use the union bound over a net of exponential size. We set $q_m = \log_2(\frac{n(\mu+1)}{\varepsilon}) = O(\ln(n))$ as $n \geq \max\{\mu, \varepsilon^{-1}\}$. We say that W_q^+ is important if $\|W_q^+\|_1 \geq \varepsilon' := \frac{\varepsilon}{\mu q_m}$ and set $Q^* = \{q \leq q_m \mid W_q \text{ is important }\}$. The idea is that the remaining weight classes can only have negligible contributions to $\|z^+\|_1$, so it suffices to analyze Q^*. To prove the contraction bound for z, i.e., that $\|(Sz)^+\|_1 \geq (1 - c\varepsilon)\|z^+\|_1$ holds for an absolute constant c, it

suffices to show that the contributions of important weight classes are preserved. For a bucket B we set $G(B) := \sum_{j \in B} z_j$ and $G^+(B) = \max\{G(B), 0\}$. In fact, we show that for each level h, there exists an 'inner' interval $Q_h = [q_h(2), q_h(3)]$ such that if W_q^+ for $q \in Q_h$ is important, then there exists a subset $W_q^* \subseteq W_q^+$ such that each element of W_q^* is sampled at level h and such that each $i \in W_q^*$ is in a bucket containing no other element of any $W_{q'}^*$ for any $q' \in [q_m]$ and $\sum_{i \in W_q^*} G(B_i) \geq (1-\varepsilon)\|W_q^+\|_1 \cdot p_h$, where B_i is the bucket at level h containing z_i. Since the weight of all buckets at level h is equal to p_h^{-1} we have that the contribution of W_q^* is indeed at least $(1-\varepsilon)\|W_q^+\|_1$. The choice of our parameters then guarantees that $\bigcup Q_h = \mathbb{N}$ and thus for any important weight class there is at least one level where it is well represented.

Net argument Finally, we construct a net of size $|\mathcal{N}_k| = \exp(O(d \log(n)))$. We ensure that the contraction bound holds for each fixed net point $z \in \mathcal{N}_k$ with failure probability at most $\frac{\delta}{|N_k|}$ which will dominate – among other parameters – the size of our sketch. By a union bound the contraction result holds for the entire net with probability at least $1 - \delta$. The net is sufficiently fine, such that we can conclude the contraction bound by relating all other points $z = X\beta \in \mathbb{R}^n$ to their closest point in the net.

Dilation bounds Our second goal is to show that with any probability $P \in [0, 1]$ it holds that $f_{1w}(X'\beta) \leq k_P f_1(z)$ for some number $k_P \in [1, \infty)$ depending only on P if $f(z) \leq (1-\varepsilon)f(0)$. More precisely we will show that the expected contribution of any weight class is at most $k\|W_q^+\|_1$ for some k which is either constant or $1 + \varepsilon$ depending on the sketch size. We then apply Markov's inequality to get k_P. In contrast to the contraction bounds we will not get an exponentially small failure probability here but we only need top apply the dilation bounds to the optimum β^*. Note that to get below $k = 2$ we apply a random shift at level 0, i.e., we choose the number of buckets at level 0 randomly. We investigate again each level separately and prove that for each level h there exists an 'outer' interval $Q'_h = [q_h(1), q_h(4)]$ such that for any $q \notin Q'_h$ the weight class W_q makes no contribution at level h at all. More specifically we show that no element of W_q appears at level h for $q < q_h(1)$ and that for any bucket B at level h that contains only elements of $\bigcup_{q > q_h(4)} W_q$ it holds that $G(B) \leq 0$. For the later fact we use that $f(z) \leq (1-\varepsilon)f(0)$ implies that the negative elements of z sum up to a larger absolute value than the positive entries of z even if some larger negative entries are removed.

Then we show that if N is large enough it holds that for each $q \in \mathbb{N}$ there are at most k levels h such that $q \in Q'_h$ and that the expected contribution of any weight class at any level is bounded by $\|W_q^+\|_1$. We conclude that the expected contribution of any weight class is at most $k\|W_q^+\|_1$. Increasing the size of N increases the size of an 'inner' interval $[q_h(2), q_h(3)] =: Q_h \subset Q'_h$ while the size of Q'_h remains (almost) unchanged such that $|Q'_h|/|Q_h|$ approaches 1. As a consequence, this also decreases the number of indices $q \in \mathbb{N}$ that appear in two intervals of the form Q'_h. More precisely, we show that for each $c \in \mathbb{N}$ we can increase N in such a way that only a $1/c$ fraction of the weight classes appear in two of those intervals. Note that all weight classes that appear only in a single Q'_h have an expected contribution of $\|W_q^+\|_1$. However those appearing in Q'_h for two different h have an an expected contribution of up to $2\|W_q^+\|_1$. Thus to reduce the expected contribution of any weight class below 2 we apply a random shift at level 0, implicitly setting

$q_0(3)$ randomly in an appropriate way such that the probability of any weight class being in two sets of the form Q'_h is at most $\frac{1}{c}$ and thus the expected contribution of *any* weight class W_q is bounded by at most $(1 + 1/c)\|W_q^+\|_1$.

Extension to variance-based regularized logistic regression We show that our algorithm also approximates the variance well under the assumption that roughly $f_1(X\beta) \leq \ln(2)$. We stress that this assumption does not rule out the existence of good approximations. Indeed, even the minimizer is contained, since we have that $\min_{\beta\in\mathbb{R}^d} f(X\beta) \leq f(0) = f_1(0) = \ln(2)$. Focusing on a single $z = X\beta$, we need to show that $\sum_{i:z_i>0} z_i^2$ is approximated well, which is done very similarly to the analysis for $\sum_{i:z_i>0} z_i$ sketched above, but with several adaptions to account for the squared loss function. We note that the increased sketching dimension in terms of $\sqrt{n}$ comes from the inter norm inequality $\|x\|_1 \leq \sqrt{n}\|x\|_2$.

We also give a lower bound by giving an example where the size of our sketch needs to be at least $\Omega(\sqrt{n})$. However this does not rule out other methods that may allow a lower sketching dimension. For example `Count`-sketch is known to work for ℓ_1 and ℓ_2 norms simultaneously within polylogarithmic size (Clarkson and Woodruff, 2015). But we stress that the standard sketches from the literature do not work for asymmetric functions since they confuse the signs of contributions leading to unbounded errors for our objective function or even for plain logistic regression, see (Munteanu et al., 2021).

3.4 Analysis

3.4.1 Assumptions

For technical reasons we make the following assumption:

Assumption 3.1. *We assume that:*

$$h_m = \min\left\{i \in \mathbb{N} \mid \frac{M_i}{bN} \leq 12\ln(n)\right\} \tag{8}$$

$$q_m = O(\ln(n)) \tag{9}$$

$$N \geq 18 \cdot 32 m_1 q_m \mu^2/\varepsilon^6 \tag{10}$$

$$b = \frac{N\varepsilon^5}{32 m_1 q_m \mu} \geq \frac{18\mu}{\varepsilon} \tag{11}$$

$$m_1 = \ln(\delta^{-1}) + O(d\ln(n))) \tag{12}$$

$$p_u \geq \frac{64\mu m_1}{\varepsilon^2 n}. \tag{13}$$

Since we want our sketch to have fewer than n rows we will also assume that $n \geq \varepsilon^{-1}, \mu, d, \delta^{-1}$. We also assume that $\varepsilon \leq 1/4$.

3.4.2 Estimating the small parts of f

Lemma 3.1 can be used in the following way: if all $x_i\beta$ make only small contributions then uniform sampling performs well. This is not the case for all parts of f but it holds for some 'small' parts of f that appear in the splitting introduced in Lemma 3.1. More precisely we get the following lemma:

Lemma 3.2. *For arbitrary $i \in [n]$ it holds that $\ell(-|x_i\beta|) < 1$ and also $2\ell(-|x_i\beta|)|x_i\beta| + \ell(-|x_i\beta|)^2 \leq 3$.*

Proof. First observe that $\ell(-|x_i\beta|) \leq \ell(0) = \ln(2) < 1$, proving the first part of the lemma.

Next note that

$$\begin{aligned}\ell(-|x_i\beta|) = \ln(1+\exp(-|x_i\beta|)) &= \int_1^{1+\exp(-|x_i\beta|)} \frac{1}{t}\,dt \\ &\leq \int_1^{1+\exp(-|x_i\beta|)} 1\,dt = \exp(-|x_i\beta|).\end{aligned}$$

Using that $\ln(t) \leq |t|$ for all $t > 0$ we conclude that

$$\ell(-|x_i\beta|)|x_i\beta| \leq \exp(\ln(|x_i\beta|) - |x_i\beta|) \leq e^0 = 1.$$

Now combining everything we get that

$$2\ell(-|x_i\beta|)|x_i\beta| + \ell(-|x_i\beta|)^2 \leq 2 + 1^2 \leq 3.$$

□

Next we note that the optimal value of $f(X\beta)$ is bounded from below:

Lemma 3.3. *For all $\beta \in \mathbb{R}^d$ it holds that $nf(X\beta) \geq nf_1(X\beta) \geq \frac{\ln(2)n}{\mu}(1+\ln(\mu)) = \Omega\left(\frac{n}{\mu}(1+\ln(\mu))\right)$.*

Proof. First we show that there is a monotonically rising function $h_\ell : [0,\infty) \to [\ln(2), 1)$ such that it holds that $\ell(-r) = h_\ell(r)\exp(-r)$. To see this note that for any $r \geq 1$ it holds that $\ln(r) = \int_1^r \frac{1}{y}dy$. Next note that the function $h_\ell(r) = \frac{\int_1^{1+e^{-r}} \frac{1}{y}dy}{e^{-r}}$ determining the average value of $\frac{1}{y}$ in the interval $[1, 1+e^{-r}]$ is a monotone rising function as $\frac{1}{y}$ is a monotonically falling function for $y \in [0,\infty)$. Further we have that $h(0) = \ln(2)$ and $h(r) \leq 1$ as $\frac{1}{y} \leq 1$ for $y \in [1,2]$. Then our first claim follows as for $r \geq 0$ we have

$$\ell(-r) = \ln(1+e^{-r}) = \int_1^{1+e^{-r}} \frac{1}{y}dy = h_\ell(r)e^{-r}.$$

Next using this fact we get

$$f_1(z) \geq \ln(2)(\sum_i \exp(\min\{z_i, 0\}) + \|z^+\|_1).$$

Since $\exp(v)$ is convex, Jensen's inequality implies

$$\sum_i \exp(\min\{z_i, 0\}) = n \sum_i \frac{1}{n} \exp(\min\{z_i, 0\}) \geq n \exp(\frac{1}{n} \sum_i \min\{z_i, 0\}).$$

Using this argument we get for $y = \frac{\|z^-\|_1}{n}$ that $\sum_i \exp(\min\{z_i, 0\}) \geq n \exp(-y)$. Recall that $\|z^+\|_1 \geq \frac{yn}{\mu}$ holds by definition of μ.

We conclude that $nf_1(z) \geq \ln(2)(n \exp(-y) + \frac{yn}{\mu})$. The function $(n \exp(-y) + \frac{yn}{\mu})$ is minimized over y if its first derivative is zero, i.e., if

$$n \exp(-y) = \frac{n}{\mu}$$

which is equivalent to $y = \ln(\mu)$. Hence $nf(z) \geq \ln(2) \left(\frac{n}{\mu} + \frac{n \ln(\mu)}{\mu} \right)$. □

We use the previous two lemmas to show that our sketch approximates the given parts of f well enough with high probability. To this end, we set $g_1(t) = \ell(-|t|)$, $g_2(t) = 2\ell(-|t|)|t| + \ell(-|2t|)$ and $g(t) = g_1(t) + \lambda g_2(t)$. We set $\mu_z = \frac{\|z^-\|_1}{\|z^+\|_1}$. Note that $\mu_z \leq \mu$.

The following Lemma is needed only in the case that $b \leq \frac{\mu}{\varepsilon}$.

Lemma 3.4. *Given any $\beta \in \mathbb{R}^d$ with failure probability at most $2 \exp(-m_1)$ the event $\mathcal{E}_0$ holds that*

$$\left| \sum_{i=1}^{n'} w_i g(x_i' \beta) - \sum_{i=1}^{n} g(x_i \beta) \right| \leq \varepsilon \cdot \max \left\{ \sum_{i=1}^{n} g(x_i \beta), \frac{n}{2\mu} \right\} \leq \varepsilon f(X\beta).$$

Proof of Lemma 3.4. We first show that the contribution to $\sum_{i=1}^{n'} w_i g(x_i' \beta)$ of the levels other than h_m is small.

Therefore note that the (relative) total weight of all buckets in a level less than h_m is at most $\sum_{h=1}^{h_{\max}} b^{-h} = b^{-1} \cdot \frac{1 - b^{-h_{\max}}}{1 - b^{-1}} \leq \frac{2}{b} \leq \frac{\varepsilon}{9\mu}$ by Assumption 3.1. Thus by Lemma 3.2 and Lemma 3.3 we have that

$$\sum_{i=1}^{n'} w_i g(x_i' \beta) \mathbf{1}_{w_i/n < 1} \leq \sum_{i=1}^{n'} 3 w_i \cdot \mathbf{1}_{w_i/n < 1} \leq \frac{\varepsilon}{3\mu} \leq \frac{\varepsilon}{3} \cdot f(X\beta).$$

Now let $k \in \{1, 2\}$. For $i \in [n]$, consider the random variable $X_i = g_k(z_i)$ if z_i is at level h_m, and $X_i = 0$ otherwise. Then we have

$$E = \mathbb{E} \left(\sum_{i=1}^{n} X_i \right) = \sum_{i=1}^{n} p_u g_k(z_i) = p_u \sum_{i=1}^{n} g_k(x_i \beta).$$

Further we have $X_i \leq 3$ by Lemma 3.2. It holds that

$$\mathbb{E} \left(\sum_{i=1}^{n} X_i^2 \right) = \sum_{i=1}^{n} p_u g_k(z_i)^2 \leq p_u \sum_{i=1}^{n} 3 g(x_i \beta) = 3E.$$

We set

$$L = p_u \cdot \max\left\{\sum_{i=1}^{n} g_k(x_i\beta), \frac{n}{2\mu}\right\} \geq E.$$

By Assumption 3.1 we have that $p_u \geq \frac{64\mu m_1}{\varepsilon^2 n}$. By Lemma 3.2 it holds that $X_i \leq 3$. Thus, using Bernstein's inequality we get that

$$\begin{aligned} P\left(\left|\sum_{i=1}^{n} X_i - E\right| \geq \frac{\varepsilon}{3} \cdot L\right) &\leq \exp\left(\frac{-\varepsilon^2 L^2/8}{3E + E}\right) \\ &= \exp\left(\frac{-\varepsilon^2 L}{32}\right) \\ &\leq \exp\left(\frac{-\varepsilon^2 p_u n/\mu}{64}\right) \\ &\leq \exp(-m_1). \end{aligned}$$

Using the union bound for $k = 1$ and $k = 2$ yields that

$$P\left(\left|\sum_{i=1}^{n'} w_i g(x_i'\beta) - \sum_{i=1}^{n} g(x_i\beta)\right| > \frac{2\varepsilon}{3} \cdot \max\left\{\sum_{i=1}^{n} g(x_i\beta), \frac{n}{2\mu}\right\}\right) \leq 2\exp(-m_1).$$

By Lemma 3.3 we have $f(X\beta) \geq \frac{n}{2\mu}$. It also holds that $f(X\beta) \geq \sum_{i=1}^{n} g(x_i\beta)$. We thus conclude that $\max\left\{\sum_{i=1}^{n} g(x_i\beta), \frac{n}{2\mu}\right\} \leq f(X\beta)$.

Combining everything gives us

$$\left|\sum_{i=1}^{n'} w_i g(x_i'\beta) - \sum_{i=1}^{n} g(x_i\beta)\right| \leq \varepsilon \cdot \max\left\{\sum_{i=1}^{n} g(x_i\beta), \frac{n}{2\mu}\right\} \leq \varepsilon f(X\beta).$$

□

3.4.3 Estimating $\|z^+\|_1$

Here we deal with the remaining 'large' parts of f. We will first analyze the approximation for a single β. To this end fix $\beta \in \mathbb{R}^d$ and set $z = X\beta$. Our goal is to approximate $\|z^+\|_1 := \sum_{i:z_i>0} z_i$ where $z^+ \in \mathbb{R}^n_{\geq 0}$ is the vector that we get by setting all negative coordinates of z to 0. We assume w.l.o.g. that $\|z\|_1 = 1$. In order to prove that $\|(Sz)^+\|_1$ approximates $\|z^+\|_1$ well, we define weight classes: given $q \in \mathbb{N}$ we set $W_q^+ = \{i \in [n] \mid z_i \in (2^{-q-1}, 2^{-q}]\}$. Before proving the contraction bounds we will analyze how weight classes can contribute at individual levels. Our analysis applies with slight adaptations to ℓ_1 regression preserving $\|z\|_1$ for the residual vector $z = X\beta - Y$. Recall that that W_q^+ is important if $\|W_q^+\|_1 \geq \varepsilon' := \frac{\varepsilon}{\mu q_m}$ and $Q^* = \{q \leq q_m \mid W_q \text{ is important }\}$ where $q_m = \log_2(\frac{n(\mu+1)}{\varepsilon})$. The following lemma shows that the total contribution of the weight classes that are not important is negligible:

Lemma 3.5. *It holds that* $\sum_{q\in Q^*} \|W_q^+\|_1 \geq (1-2\varepsilon)\|z^+\|_1$.

Proof of Lemma 3.5. First note that

$$\sum_{|z_i|<2^{-q_m}} |z_i| \leq n \cdot \frac{\varepsilon}{(\mu+1)n} = \varepsilon/(\mu+1).$$

Second note that $\sum_{q\leq q_m, q\notin Q^*} \|W_q^+\|_1 \leq q_m \cdot \frac{\varepsilon}{(\mu+1)q_m} \leq \varepsilon/(\mu+1)$. By the μ-condition we have that $\|z^-\|_1 \leq \mu\|z^+\|_1$ and thus we get that $1 = \|z^-\|_1 + \|z^+\|_1 \leq \mu\|z^+\|_1 + \|z^+\|_1$. Consequently, $\|z^+\|_1 \geq \frac{1}{\mu+1}$ and

$$\sum_{q\in Q^*} \|W_q^+\|_1 = \sum_{q\in\mathbb{N}} \|W_q^+\|_1 - \sum_{q\notin Q^*, q\leq q_m} \|W_q^+\|_1 - \sum_{q>q_m} \|W_q^+\|_1 \geq \|z^+\|_1 - \frac{2\varepsilon}{(\mu+1)} \geq (1-2\varepsilon)\|z^+\|_1.$$

□

3.4.4 Analysis for a single level

Fix $h \in [0, h_m]$. Our goal in this subsection and the following subsection is to analyze one level individually. More precisely we look at the number of rows ending up in the level, as well as which weight classes can have a notable contribution as well as the weight classes that are likely to have their contribution preserved up to small error, if they contain enough elements or contain heavy hitters, which we show is the same as being important. More precisely we establish bounds $q_h(1) < q_h(2) < q_h(3) < q_h(4)$ such that if $q \in [q_h(2), q_h(3)]$ then, if W_q is important, then its contribution is preserved in this level up to a small error for $q \notin Q'_h = [q_h(1), q_h(4)]$ then the contribution of the weight class to the level is negligible. We will later use this to determine the probabilities p_{h_i} in such a way that $q_0(2) = 0$ and we have $q_{h_i}(3) = q_{h_{i+1}}(2)$ to guarantee that for any important weight class there is a level where its contribution is preserved. For $q \in Q'_h \setminus Q_h$ we will show later that the expected contribution is bounded by $\|W_q\|_1$. However it is likely or at least possible that the contribution is both lower or higher so in particular we cannot get a sufficient bound failure probability of having a preserved contribution here. Thus our goal here is to make $\frac{|Q'_h|}{|Q_h|}$ as small as possible.

As the weight classes with heavy hitters do not need to contain a large number of elements we will deal with them in the subsection that follows after this one, where we will have look at level 0 specifically.

First consider the number of elements at a fixed level h. We can view it as a binomial random variable with parameters n and p_h since the probability for any row to appear at level h is p_h. Since we fix h in this subsection, we set $M = M_h = p_h n$, $p = p_h = \frac{M}{n}$ and $N = N_h$. We set $U \subset [n]$ to be the set of elements that are sampled at level h. We also set $\mu_z = \frac{\sum_{z_i<0} |z_i|}{\sum_{z_i>0} |z_i|} \leq \mu$.

Our main lemma is the following:

Lemma 3.6. *With probability at least* $1 - \frac{\delta}{h_m}$ *the weight classes* W_q *with either*

$$q \geq q_{(M,N)}(4) := \log_2\left(\gamma_2^{-1}\right) := \log_2\left(\frac{2N\ln(Nh_{\max}/\delta)}{p\varepsilon^2}\right) \text{ or}$$
$$q \leq q_{(M,N)}(1) := \log_2\left(\frac{\mu_z\delta}{ph_m}\right)$$

have zero contribution to $\sum_B G^+(B)$*, i.e., for any bucket* B *we have* $\sum_{z_i \in B \setminus I_r} z_i \leq 0$ *where* $I_r = \{i \in [n] \mid z_i \in W_q, q \in [q_{(M,N)}(1), q_{(M,N)}(4)]\}$.

Further, with failure probability at most $\exp(-\Omega(m_1))$ *there exists, for each* $q \in [q_{(M,N)}(2), q_{(M,N)}(3)]$*, where*

$$q_{(M,N)}(2) := \log_2\left(\frac{8q_m\mu_z m_1}{\varepsilon^3 p}\right) \text{ and}$$
$$q_{(M,N)}(3) := \log_2\left(\frac{N\varepsilon^2}{4p}\right),$$

a set W_q^* *such that* $\sum_{i \in W_q^*} G(B_i) \geq (1-\varepsilon)^2\|W_q^+\|_1 \cdot \frac{M}{n}$.

It thus holds that

$$q_{(M,N)}(2) - q_{(M,N)}(1) = \log_2\left(\frac{8q_m m_1 h_m}{\varepsilon^3\delta}\right)$$
$$q_{(M,N)}(3) - q_{(M,N)}(2) = \log_2\left(\frac{N\varepsilon^5}{32m_1\mu q_m}\right) =: \log_2(b)$$
$$q_{(M,N)}(4) - q_{(M,N)}(3) = \log_2\left(\frac{8\ln(Nh_m/\delta)}{\varepsilon^4}\right).$$

If $N = M$ then we set $q_{(M,N)}(3) = q_{(M,N)}(4) = \infty$. If $M = n$ then we set $q_{(M,N)}(1) = q_{(M,N)}(2) = 0$. We set $q_h(i) = q_{(M_h,N_h)}(i)$ for $i \in \{1,2,3,4\}$ and $Q_h = [q_h(2), q_h(3)]$ to be the well-approximated weight classes, and $Q'_h = [q_h(1), q_h(4)]$ to be the relevant weight classes at level h.

We further define the following threshold and set:

$$\gamma_1 := \frac{p}{3m_1}$$
$$Y_1 := \{i \in [n] \mid |z_i| \geq \gamma_1\}$$

Here Y_1 is the 'set of large elements'. We set $\mathcal{B}_h$ to be the set of all buckets at level h. Recall that $m_1 \in \mathbb{R}$ is a lower bound on the negative logarithm of the failure probability, which we will need later when union bounding over all failure probabilities. Also recall that $G(B) = \sum_{i \in B} z_i$ is the sum of all rows in a bucket B. The following lemma yields the inner bounds, i.e., bounds for $q_h(2)$ and $q_h(3)$, which are the weight class indices that are well represented by U and will later be used to prove the contraction bound. The first two items show that there are at most εN buckets at level h that either contain a large element or have a large sum of small contributions. The third item shows that if W_q has sufficiently many elements, then there exists

a large subset W_q^* where each element is in a bucket with no other large entry such that $\|W_q^*\|_1$ is close to $\|W_q^+\|_1 \cdot \frac{M}{n}$. The fourth item shows that $\sum_{z_i \in W_q^*} G(B_i)$ is close to $\|W_q^*\|_1$.

Lemma 3.7. *The following claims hold:*

1) $|Y_1 \cap U| \leq \varepsilon N/2$ *with failure probability at most* $\exp(-m_1)$*;*

2) Let $\mathcal{B} = \{B \in \mathcal{B}_h \mid \sum_{i \in B \setminus Y_1} |z_i| \leq \frac{4p}{\varepsilon N}\}$*. Then* $|\mathcal{B}| \geq (1 - \frac{\varepsilon}{2})N$ *with failure probability at most* $\exp(-m_1)$*;*

3) Assume that $q \geq \log_2(\frac{8q_m\mu_z m_1}{\varepsilon^3 p})$ *and that* W_q^+ *is important or* $|W_q| \geq 8m_1\varepsilon^{-2} \cdot p^{-1}$*. Then with failure probability at most* $\exp(-m_1)$ *there exists* $W_q^* \subset W_q^+ \cap \mathcal{B}$ *such that* $\|W_q^*\|_1 \geq (1-\varepsilon)^2\|W_q^+\|_1 \cdot p$ *and each element of* W_q^* *is in a bucket in* $\mathcal{B}$ *containing no other element of* Y_1*;*

4) If $q \leq \log_2(\frac{N\varepsilon^2}{4p})$ *and* W_q^* *as in 3) exists, then with failure probability at most* $\exp(-m_1)$ *it holds that* $\sum_{i \in W_q^*} G(B_i) \geq (1-\varepsilon)\|W_q^*\|_1$*.*

Proof. 1) Note that $|Y_1| \leq \gamma_1^{-1}$ since $\|z\|_1 = 1$ and that we can view $|Y_1 \cap U|$ as a binomial random variable with parameters $|Y_1|$ and $p = \frac{M}{n}$. Thus, the expected number of elements of Y_1 at level h is bounded by $|Y_1| \cdot \frac{M}{n} \leq \frac{p}{\gamma_1} = 3m_1 \leq \frac{\varepsilon N}{4}$ since $N \geq 12m_1$ (see Assumption 3.1). Thus, we get by Lemma 2.14 that

$$\begin{aligned} P\left(|Y_1 \cap U| \geq \frac{\varepsilon N}{2}\right) &\leq P\left(|Y_1 \cap U| - |Y_1| \cdot p \geq \frac{\varepsilon N}{4}\right) \leq P\left(|Y_1 \cap U| - |Y_1| \cdot p \geq 3m_1\right) \\ &\leq \exp\left(-3m_1/3\right) \leq \exp(-m_1). \end{aligned}$$

2) For $i \in T = [n] \setminus Y_1$ we set $X_i = |z_i|$ if $i \in U$ and $X_i = 0$ otherwise. Since $\sum_{i \in T} |z_i| \leq \|z\|_1 = 1$ we have that $\mathbb{E}(\sum_{i \in T} X_i) = p \cdot \sum_{i \in T} |z_i| \leq p$. Since all 'large elements' are in Y_1 we have that $X_i < \gamma_1$ for all $i \in [n]$ and thus

$$\mathbb{E}\left(\sum_{i \in T} X_i^2\right) = \sum_{i \in T} p|z_i|^2 \leq \sum_{i \in T} p\gamma_1|z_i| = p\gamma_1 \sum_{i \in T} |z_i| \leq p\gamma_1.$$

Using Bernstein's inequality we get

$$P\left(\sum_{i \in T} X_i \geq 2p\right) \leq \exp\left(-\frac{p^2/2}{p\gamma_1 + p\gamma_1/3}\right) \leq \exp\left(-\frac{p}{3\gamma_1}\right) = \exp(-m_1).$$

This implies that $\sum_{i \in T} X_i \leq 2p$ with failure probability at most $\exp(-m_1)$. Now if $\sum_{i \in T} X_i \leq 2p$ then there can be at most $\frac{\varepsilon N}{2}$ buckets B with $G(B \setminus Y_1) \geq \frac{4p}{\varepsilon N}$.

3) First note that if $q \geq \log_2(\frac{8q_m\mu_z m_1}{\varepsilon^3 p})$ is important then $2^{-q} \cdot |W_q^+| \geq \|W_q^+\|_1 \geq \frac{\varepsilon}{q_m\mu_z}$, which implies that $|W_q^+| \geq \frac{2^q\varepsilon}{q_m\mu_z} \geq 8m_1\varepsilon^{-2} \cdot p^{-1}$. Assume that all entries of $Y_1 \setminus W_q^+$ have been assigned and let $\mathcal{B}' \subset \mathcal{B}$ be the buckets of $\mathcal{B}$ with no elements from $Y_1 \setminus W_q^+$. By 1) and 2) there are at least $(1-\varepsilon)N$ buckets in $\mathcal{B}'$. For $z_i \in W_q^+$ consider the random variable that takes the value $Z_i = z_i$ if $i \in \bigcup_{B \in \mathcal{B}'} B$ and $Z_i = 0$ otherwise.

Set $Z = \sum_{z_i \in W_q^+} Z_i$. We have $Z_i = z_i$ if element i is sampled at level h and sent to a bucket in $\mathcal{B}'$, which happens with probability at least $p \cdot \frac{(1-\varepsilon)N}{N} = (1-\varepsilon)p$. We thus have for the expected value of Z that

$$\begin{aligned}\mathbb{E}(Z) \geq (1-\varepsilon)p \cdot \|W_q^+\|_1 \geq (1-\varepsilon)p \cdot 2^{-q-1} \cdot |W_q^+| &\geq (1-\varepsilon) \cdot 2^{-q-1} \cdot 8m_1\varepsilon^{-2} \\ &\geq 2^{-q} \cdot 3m_1\varepsilon^{-2}.\end{aligned}$$

Further, the maximum value of any Z_i is 2^{-q} and the probability that $Z_i = z_i$ is upper bounded by p. Consequently, the variance of Z is bounded by

$$\sum_{z_i \in W_q^+} \mathbb{E}(Z_i^2) \leq \sum_{z_i \in W_q^+} p z_i^2 \leq 2^{-q} \sum_{z_i \in W_q^+} p z_i = 2^{-q}\mathbb{E}(Z).$$

Using Bernstein's inequality we get that

$$\begin{aligned}P\left(Z < (1-\varepsilon)^2 p \cdot \|W_q^+\|_1\right) &\leq P\left(Z - \mathbb{E}(Z) > \varepsilon\mathbb{E}(Z)\right) \\ &\leq \exp\left(\frac{-\varepsilon^2\mathbb{E}(Z)^2/2}{2^{-q}\mathbb{E}(Z) + 2^{-q}\varepsilon\mathbb{E}(Z)/3}\right) \\ &\leq \exp\left(\frac{-\varepsilon^2\mathbb{E}(Z)}{3 \cdot 2^{-q}}\right) \\ &\leq \exp\left(-m_1\right).\end{aligned}$$

We set $W_q^* = \{z_i \in W_q^+ \mid Z_i = z_i\}$.

4) By 2) and 3) we have that any entry $z_i \in W_q^*$ is in a bucket B with $\sum_{j \in B \setminus \{i\}} |z_j| \leq \frac{4p}{\varepsilon N}$. Thus, we have for $z_i \geq \frac{4p}{\varepsilon^2 N}$ that $\sum_{j \in B_i} z_j \geq z_i - \frac{4p}{\varepsilon N} \geq (1-\varepsilon)z_i$. Now we conclude

$$\sum_{i \in W_q^*} G(B_i) \geq \sum_{i \in W_q^*} (1-\varepsilon)z_i = (1-\varepsilon)\|W_q^*\|_1.$$

□

Note that if all buckets contain only a single element then we can remove the condition $q \leq \log_2(\frac{N\varepsilon^2}{4p})$. Hence, we can set $q_{(M,N)}(3) = q_{(M,N)}(4) = \infty$ if $N = M$ (respectively, $h = h_m$).

Next we are going to prove the outer bounds $q_h(1)$ and $q_h(4)$, which are important for the dilation bound. For the outer bounds, i.e., the borders of the interval of weight classes that can have a non-negligible contribution to U, we need the following parameters defining the set of small elements:

$$\begin{aligned}\gamma_2 &:= \frac{p\varepsilon^2}{3N\ln(Nh_{\max}/\delta)} \\ Y_2 &= \{i \in [n] \mid |z_i| \leq \gamma_2\}\end{aligned}$$

We further set E to be the expected value of an entry chosen uniformly at random from Y_2.

The following Lemma contains two parts. The first part shows that if the sum of all small entries of z (here the term 'small' is depending on the level) is negative and the absolute value is at least ε/n, or in other words if restrict to the small elements of z the negative terms outweight the positive parts at least slightly, then any bucket containing only those elements will have a negative value as well with high enough probability. The second part shows that with high probability the level does not contain any 'large' elements which is due to the fact that the number of large elements is limited.

Lemma 3.8. *The following hold:*

1) If $E \leq -\varepsilon/n$, then for any bucket B that contains only elements of Y_2, we have $G(B) = \sum_{i \in B} z_i \leq 0$ with failure probability at most $\frac{\delta}{N h_{\max}}$.

2) U contains no element i with $z_i \geq \frac{p h_{\max}}{\delta}$ with failure probability at most $\frac{\mu_z \delta}{h_{\max}}$.

Proof. 1) First consider a single bucket B containing only elements of Y_2. For $i \in [n]$, let X_i be a random variable that attains the value $X_i = z_i$ if $i \in B$ and $X_i = 0$ otherwise. The expected value of $G(B) = \sum_{i \in [n]} X_i$ is $E' := n \cdot \frac{p}{N} \cdot E \leq -\frac{p\varepsilon}{N}$. Further, we have that

$$\mathbb{E}\left(\sum_{i \in [n]} X_i^2\right) = \sum_{i \in Y_2} \frac{p}{N} \cdot z_i^2 \leq \gamma_2 \cdot \sum_{i \in Y_2} \frac{p}{N} \cdot |z_i| = \gamma_2 \frac{p}{N}$$

since all X_i are bounded by γ_2 by assumption. Thus, applying Bernstein's inequality yields

$$\begin{aligned}
P(G(B) > 0) \leq P\left(\sum_{i \in [n]} X_i - E' \geq |E'|\right) &\leq \exp\left(\frac{-|E'|^2/2}{\gamma_2 \frac{p}{N} + \gamma_2 |E'|/3}\right) \\
&\leq \exp\left(\frac{-\varepsilon \cdot p/(N)}{2\gamma_2(p/(N|E'|) + 1/3)}\right) \\
&\leq \exp\left(\frac{-\varepsilon \cdot p/(N)}{2\gamma_2(\varepsilon^{-1} + 1/3)}\right) \\
&\leq \exp\left(\frac{-\varepsilon^2 \cdot p/(N)}{3\gamma_2}\right) \\
&\leq \exp\left(-\ln\left(\frac{N h_{\max}}{\delta}\right)\right) \\
&= \frac{\delta}{N h_{\max}}.
\end{aligned}$$

2) Recall that $\sum_{z_i > 0} z_i \leq 1/\mu_z$. Thus, there are at most $\frac{n\delta}{\mu_z M h_{\max}}$ entries with $z_i \geq \frac{M h_{\max}}{n\delta}$. The expected number of those entries in U is thus at most $\frac{n\delta}{\mu_z M h_{\max}} \cdot \frac{M}{n} \leq \frac{\delta}{h_{\max}}$, which also upper bounds the probability of at least one entry with $z_i \geq \frac{M h_{\max}}{n\delta}$ being contained in U. □

Putting both lemmas together we get all bounds $q_h(i)$ except $q_0(2)$. Since the weight classes containing the heavy hitters are not necessarily large enough to get exponentially small probabilities we handle those in

the next subsection.

3.4.5 Heavy hitters

In this subsection we will analyze the level containing all entries. Our goal is to show that we can indeed set $q_0(2) = 0$ in Lemma 3.6. Let U be as before and assume that $M = n$. If for $q \geq \log_2(\frac{8q_m\mu m_1}{\varepsilon^3})$ the weight class W_q is important, we have seen that there is a W_q^* which represents W_q. Thus we only need to look at the remaining weight classes which can be important even though they do not contain enough elements to guarantee for a subset W_q^* to exist with high enough probability.

Let $Q_0 = \{q \leq \log_2(\frac{8q_m\mu m_1}{\varepsilon^3})\}$ be the the set of indices of weight classes containing only large elements. We set $H = \bigcup_{q\in Q_0} W_q$ to be the class of heavy hitters.

We first give properties of the ℓ_p leverage scores for $p \in [1, \infty)$. Note that the important cases for us in this section are $p = 1$ and in the later part of the section also $p = 2$.

We let $u \in \mathbb{R}^n_{\geq 0}$ denote the vector whose coordinates u_i denote the i-th ℓ_1-leverage scores, i.e., $u_i = \max_{\beta\in\mathbb{R}^d} \frac{|x_i\beta|}{\sum_{j\in[n]}|x_j\beta|}$. For $p \in [1, \infty)$ we set $G_p(z) = \sum_{i=1}^n |z_i|^p$

Lemma 3.9. *Let $u^{(p)} \in \mathbb{R}^n$ be the vector whose coordinates $u_i^{(p)}$ denote the i-th ℓ_p-leverage scores, i.e., $u_i = \max_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{|x_i\beta|^p}{\sum_{j\in[n]}|x_j\beta|^p}$. If $u_i^{(p)}$ is the k-th largest coordinate of $u^{(p)}$, then for z in the subspace spanned by the columns of X it holds that $|z_i|^p \leq \frac{d^p}{k} G_p(z)$ and $\sum_{i=1}^n u_i^{(p)} = d^p$. In particular we have that $|z_i| \leq \frac{d}{k} G(z)$ and $\sum_{i=1}^n u_i = d$. If $p = 2$ then it holds that $|z_i| \leq \frac{d}{k} G_2(z)$ and $\sum_{i=1}^n u_i^{(2)} = d$.*

Proof. By Dasgupta et al. (2009) there exists a so-called *Auerbach* basis Q of A with the following properties. It holds that $\|Qx\|_p \geq \|x\|_q$ for all $x \in \mathbb{R}^d$ where q is the dual norm of p and $\sum_{ij} |Q_{ij}|^p \leq d$. Note that by a change of basis

$$u_i = \max_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|(Ax)_i|^p}{\|Ax\|_p^p} = \max_{x\in\mathbb{R}^d\setminus\{0\}} \frac{|(Qx)_i|^p}{\|Qx\|_p^p}.$$

Thus $|z_i|^p = |Q_i x|^p \leq (\|Q_i\|_p \|x\|_q)^p \leq (\|Q_i\|_p \|Qx\|_p)^p$ and it follows that $\sum_i u_i \leq \sum_i \|Q_i\|_p^p = \sum_{ij} |Q_{ij}|^p \leq d^p$.

Consequently the k-th largest coordinate of z can be at most $|z_p| \leq u_p G(Qx) \leq \frac{d}{k} G(Qx) = \frac{d}{k} G(z)$. For $p = 2$ an orthonormal basis Q fulfills $\|Qx\|_p \geq \|x\|_q$ for all $x \in \mathbb{R}^d$ and $\sum_{ij} |Q_{ij}|^p \leq d$. Thus using the same proof as above gives us the desired result for $p = 2$. □

Lemma 3.10. *Let $Y_3 = \{i \mid u_i \geq \gamma_3\}$ where $\gamma_3 = \frac{\varepsilon^3}{8q_m\mu m_1}$. Further, for $j \in Y_3$ let $\mathcal{C}_j = \{B \mid \sum_{i\in B\setminus\{j\}} u_i \geq \varepsilon\gamma_3\}$. Let $Y_3' = \{j \in Y_3 \mid j \in \mathcal{B}_0 \setminus \mathcal{C}_j\}$ If $N_0 \geq \max\{\frac{2d^2\mu}{\gamma_3\varepsilon^2}, \frac{2\ln(\kappa^{-1})d^2}{\gamma_3\varepsilon\kappa}\}$ for $\kappa \in (0, 1/2)$, then with probability $1 - 2\kappa$, it holds that $\sum_{j\in Y_3\setminus Y_3'} u_j \leq 2\varepsilon\mu^{-1}$.*

Proof. We split the set Y_3 into two parts. One contains the elements with the largest leverage scores and the other one consisting of the remaining elements with smaller leverage scores. Consider the set

$L_0 = \{j \in Y_3 \mid u_j \geq \gamma\}$ where $\gamma = \frac{\varepsilon}{\mu \ln(\kappa)}$. By Lemma 3.9 it holds that $\sum_{i=1}^{n} u_i \leq d$ and thus there can be at most d/γ elements in L_0 and $\frac{d}{\varepsilon\gamma_3}$ buckets B with $\sum_{i\in B} u_i \geq \varepsilon\gamma_3$. Thus the probability of any element $j \in L_0$ being assigned to a bucket of C_j is bounded by $\frac{d}{\varepsilon\gamma_3 N_0} \cdot \frac{d}{\gamma} \leq \kappa$.

Let $L_1 = \{j \in Y_3 \mid u_j < \gamma\}$. For $j \in L_1$ let X_j be the random variable with $X_j = u_j$ if j is placed in a bucket in C_j and $X_j = 0$ else. We have that the probability that $X_j = u_j$ is at most $\frac{d}{\gamma_3\varepsilon N_0}$ and that $|L_1| \leq \frac{d}{\gamma_3}$. Thus using Lemma 3.9 we have that

$$\mathbb{E}(\sum_{j\in L_1} X_j) \leq \sum_{j\in L_1} u_j \cdot \frac{d}{\gamma_3\varepsilon N_0} = \frac{d^2}{\gamma_3\varepsilon N_0} \leq \varepsilon\mu^{-1}.$$

Further we have that

$$\mathbb{E}(\sum_{j\in L_1} X_j^2) \leq \sum_{j\in L_1} u_j^2 \cdot \frac{d}{\gamma_3\varepsilon N_0} \leq \sum_{j\in L_1} u_j \cdot \frac{d\gamma}{\gamma_3\varepsilon N_0} \leq \frac{d^2\gamma}{\gamma_3\varepsilon N_0} \leq \varepsilon\mu^{-1}\gamma/2.$$

using the fact that $d^2/(\varepsilon N_0) \leq \varepsilon\gamma/2$. Further using $\ln(\kappa) \geq \varepsilon\mu^{-1}\gamma$ and Bernstein's inequality we have that

$$\Pr(\sum_{j\in L_1} X_j \geq 2\varepsilon\mu^{-1}) \leq \exp\left(-\frac{\varepsilon^2\mu^{-2}}{\varepsilon\mu^{-1}\gamma/2 + \varepsilon\mu^{-1}\gamma/3}\right) \leq \exp(-\varepsilon\mu^{-1}\gamma) \leq \exp(\ln(\kappa)) = \kappa.$$

□

We apply Lemma 3.10 with $\kappa = \delta$. We denote by $\mathcal{E}_1$ the event that $\sum_{j\in Y_3\setminus Y_3'} u_j \geq 2\varepsilon$. By Lemma 3.10 $\mathcal{E}_1$ holds with probability at least $1-\delta$ for an appropriate $N = N_0 = \max\{\frac{d^2\mu}{\gamma_3\varepsilon^2}, \frac{2\ln(\delta)d^2}{\gamma_3\varepsilon\delta}\} = \frac{d^2 q_m \mu^2 m_1}{\delta\varepsilon^5} = \mathcal{O}(\frac{d^2 q_m \mu m_1}{\delta\varepsilon^5})$. For any entry $z_i \in H$ we have $z_i \geq \gamma_3$ and thus by Lemma 3.9, we have $i \in Y_3$. It remains to show that the remaining entries in the buckets containing a heavy hitter only have a small contribution.

Lemma 3.11. *Assume $\mathcal{E}_1$ holds. For any $i \in H \cap Y_3'$ we have $G(B_i) \geq (1-\varepsilon)z_i$. Further it holds that $\sum_{i\in Y_3\setminus Y_3'} |z_i| \leq 2\varepsilon\mu^{-1}$.*

Proof. Let $z_i \in H \cap Y_3'$. By $\mathcal{E}_1$ we have that $\sum_{j\in B\setminus\{i\}} u_j \leq \varepsilon\gamma_3 \leq \varepsilon z_i$. We conclude that

$$G(B_i) \geq z_i - \sum_{j\in B_i\setminus\{i\}} |z_j| \geq z_i - \sum_{j\in B_i\setminus\{i\}} u_j \geq z_i - \varepsilon z_i \geq (1-\varepsilon)z_i.$$

The second part of the claim follows as $|z_j| \leq u_j$ for any $j \in [n]$. □

Heavy hitters - alternative version There is another way of handling heavy hitters. Using it we can reduce the sketch size at the cost of running time. The idea is that each row gets sampled multiple times. We will look into two versions here with a trade off between running time and sketch size.

In the first version we replace level 0 by the following sketch: We have h_0 sub levels $0.1, \ldots 0.h_0$ for $h_0 = \lceil \frac{3\ln(\gamma_3^{-1}\delta^{-1}))}{\varepsilon} \rceil$. Each sub level consists of $N \geq N_0'' := 6\gamma_3^{-1}d\varepsilon^{-1} \geq 6|Y_3|d\varepsilon^{-1} = O(\frac{dm_1\mu q_m}{\varepsilon^4})$ buckets. Now each row i gets mapped to exactly one bucket of each sub level. As a consequence we no longer need to

guarantee that all heavy hitters are separated from all other big elements but instead we guarantee that for any heavy hitter z_i there exist at least $(1-\varepsilon)h_0$ sub levels where z_i is in a bucket with no other big element. To compensate each row being in multiple buckets we set the weight of each bucket to be $\frac{1}{h_0}$.

We denote by $B_i(\ell)$ the bucket of sub level ℓ containing z_i.

Lemma 3.12. *With failure probability at most δ the event $\mathcal{E}_1'$ holds, that for any $i \in Y_3$, there exists a set $L \subseteq [h_0]$ of at least $(1-\varepsilon)h_0$ sub levels $\ell \in [h_0]$ such that the bucket $B_i(\ell) \notin \mathcal{C}_j = \{B \mid \sum_{i \in B \setminus \{j\}} u_i \geq \varepsilon\gamma_3\}$.*

Proof. Fix $i \in Y_3$. Since there are $N = 2|Y_3|d\varepsilon^{-1}$ using Lemma 3.10 there can be at most a fraction of $1/2$ of the buckets at each sub level in C_i. Let X_j be the random variable with $X_j = 1$ if $B_i(j) \in C_j$ and $X_j = 0$ otherwise. Note that X_j is a Bernoulli random variable with $p = \varepsilon$. Thus we can apply the Chernoff bound to $\sum_{j=1}^{h_0} X_j$:

$$P\left(\sum_{j=1}^{h_0} X_j > \varepsilon h_0\right) \leq \exp(-h_0\varepsilon/6) \leq \frac{\delta}{|Y_3|}$$

and thus there exists a set $L \subseteq [h_0]$ of at least $(1-\varepsilon)h_0$ sub levels $\ell \in [h_0]$ such that the bucket $B_i(\ell) \notin \mathcal{C}_j$. Now using the union bound we get that with failure probability at most δ this holds for all $i \in Y_3$ that i. □

Then we get using the same proof as for Lemma 3.11:

Lemma 3.13. *Assume $\mathcal{E}_1'$ holds. Then for any element $i \in H$ it holds that $\sum_{\ell=1}^{h_0} G^+(B_i(\ell)) \geq (1-2\varepsilon)h_0 z_i$.*

Heavy hitters - alternative version 2 The idea of the second alternative version is that each row gets sampled multiple times as in the alternative version but this time we do have different sub levels but instead just sample each element multiple times. More precisely, each row i gets sampled in $s = 8m_1q_m\mu/\varepsilon^2$ buckets at level 0. Technically we are getting rid of heavy hitters this way since the maximum leverage score in the instance created this way is at most $1/s$ as we are sketching $\tilde{X}$ where each row of X appears s times and thus each entry of $\tilde{z} = \tilde{X}\beta$ appears at least s times. To compensate the fact that each element appears multiple times, we set the weight of buckets of level 0 to $w_0 = 1/s$.

3.4.6 Contraction bounds for a single point

We set U_h to be the rows z_i sampled at level h. Combining previous subsections we get the following lemma:

Lemma 3.14. *Assume that $\mathcal{E}_1$ holds. Denote by z_i' the i-th row of $SX\beta$ for $i \in n'$. Then with failure probability at most $(2h_m + 2q_m)e^{-m_1}$ it holds that*

$$\sum_{i \in n', z_i' \geq 0} w_i z_i' \geq (1-6\varepsilon)\|(X\beta)^+\|_1.$$

Proof. By Lemma 3.7 and Lemma 3.11 we have that for each important weight class W_q^+ there exists a subset $W_q^* \subseteq U_h$ with $\sum_{i \in W_q^*} G(B_i) \geq (1-\varepsilon)^2 \|W_q^+\|_1 p_h$ with failure probability at most $(2h_m + 2q_m)e^{-m_1}$. For $q \in Q_H$ we can set $W_q^* = W_q^+ \cap Y_3'$. Note that it holds that $\sum_{i \in Y_3 \setminus Y_3'} |z_i| \leq 2\varepsilon\mu^{-1}$ by Lemma 3.11. Then using Lemma 3.5 we get

$$\begin{aligned}
\sum_{i \in n', z_i' \geq 0} w_i z_i' &\geq \sum_{q \in Q^*} p_h^{-1} \sum_{i \in W_q^*} G(B_i) \\
&\geq \sum_{q \in Q^*} (1-\varepsilon)^2 \|W_q^+\|_1 - 2\varepsilon\mu^{-1} \\
&\geq (1-2\varepsilon)(1-\varepsilon)^2 \|(X\beta)^+\|_1 - 2\varepsilon\mu^{-1} \geq (1-6\varepsilon)\|(X\beta)^+\|_1.
\end{aligned}$$

□

3.4.7 Net argument

To get a weak weighted sketch we need the contraction bounds not just for a single solution but for all $\beta \in \mathbb{R}^d$. Thus we will construct a net $\mathcal{N}$ such that if a slightly stronger contraction bound holds for any $\beta \in \mathcal{N}$ then the contraction bound holds for any $\beta \in \mathbb{R}^d$. For now we ignore the variance regularization and focus only on f_1, i.e., on plain logistic regression. We first show that if the distance of two vectors $v, v' \in \mathbb{R}^n$ is small then $|f_1(v) - f_1(v')|$ is also small.

The next two lemmas show that if the contraction bound holds for some β then it holds for any β' close to β.

Lemma 3.15. *For any $v, v' \in \mathbb{R}^n$ with $\|v - v'\|_1 \leq \varepsilon$ it holds that $|f_1(v) - f_1(v')| \leq \varepsilon$.*

Proof. Since $\ell'(v) = \frac{e^v}{e^v+1} \leq 1$ we get that

$$|f_1(v) - f_1(v')| \leq \sum_{i=1}^n |\ell(v_i) - \ell(v_i')| \leq \sum_{i=1}^n |v_i - v_i'| = \|v - v'\|_1$$

which proves the lemma. □

Lemma 3.16. *Assume that for $\beta \in \mathbb{R}^d$ it holds that $|f_1(X'\beta) - f_1(X\beta)| \leq \varepsilon$. Then for any $\beta' \in \mathbb{R}^d$ with $\|X\beta - X\beta'\|_1 \leq \varepsilon/(b^{h_m} h_m)$ it holds that $|f_1(X\beta') - f_1(X'\beta')| \leq 3\varepsilon$.*

Proof. It holds that $\|X'(\beta - \beta')\|_1 = \|SX(\beta - \beta')\|_1 \leq b^{h_m} h_m \|X(\beta - \beta')\|_1 \leq \varepsilon$ since for each $i \in [n]$ there are at most h_m columns j such that $S_{ij} \neq 0$ and each entry of S is bounded by b^{h_m}. Thus, using the triangle inequality and applying Lemma 3.15 yields

$$\begin{aligned}
|f_1(X\beta') - f_1(X'\beta')| &\leq |f_1(X'\beta') - f_1(X'\beta)| + |f_1(X'\beta) - f_1(X\beta)| + |f_1(X\beta) - f_1(X\beta')| \\
&\leq \varepsilon + \varepsilon + \varepsilon \leq 3\varepsilon.
\end{aligned}$$

□

We are now ready to construct our net:

Lemma 3.17. *There exists a net $\mathcal{N} \subset \mathbb{R}^d$ of size $|\mathcal{N}| = \exp(\mathcal{O}(d\ln(n)))$ such that for any point $y \in \mathbb{R}^d$ with $\|Xy\|_1 \leq n\mu$ there exists a point $y' \in \mathcal{N}$ such that $\|Xy' - Xy\|_1 \leq \frac{\varepsilon}{\mu b^{h_{\max}} h_m}$.*

Proof. We set

$$\mathcal{N} = \left\{ \beta = v \cdot \frac{\varepsilon}{db^{h_m} h_m} \mid v \in \mathbb{Z}^d \text{ with } \|v\|_\infty \leq \frac{dn\mu b^{h_{\max}} h_m}{\varepsilon} \right\}. \tag{14}$$

Then for any $y \in \mathbb{R}$ with $\|Xy\|_1 \leq n\mu$ the point $Xy' = \lfloor \frac{db^{h_m} h_m}{\varepsilon} \cdot Xy \rfloor \cdot \frac{\varepsilon}{db^{h_m} h_m}$ is in $\mathcal{N}$ and it holds that $\|Xy - Xy'\|_1 \leq d \cdot \frac{\varepsilon}{db^{h_m} h_m} = \frac{\varepsilon}{b^{h_m} h_m}$. Further we have $|\mathcal{N}| \leq \left(\frac{d^2 n\mu b^{2h_m} h_m^2}{\varepsilon^2}\right)^d = \exp(\mathcal{O}(d\ln(n))$. □

Combining Lemma 3.16 and Lemma 3.17 we get:

Lemma 3.18. *There exists a net $\mathcal{N} \subset \mathbb{R}^d$ with $|\mathcal{N}| = \exp(\mathcal{O}(d\ln(n)))$ such that if $|f_1(X'\beta) - f_1(X\beta)| \leq \varepsilon$ holds for any $\beta \in \mathcal{N}$, then for any $\beta' \in \mathbb{R}^d$ with $\|X\beta'\|_1 \leq n\mu$ it holds that $|f_1(X'\beta') - f_1(X\beta')| \leq 3\varepsilon$.*

3.4.8 Dilation bounds

In this subsection we prove the dilation bounds. More precisely we will will show that for any weight class W_q the expected contribution to $\|(X'\beta)^+\|$ is bounded by $k\|W_q\|_1$ for some k.

First we show that for any good $\beta \in \mathbb{R}$, i.e. β with $f_1(X\beta) \leq n\ln(2)$, assumption of Lemma 3.8 1) is fulfilled. Given $\beta \in \mathbb{R}^d$ and $z = X\beta$ set $Z_0 = Z_0(\beta) \subset Z = \{z_1, \ldots, z_n\}$ to be the set of the $(1-\varepsilon)n$ largest entries ordered by absolute value. In other words, we remove the εn smallest entries. Similarly we set $Z_1 = Z_1(\beta) \subset Z$ to be the set of the $(1-2\varepsilon)n$ largest entries. Again we assume that $\|z\|_1 = 1$. Our next goal is to show that if $f_1(z)$ is small then $\sum_{z_i \in Z_0} z_i$ remains negative even if we remove the smallest entries. Here small means negative with large absolute value.

Lemma 3.19. *If $f_1(X\beta) < (1-2\varepsilon) f_1(0)$ then it holds that*

$$\sum_{z_i \in Z_0, z_i \leq 0} |z_i| \geq (1+\varepsilon) \sum_{z_i \geq 0} |z_i|$$

Proof. Let X_1 denote the matrix X where the columns not corresponding to an entry of Z_1 are removed. We denote by $\tilde{f}_1$ the function nf_1 restricted to $|Z_1|$ entries, i.e., $\tilde{f}_1(X\beta) = \sum_{x_i \in X_1} \ell(x_i\beta)$. Since ℓ is always larger than 0, removing $2\varepsilon n$ entries can only reduce nf_1. We thus have that

$$\tilde{f}_1(0) = (1-2\varepsilon) n f_1(0) \geq n f_1(X\beta) = n f_1(Z) \geq \tilde{f}_1(Z_1).$$

Now consider the function $\varphi(r) = \tilde{f}_1(r \cdot X\beta)$. Note that the derivative of φ at zero is given by $\varphi'(0) = \sum_{x_i \in X_1} \frac{e^0}{e^0+1} \cdot x_i\beta = \frac{1}{2} \cdot \sum_{z_i \in Z_1} z_i$. Since $\tilde{f}_1$ is convex φ is also convex. In particular this means that

$\tilde{f}_1(X\beta) < \tilde{f}_1(0)$ implies $\varphi'(0) < 0$. Thus it must hold that $\sum_{z_i \in Z_1} z_i < 0$, or equivalently, $\sum_{z_i \in Z_1, z_i < 0} |z_i| > \sum_{z_i \in Z_1, z_i > 0} |z_i|$. Since all entries in $Z_0 \setminus Z_1$ are less than or equal to any entry in Z_0, we have that

$$\sum_{z_i \in Z_0, z_i < 0} |z_i| \geq \frac{1}{1-\varepsilon} \sum_{z_i \in Z_1, z_i < 0} |z_i| \geq (1+\varepsilon) \sum_{z_i > 0} |z_i|.$$

□

Next we show how we can use the outer bound $q_h(1)$ and $q_h(4)$ to bound the expected contribution of W_q by proving that the expected contribution of W_q at any level with $q_h(1) < q < q_h(4)$ is bounded by $\|W_q\|_1$ and by 0 otherwise. The following lemma gives us an upper bound on the expected value of $G^+(Z)$.

Lemma 3.20. *If for all $i \leq h_m - 1$ it holds that $q_{(M_i N_i)}(4) < q_{(M_{i+k} N_{i+k})}(1)$ and $N_0 \geq N_0'$, then the expected contribution of any weight class W_q is at most $k \cdot \|W_q\|_1$.*

Proof. Consider a weight class W_q^+. For any level h it follows by Lemma 3.6 that if $q \notin [q_{(M_h N_h)}(1), q_{(M_h N_h)}(4)]$ then W_q^+ has zero contribution at level h, i.e., either there are no elements of W_q^+ at level h or we have $W_q^+ \subset Y_1$ and for any bucket B of level h it holds that $\sum_{i \in Y_1 \cap B} z_i \leq 0$. At any level the expected contribution of W_q^+ is bounded by $p_h^{-1} \cdot \sum_{i \in W_q^+} p_h z_i = \|W_q^+\|_1$. This upper bound would be tight if all entries of Z were positive. Hence, the expected contribution of W_q^+ is upper bounded by the number of levels h with $q \in [q_{(M_h N_h)}(1), q_{(M_h N_h)}(4)]$. Since $q_{(M_h N_h)}(1)$ and $q_{(M_h N_h)}(4)$ are monotonically increasing in h, it follows that if $q_{(M_i N_i)}(4) < q_{(M_{i+k} N_{i+k})}(1)$ then any q can be contained in at most k intervals of the form $[q_{(M_h N_h)}(1), q_{(M_h N_h)}(4)]$, concluding the lemma. See Figure 2 for an illustration. □

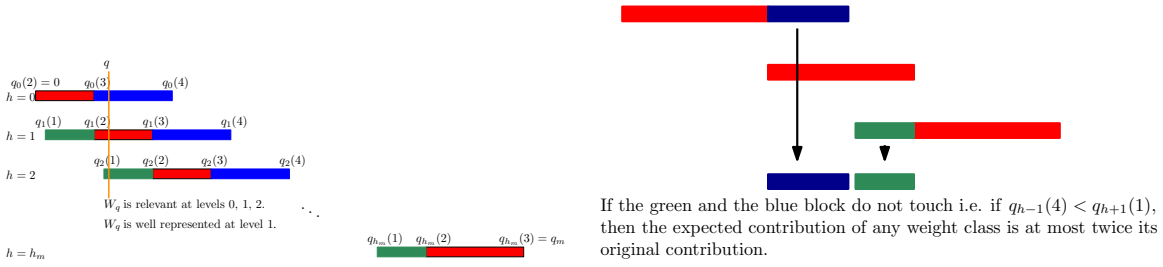

Figure 2: Illustration of Lemma 3.20 and Lemma 3.21.

Lemma 3.20 can be used to show that the expected contribution of any weight class to $G^+(Z)$ is at most twice its total weight:

Lemma 3.21. *If we choose $N_i = N := \max\{N_0', \frac{2048 m_1^2 \mu \ln(N h_m/\delta) q_m^2 h_m}{\varepsilon^{12} \delta}\}$ for all $i \in [h_m]$, and M_i solving the equation $q_{(M_{i-1}, N)}(3) = q_{(M_i, N)}(2)$ then the expected contribution of any weight class W_q is at most $2\|W_q\|_1$.*

Proof. We set $q_i(j) = q_{(M_i N_i)}(j)$. We first show that $q_{(i+2)}(1) - q_i(4)$ can be expressed using the terms $q_{i+1}(3) - q_{i+1}(2)$, $(q_{i+2}(2) - q_{i+2}(1))$ and $(q_i(4) - q_i(3))$, which are the same for each i if the number of

buckets at each level is identical, i.e., for all $j \leq h_q$ it holds that $N_j = N_i$. Observe that

$$\begin{aligned} q_{(i+2)}(1) - q_i(4) &= q_{i+2}(2) + q_{i+2}(1) - q_{i+2}(2) - (q_i(3) + q_i(4) - q_i(3)) \\ &= q_{i+2}(2) - q_i(3) - (q_{i+2}(2) - q_{i+2}(1)) - (q_i(4) - q_i(3)) \\ &= q_{i+1}(3) - q_{i+1}(2) - (q_{i+2}(2) - q_{i+2}(1)) - (q_i(4) - q_i(3)). \end{aligned}$$

Figure 2 illustrates those three terms. Using Lemma 3.6 we can bound the sum of the two subtracted terms by

$$\begin{aligned} (q_{i+2}(2) - q_{i+2}(1)) + (q_i(4) - q_i(3)) &= \log_2\left(\frac{8q_m m_1 h_m}{\varepsilon^3 \delta}\right) + \log_2\left(\frac{8\ln(Nh_m/\delta)}{\varepsilon^3}\right) \\ &= \log_2\left(\frac{64 m_1 \ln(Nh_m/\delta) q_m h_m}{\varepsilon^7 \delta}\right). \end{aligned}$$

By Lemma 3.6 we have that $q_{i+1}(3) - q_{i+1}(2) \geq \log_2\left(\frac{N\varepsilon^5}{32 m_1 \mu q_m}\right)$. Thus, combining both equations we get that

$$\begin{aligned} q_{(i+2)}(1) - q_i(4) &= \log_2\left(\frac{N\varepsilon^5}{32 m_1 \mu q_m}\right) - \log_2\left(\frac{64 m_1 \ln(Nh_m/\delta) q_m h_m}{\varepsilon^7 \delta}\right) \\ &= \log_2\left(\frac{N\varepsilon^{12}\delta}{2048 m_1^2 \mu \ln(Nh_m/\delta) q_m^2 h_m}\right). \end{aligned}$$

If $N \geq \frac{2048 m_1 \mu \ln(Nh_m/\delta) q_m^2 h_m}{\varepsilon^{12}\delta}$ then we have $q_{(i+2)}(1) - q_i(4) \geq 0$ and thus by Lemma 3.20, the expected contribution of any weight class W_q is at most $2\|W_q\|_1$. □

If $N < \frac{2048 m_1^2 \mu \ln(Nh_m/\delta) q_m^2 h_m}{\varepsilon^{12}\delta}$ we have the following adaptation of previous lemma:

Lemma 3.22. *If for some $k \in \mathbb{N}$ we choose $N_i = N \geq \frac{32 m_1 \mu q_m}{\varepsilon^5} \cdot \left(\frac{64 m_1 \ln(Nh_m/\delta) q_m h_m}{\varepsilon^7 \delta}\right)^{1/(k-1)}$ for all $i \in [h_m]$, and M_i solving the equation $q_{(M_{i-1},N)}(3) = q_{(M_i,N)}(2)$, then the expected contribution of any weight class W_q is at most $k\|W_q\|_1$.*

Proof. We generalize the proof of Lemma 3.21. We can substitute $q_{(i+k)}(1) - q_i(4)$ as follows:

$$\begin{aligned} q_{(i+k)}(1) - q_i(4) &= q_{(i+k)}(1) - q_{(i+k)}(2) + q_i(3) - q_i(4) + q_{(i+k)}(2) - q_i(3) \\ &= q_{(i+k)}(2) - q_i(3) - (q_{(i+k)}(2) - q_{(i+k)}(1)) - (q_i(4) - q_i(3)) \\ &= q_{(i+k-1)}(3) - q_{i+1}(2) - (q_{(i+k)}(2) - q_{(i+k)}(1)) - (q_i(4) - q_i(3)) \\ &= \sum_{j=1}^{k-1} q_{(i+j)}(3) - q_{i+j}(2) - (q_{(i+k)}(2) - q_{(i+k)}(1)) - (q_i(4) - q_i(3)). \end{aligned}$$

The difference to the proof of Lemma 3.21 is the telescoping sum. We have that

$$\sum_{j=1}^{k-1} q_{(i+j)}(3) - q_{i+j}(2) = (k-1) \cdot \log_2\left(\frac{N\varepsilon^5}{32m_1\mu q_m}\right) = \log_2\left(\left(\frac{N\varepsilon^5}{32m_1\mu q_m}\right)^{k-1}\right).$$

Thus if $N \geq \frac{32m_1\mu q_m}{\varepsilon^5} \cdot \left(\frac{64m_1 \ln(Nh_m/\delta)q_m h_m}{\varepsilon^7\delta}\right)^{1/(k-1)}$ we have that

$$\sum_{j=1}^{k-1} q_{(i+j)}(3) - q_{i+j}(2) \geq \log_2\left(\frac{64m_1 \ln(Nh_m/\delta)q_m h_m}{\varepsilon^7\delta}\right).$$

Further note that $(q_{i+2}(2) - q_{i+2}(1)) + (q_i(4) - q_i(3)) = \log_2\left(\frac{64m_1 \ln(Nh_m/\delta)q_m h_m}{\varepsilon^7\delta}\right)$ as before. We conclude that $q_{(i+k)}(1) - q_i(4) > 0$. Consequently, applying Lemma 3.20 finishes the proof. □

Next we want to show how we can reduce the expected contribution of all weight classes below $2\|W_q\|_1$. To this end we first increase the number of buckets at each level so as to get

$$\log_2\left(\frac{N\varepsilon^5}{32m_1\mu q_m}\right) \geq k \log_2\left(\frac{64m_1 \ln(Nh_m/\delta)q_m h_m}{\varepsilon^7\delta}\right).$$

Note that the expected contribution of any important weight class W_q^+ is at least $\|W_q^+\|_1$. Moreover, the above choice ensures that all but a k-th fraction of weight classes have an expected contribution of exactly $\|W_q^+\|_1$, and only the remaining k-th fraction has a larger expected contribution that crucially is still bounded by $2\|W_q^+\|_1$. Then the last step is to add a random shift so that the probability of each weight class W_q^+ for having an expected contribution of $2\|W_q^+\|_1$ is at most $\frac{1}{k}$. To simplify notation we set $N_1' = \frac{32m_1\mu q_m}{\varepsilon^5}$ and $N_2' = \frac{64m_1 \ln(n)q_m h_m}{\varepsilon^7\delta}$ and assume that $n \geq N^k h_m/\delta$.

Lemma 3.23. *Let $\gamma = \frac{1}{k} < 1$ for some $k \in \mathbb{N}$. Assume that N_0 is chosen uniformly at random from $N^{(1)}, \dots N^{(1/\gamma)}$ where $N^{(i)} = N_0' \cdot N_2'^i$. Further let $N_i = N = N_1' \cdot N_2'^{k+1}$ for any $i > 0$. Then the expected contribution of any weight class W_q^+ is at most $(1+\gamma)\|W_q^+\|_1$.*

Proof. First note that

$$\log_2\left(\frac{N\varepsilon^5}{32m_1\mu q_m}\right) - k \log_2\left(\frac{64m_1\mu \ln(n)q_m h_m}{\varepsilon^7\delta}\right) = \log_2(N/N_1') - \log_2(N_2'^k) \geq 0.$$

This shows that the relation of weight classes that are relevant on two levels to the weight classes that are relevant on only one level is $1 : k$. By choosing N_0 at random we introduce a shift by $i \log_2(N_2')$, which is the maximal length of a block $[q_{i-1}(1), q_i(4)]$. Hence, for each $q \in \mathbb{N}$ there can be only one i such that q is relevant in two levels. This implies that the expected contribution of W_q^+ is at most $\frac{k-1}{k} \cdot \|W_q^+\|_1 + \frac{1}{k} \cdot 2\|W_q^+\|_1 = (1 + \frac{1}{k})\|W_q^+\|_1$. □

3.5 Main result

Our main result is the following:

Theorem 1. *Let $X \in \mathbb{R}^{n\times d}$ be a μ-complex matrix for bounded $\mu < n$. Let $\varepsilon, \delta > 0$ and let $a > 1$ and let $\mathrm{nnz}(X)$ be the number of non zero entries of X. Then there is a distribution over sketching matrices $S \in \mathbb{R}^{r\times n}$ and a corresponding weight vector $w \in \mathbb{R}^r$, for which $X' = SX$ can be computed in T time in a single pass over a turnstile data stream such that (X', w) is a weak weighted $(\mathbb{R}^d, \alpha, \varepsilon)$-sketch for f_1 with failure probability at most P, where*

1. *$r = O(\mu^2 d^{1+c}\ln(n)^{2+4c})$ for any constant $c > 0$, $T = O(d\ln(n)\mu \cdot \mathrm{nnz}(X))$, and α and P are constant,*
2. *$r = O(\frac{\mu^2 d^3 \ln(n)^9}{\varepsilon^{11}\delta})$, $T = O(\mathrm{nnz}(X))$, $\alpha = 1 + (1+\varepsilon)a$ and $P = \delta + \frac{1}{a}$;*
3. *$r = O(\frac{d^3\ln(n)^5\mu^2}{\delta\varepsilon^6}) + \frac{32d\mu\ln(n)^2}{\varepsilon^5} \cdot (\frac{64d\ln(n)^4}{\varepsilon^7\delta})^{1+\varepsilon^{-1}}$, $T = O(\mathrm{nnz}(X))$, $\alpha = (1+a\varepsilon)$, and $P = \delta + \frac{1}{a}$.*

Proof. If $\beta = 0$ is a $1-2\varepsilon$ approximation, then we get the dilation bounds for free since $f_{1w}(X'\beta) = \ln(2) = f_1(X\beta)$. Otherwise let β^* be the minimizer of $f_1(X\beta)$. Note that β^* satisfies the assumption of Lemma 3.19.

1) We fix constants $\varepsilon = 1/8$ and $\delta = 1/8$.

We use the second alternative approach for handling heavy hitters and define M_i and N_i as in Lemma 3.22 for some constant $k = 1 + \frac{1}{c}$ and set $h_m = \min\{i \mid M_i \le N\}$.

By Lemma 3.22 the expected contribution of any weight class is at most $k\|W_q\|_1$. Thus using Markov's inequality we can bound $f_{1w}(SX\beta^*) \le akf_1(X\beta^*)$ with probability $\frac{1}{a}$ for any $a \in \mathbb{N}$. In other words, it is constant with constant probability. By our choice of M_i and N_i, the contraction bounds hold for any $X\beta$ with failure probability at most $(2h_m + 2q_m + 2)e^{-m_1}$ by combining Lemma 3.14 and Lemma 3.4. Setting $m_1 = \mathcal{O}(d\ln(n))$ and using Lemma 3.18 we get that the contraction bounds hold for all $\beta \in \mathbb{R}^d$ with $\|X\beta\|_1 \le n\mu$. We note that the contraction bounds can be extended to any $\beta \in \mathbb{R}^d$ since $f_1(X\beta) \approx \|X\beta\|_1$ if $\|X\beta\|_1 > n\mu$. We refer to (Munteanu et al., 2021) for details. Further note that $q_i(2) < q_i(3)$, and thus $h_m \le \log_2(2^{q_m}) = O(\ln(n))$. The number of buckets at each level is $N = \frac{32m_1\mu q_m}{(1/8)^5} \cdot (\frac{64m_1\ln(Nh_m/\delta)q_m h_m}{(1/8)^7})^c$. We specify the number r of rows of SX, which is $r = h_m N$. Since $h_m, q_m = O(\ln(n))$ and $m_1 = \mathcal{O}(d\ln(n))$ we get that $r = O(\mu^2 d^{1+c}\ln(n)^{2+4c})$. The running time of our algorithm is $O(\mu d\ln(n)\mathrm{nnz}(X))$ since each row x_i gets assigned to $O(\mu d\ln(n))$ buckets.

2) Next we show that with $r = O(\frac{\mu^2 d^4\ln(n)^7}{\varepsilon^{12}\delta})$ and $T = O(\mathrm{nnz}(X))$ we can get an approximation factor of $\alpha = 1 + (1+\varepsilon)a$ and failure probability of $P = \delta + \frac{1}{a}$. There are only a few differences compared to the proof of the first part: instead of Lemma 3.22 we use Lemma 3.21. Hence we need the number of buckets to be

$$N = \max\left\{N_0', \frac{2048m_1^2\mu^2\ln(Nh_m/\delta)q_m^2 h_m}{\varepsilon^{12}\delta}\right\} = \frac{2048m_1^2\mu^2\ln(Nh_m/\delta)q_m^2 h_m}{\varepsilon^{12}\delta}.$$

Consequently we have that $r = h_m N = O(\frac{\mu^2 d^4\ln(n)^7}{\varepsilon^{12}\delta})$. Since every row gets assigned to $O(1)$ buckets the running time is $O(\mathrm{nnz}(X))$. Now assume that the contraction bound holds for β^*. Then $Y = f_{1w}(SX\beta^*) -$

$(1-\varepsilon)f_1(X\beta^*)$ is a positive random variable with expected value at most $(1+\varepsilon)f_1(X\beta^*)$, and thus using Markov's inequality gives us that $Y > a(1+\varepsilon)f_1(X\beta^*)$ holds with probability at most $\frac{1}{a}$. Hence it follows that $f_{1w}(SX\beta^*) \leq f_1(X\beta^*) + a(1+\varepsilon)f_1(X\beta^*)$ with failure probability at most $\frac{1}{a}$.
3) The proof is again similar to 2). The only difference is that we use Lemma 3.23 instead of Lemma 3.21. Hence the number of buckets at each level is bounded by $N = \max\{N_0', N_1' \cdot N_2'^{1+\varepsilon^{-1}}\}$. Thus $r = h_m N = O(\frac{d^2 h_m q_m^2 \mu^2 m_1^2}{\delta\varepsilon^7} + \frac{32d\mu\ln(n)^2}{\varepsilon^5} \cdot (\frac{64d\ln(n)^4}{\varepsilon^7\delta})^{1+\varepsilon^{-1}})$. □

Note that if μ is not too large, i.e. $\mu \in O((d\log^3(n))^c)$ then we can get a linear upper bound of the sketch size also in terms of μ. This has been worked out in (Munteanu et al., 2023).

3.6 Extension to linear ℓ_1-regression

For ℓ_1 regression, where the objective is $\|X\beta - y\|_1$, we have

Theorem 2. *Let $X \in \mathbb{R}^{n\times d}$ and let $Y \in \mathbb{R}^n$. Let $\varepsilon, \delta > 0$ and let $a > 1$. Then there is a distribution over sketching matrices $S \in \mathbb{R}^{r\times n}$ and a corresponding weight vector $w \in \mathbb{R}^r$, for which $X' = S[X, Y]$ can be computed in T time in a single pass over a turnstile data stream such that (X', w) is a weak weighted $(\mathbb{R}^d, \alpha, \varepsilon)$-sketch for ℓ_1-regression with failure probability at most P, where*

1. *$r = O(d^{1+c}\ln(n)^{3+5c})$ for any constant $1 \geq c > 0$, $T = O(d\ln(n)\mathrm{nnz}(X))$, and $\alpha = 1 + \frac{1}{c}$ and P are constant,*
2. *$r = O(\frac{d^3\ln(n)^5}{\delta\varepsilon^7}) + \frac{32d\ln(n)^3}{\varepsilon^5} \cdot (\frac{64d\ln(n)^5}{\varepsilon^6\delta})^{1+\varepsilon^{-1}}$, $T = O(\mathrm{nnz}(X))$, $\alpha = (1+a\varepsilon)$, and $P = \delta + \frac{1}{a}$.*

In the following we discuss the changes in the proofs:

The sketching algorithm is the same as before and also the analysis is very similar to the previous part. We start with a fixed point $z = (X, -y)\beta'$, where $\beta' = (\beta, 1) \in \mathbb{R}^d$ and analyze Sz. Again we assume that $\|z\|_1 = 1$. Instead of weight classes W_q^+ we use weight classes $W_q = \{i \in [n] \mid |z_i| \in (2^{-q-1}, 2^{-q}]\}$. Since we are only dealing with absolute values, which are symmetric, we no longer need to parameterize by μ. We can continue to use the same definitions for $q_h(1)$, $q_h(2)$ and $q_h(3)$ when setting μ in those bounds to be 1. We will only slightly change $q_h(3)$ since we will need another trick to prove the second outer bound $q_h'(4)$.

3.6.1 Dilation bounds for ℓ_1

When looking at logistic regression we had that for any any good $\beta \in \mathbb{R}$, i.e. β with $f_1(X\beta) \leq n\ln(2)$, assumption of Lemma 3.8 1) is fulfilled. For ℓ_1-regression this is no longer the case. Thus for approximating ℓ_1 we need a different approach for $q_h(4)$ when bounding the contribution of small entries at each level. The idea is, similar as in (Clarkson and Woodruff, 2015), to use a Ky-Fan norm argument to remove the smallest contributions from the ℓ_1-norm. At a fixed level h we put $\mathcal{B}_h$ to be the set of buckets at level h and $\mathcal{B}_h'$ to be the set of buckets with the $p2^{q_h'(3)} \leq \frac{\varepsilon N}{h_m}$ largest entries with respect to $|G(B)|$ where $q_h'(3) := \ln(\min\{\frac{\varepsilon N}{2h_m}, \frac{N\varepsilon^2}{4}\})$.

We further define

$$K(h) = \sum_{B \in \mathcal{B}'_h} |G(B)|.$$

Since $\bigcup_{q \in [q_h(2), q'_h(3)]} W_q^*$ contains at most $p2^{q'_h(3)}$ elements, we have that $K(h) \geq \|W_q^*\|_1$. We set $q'_h(4) = \ln(\frac{3Nh_m \ln Nh_m/\delta}{p\varepsilon})$. Set $Y_2 = Y_2(h) = \{i \in [n] \mid |z_i| \leq \gamma_2 := \frac{cp}{N \ln(Nh_m/\delta)}\}$ to be the set of small elements at level h.

Lemma 3.24. *With failure probability at most $\frac{\delta}{h_m N}$ it holds that for any bucket B at level h we have that*

$$\sum_{i \in B \cap Y_2} |z_i| \leq \max\left\{2 \cdot \frac{p \cdot \|Y_2\|_1}{N}, \frac{p}{N}\left(\|Y_2\|_1 + \frac{\varepsilon}{h_m}\right)\right\}$$

Proof. Fix a bucket B at level h. For $i \in Y_2$ let $X_i = z_i$ if $i \in B$ and $X_i = 0$ otherwise. Then we have $E := \mathbb{E}(\sum_{i \in Y_2} X_i) = \frac{p \cdot \|Y_2\|_1}{N}$. Further we have $\mathbb{E}(\sum_{i \in Y_2} X_i^2) = \sum_{i \in Y_2} \frac{p}{N} \cdot z_i^2 \leq \frac{\gamma_2 p}{N} \cdot \sum_{i \in Y_2} |z_i| = \gamma_2 E$. We set $\lambda = \max\{E, \frac{\varepsilon}{Nh_m}\}$ Then using Bernstein's inequality we get that

$$\begin{aligned}
P(\sum_{i \in Y_2} X_i \geq E + \lambda) &\leq \exp\left(\frac{-\lambda^2/2}{\gamma_2 E + \gamma_2 E/3}\right) \\
&\leq \exp\left(\frac{-\lambda^2/2}{\gamma_2 \lambda + \gamma_2 \lambda/3}\right) \\
&\leq \exp\left(\frac{-\lambda}{3\gamma_2}\right) \\
&\leq \exp\left(\frac{-p\varepsilon}{3Nh_m\gamma_2}\right) \\
&\leq \exp\left(-\ln(Nh_m/\delta)\right) \leq \frac{\delta}{h_m N}.
\end{aligned}$$

□

Lemma 3.25. *With failure probability at most δ it holds that*

$$\sum_{h \leq h_m} \sum_{i \in Y_2(h) \cap \bigcup_{B \in \mathcal{B}'_h} B} z_i \leq \varepsilon$$

Proof. Using the union bound over the event from Lemma 3.24 over all Nh_m buckets, using that $|\mathcal{B}'_h| \leq \varepsilon N/2h_m$ and $\max\{2 \cdot \|Y_2\|_1, \left(\|Y_2\|_1 + \frac{\varepsilon}{h_m}\right)\} \leq 2$ we get that

$$\sum_{i \in Y_2(h) \cap \bigcup_{B \in \mathcal{B}'_h} B} z_i \leq \frac{\varepsilon N}{2h_m} \cdot \frac{2p}{N} \leq \frac{\varepsilon}{h_m}.$$

holds for every level h with failure probability at most δ. Summing up over all levels we get that

$$\sum_{h\in h_m}\sum_{i\in Y_2(h)\cap\bigcup_{B\in\mathcal{B}'_h}B} z_i \leq \varepsilon.$$

□

We have the following lemmas using similar proofs as in the previous section:

Lemma 3.26. *If for some $k\in\mathbb{N}$ we choose $N_i = N \geq \frac{32m_1q_mh_m}{\varepsilon^5}\cdot\left(\frac{64m_1\ln(Nh_m/\delta)q_mh_m^2}{\varepsilon^6\delta}\right)^{1/(k-1)}$ for all $i\in[h_m]$ and M_i solving the equation $q_{(M_{i-1},N)}(3) = q_{(M_i,N)}(2)$, then the expected contribution of any weight class W_q is at most $(k+\varepsilon)\|W_q\|_1$.*

Here the additional ε comes from Lemma 3.24.

We set$N''_0 = N'_0$, $N''_1 = \frac{32m_1q_mh_m}{\varepsilon^5}$ and $N''_2 = \frac{64m_1\ln(Nh_m/\delta)q_mh_m^2}{\varepsilon^6\delta}$ and assume that $n\geq N^kh_m/\delta$.

Lemma 3.27. *Let $\gamma = \frac{1}{k} < 1$ for some $k\in\mathbb{N}$. Assume that N_0 is chosen uniformly at random from $N^{(1)},\dots N^{(1/\gamma)}$ where $N^{(i)} = N''_0\cdot N''^i_2$. Further let $N_i = N = N''_1\cdot N''^{k+1}_2$ for any $i>0$. Then the expected contribution of any weight class W_q is at most $(1+\gamma)\|W_q\|_1$.*

3.6.2 Net argument

For $\beta\in\mathbb{R}^{d+1}$ we set $g_1(\beta) = \|(X,-y)\beta)\|_1$ and $g_2(\beta) = \|(SX,-Sy)\beta)\|_1$

Lemma 3.28. *Assume that for $\beta\in\mathbb{R}^{d+1}$ it holds that $|g_1(\beta)-g_2(\beta)|\leq\varepsilon$. Then for any $\beta'\in\mathbb{R}^d$ with $\|X\beta - X\beta'\|_1\leq\varepsilon/(b^{h_m}h_m)$ it holds that $|g_1(\beta')-g_2(\beta')|\leq 3\varepsilon$.*

Proof. It holds that $\|X'(\beta-\beta')\|_1 = \|SX(\beta-\beta')\|_1 \leq b^{h_m}h_m\|X(\beta-\beta')\|_1\leq\varepsilon$ since for each $i\in[n]$ there are at most h_m columns j such that $S_{ij}\neq 0$ and each entry of S is bounded by b^{h_m}. Also note that $\|g_i(v)-g_i(v')\|_1\leq\|v-v'\|_1$ holds for any two vectors $v,v'\in\mathbb{R}^{d+1}$. Thus, using the triangle inequality yields

$$\begin{aligned}|g_1(\beta')-g_2(\beta')| &\leq |g_2(\beta')-g_2(\beta)| + |g_2(\beta)-g_1(\beta)| + |g_1(\beta)-g_1(\beta')|\\ &\leq \varepsilon+\varepsilon+\varepsilon\leq 3\varepsilon.\end{aligned}$$

□

Lemma 3.29. *There exists a net $\mathcal{N}\subset\mathbb{R}^d$ with $|\mathcal{N}| = \exp\left(\mathcal{O}(d\ln(n))\right)$ such that if $|g_1(\beta)-g_2(\beta)|\leq\varepsilon g_1(\beta)$ holds for any $\beta\in\mathcal{N}$ then for any $\beta'\in\mathbb{R}^{d+1}$ it holds that $|g_1(\beta')-g_2(\beta')|\leq 3\varepsilon g_1(\beta')$.*

Proof. We set

$$\mathcal{N} = \left\{\beta = v\cdot\frac{\varepsilon}{db^{h_m}h_m} \mid v\in\mathbb{Z}^d \text{ with } \|v\|_\infty\leq\frac{db^{h_m}h_m}{\varepsilon}\right\}. \tag{15}$$

Then it holds that for any $\beta \in \mathbb{R}^{d+1}$ with $g_1(\beta) = 1$ the point $(X, -y)\beta' = \lfloor \frac{db^{h_m} h_m}{\varepsilon} \cdot (X, -y)\beta) \rfloor \cdot \frac{\varepsilon}{db^{h_m} h_m}$ is in $\mathcal{N}$ and it holds that $\|(X, -y)\beta'\|_1 \leq d \cdot \frac{\varepsilon}{db^{h_m} h_m} = \frac{\varepsilon}{b^{h_m} h_m}$. Using Lemma 3.28 it holds that $|g_1(\beta') - g_2(\beta')| \leq 3\varepsilon \leq 3\varepsilon g_1(\beta)$. Further we have $|\mathcal{N}| \leq \left(\frac{db^{h_m} h_m}{\varepsilon}\right)^{2d} = \exp\left(\mathcal{O}(d \ln(n)\right)$. Now for any $r \in \mathbb{R}$ and $\beta \in \mathbb{R}^{d+1}$ with $g_1(\beta) = 1$ we have that $|g_1(r\beta) - g_2(r\beta)| = |rg_1(\beta) - rg_2(\beta)| = r|g_1(\beta) - g_2(\beta)| \leq 3\varepsilon r$. □

3.7 Extension to logistic regression with variance-based regularization

For the variance-based regularization, where we consider the full objective function $f(X\beta)$, we have

Theorem 3. *Let $X \in \mathbb{R}^{n \times d}$ be a μ-complex matrix for bounded $\mu < n$. Let $\varepsilon, \delta > 0$, let $a > 1$ and set $V = \{X\beta \mid f_1(X\beta) \leq \ln(2)(1 - \varepsilon)\}$. Then there is a distribution over sketching matrices $S \in \mathbb{R}^{r \times n}$ and a corresponding weight vector $w \in \mathbb{R}^r$, for which $X' = SX$ can be computed in T time in a single pass over a turnstile data stream such that (X', w) is a weak weighted (V, α, ε)-sketch for f with failure probability at most P, where*

- $r = O(\frac{n^{0.5+c}\mu d^2 \ln^3(n)}{\varepsilon^5} \cdot \max\{d, \ln(n), \varepsilon^{-1}, \delta^{-1}, \mu\} + \frac{d^3\mu^2 \ln(n)^3\sqrt{n}}{\delta\varepsilon^6})$, *for arbitrary constant* $1 \geq c > 0$, $T = O(\text{nnz}(X))$, $\alpha = 1 + \frac{1}{c}$, *and* $P = \delta + \frac{1}{a}$.

In this section we show that our algorithm also approximates the variance well under the assumption that roughly $f_1(X\beta) \leq \ln(2)$. We stress that this assumption does not rule out the existence of good approximations. Indeed, even the minimizer is contained, since we have that $\min_{\beta \in \mathbb{R}^d} f(X\beta) \leq f(0) = f_1(0) = \ln(2)$ and $f(X\beta) \geq f_1(X\beta)$ holds for any $\beta \in \mathbb{R}^d$. Again we focus on a single $z = X\beta$ first. What remains to show is that $\sum_{i:z_i>0} z_i^2$ is approximated well. We set $H(z) = \sum_{i=1}^n z_i^2$, $H^+(z) = \sum_{i:z_i>0} z_i^2$ and $h(y) = \frac{y^2}{H^+(z)}$. By μ-complexity we get that $H^+(z) \geq \frac{H(z)}{\mu}$. We define $W_q^2 = \{i \in [n] \mid h(z_i) \in (2^{-q-1}, 2^q]\}$ and $W_q^1 = \{i \in [n] \mid \frac{z_i}{\|z\|_1} \in (2^{-q-1}, 2^q]\}$. As the argument is almost the same as in the section before, we will only note the differences. We will also use the same definition of importance, i.e., a weight class W_q^2 is important if $H^+(W_q^2) \geq \frac{\varepsilon}{q_m \mu}$. Similar to the previous analysis we have that if W_q^2 is important then $|W_q^2| \geq \frac{\varepsilon 2^q}{q_m \mu}$. With those adapted definitions we proceed by adapting the main lemmas of Section 3.4.3 that finally yield Theorem 3.

Lemma 3.30. *For any $z_i \in W_q^2$ there exists $q' \leq (q-1)/2 + \ln(n)/2$ such that $z_i \in W_{q'}^1$.*

Proof. It is well known that $\|z\|_1 \leq \sqrt{n}\|z\|_2$. We conclude that

$$\frac{z_i}{\|z\|_1} \geq \frac{z_i}{\sqrt{n}\|z\|_2} = \frac{1}{\sqrt{n}} \frac{z_i^2}{\|z\|_2^2} \Big/ \sqrt{\frac{z_i^2}{\|z\|_2^2}} \geq \frac{1}{\sqrt{n}} \cdot \frac{2^{-q-1}}{2^{-(q-1)/2}}.$$

Now taking the logarithm proves the lemma. □

Contraction bounds Recall that:

$$\gamma_1 := \frac{p}{3m_1}$$
$$Y_1 := \{i \in [n] \mid |z_i| \geq \gamma_1\}$$

Here Y_1 is the set of 'large elements'. We redefine $\mu_z = \frac{\sum_{z_i>0} z_i^2}{\sum_{z_i<0} z_i^2}$

Lemma 3.31. *The following hold:*

1) $|Y_1 \cap U| \leq \varepsilon N/2$ *with failure probability at most* $\exp(-m_1)$*;*

2) *Let* $\mathcal{B} = \{B \in \mathcal{B}_h \mid \sum_{i \in B \setminus Y_1} |z_i| \leq \frac{4p}{\varepsilon N}\}$. *Then* $|\mathcal{B}| \geq (1-\varepsilon/2)N$ *with failure probability at most* $\exp(-m_1)$*;*

3) *Assume that* $q \geq \log_2(\frac{8q_m\mu_z m_1}{\varepsilon^3 p}))$ *and that* W_q^2 *is important or that* $|W_q| \geq 8m_1\varepsilon^{-2} \cdot p^{-1}$. *Then with failure probability at most* $\exp(-m_1)$ *there exists* $W_q^* \subset W_q^2 \cap \mathcal{B}$ *such that* $\|W_q^*\|_1 \geq (1-\varepsilon)^2\|W_q^+\|_1 \cdot p$ *and each element of* W_q^* *is in a bucket in* $\mathcal{B}$ *containing no other element of* Y_1*;*

4) *If* $q \leq \log_2(\frac{N\varepsilon^2}{\sqrt{n}4p})$ *and* W_q^* *as in 3) exists, then with failure probability at most* $\exp(-m_1)$ *it holds that* $\sum_{i \in W_q^*} G(B_i) \geq (1-\varepsilon)\|W_q^*\|_1$.

The proof is verbatim to the proof of Lemma 3.7. For the 4th part we use Lemma 3.30 to reduce the problem to the weight class W_q^1. This causes an additional term of $\frac{1}{\sqrt{n}}$ in the logarithm of $q_3(M,N)$.

We also have a change in $q_4(M,N)$. More precisely we need two additional factors of ε in γ_2:

$$\gamma_2 := \frac{M\varepsilon^4}{2Nn\ln(Nh_{\max}/\delta)}$$
$$Y_2 = \{i \in [n] \mid |z_i| \leq \gamma_2\}\text{: Set of small elements;}$$
$$Y_2^+ = \{i \in [n] \mid |z_i| \leq \gamma_2, z_i \leq 0\}\text{: Set of small negative elements;}$$
$$Y_2^- = \{i \in [n] \mid z_i \leq \gamma_2, z_i \geq 0\}\text{: Set of small positive elements;}$$

Further we set $A := \sum_{z_i \geq 0} z_i$, $A' = \sum_{z_i \in Y_2^-} |z_i|$, $A_1 = \sum_{z_i \in Y_2^+} |z_i|$ and $A_2 = A - A_1 \geq 0$.

Lemma 3.32. *If* $A' \geq A(1+\varepsilon)$ *then for any bucket* B *that contains only elements of* Y_2 *we have that* $G(B) = \sum_{i \in B} z_i \leq \frac{M}{Nn} \cdot (-A_2)$ *with failure probability at most* $\frac{\delta}{Nh_{\max}}$.

Proof. Let X_i be the random variable attaining value z_i if $i \in B$ and 0 otherwise, for $i \in [n]$. The expected value for $G(B) = \sum_{i \in [n]} X_i$ is $E' := \frac{M}{nN} \cdot (A' - A_1)$. Further we have that

$$\mathbb{E}(\sum_{i \in [n]} X_i^2) = \sum_{i \in Y_2} \frac{M}{nN} \cdot z_i^2 \leq \frac{M}{nN} \cdot \sum_{i \in Y_2} \gamma_2 z_i \leq \frac{\gamma_2 M}{nN}$$

since all X_i are bounded by γ_2 by assumption. Applying Bernstein's inequality thus yields

$$\begin{aligned}
P(G(B) > 0) \leq P\left(\sum_{i\in[n]} X_i - E' \geq \varepsilon|E'|\right) &\leq \exp\left(\frac{-\varepsilon^2|E'|^2/2}{\gamma_2 \cdot M/(nN) + \varepsilon\gamma_2|E'|/3}\right) \\
&\leq \exp\left(\frac{-\varepsilon^3 \cdot M/(nN)/2}{\gamma_2(M/(nNE') + \varepsilon/3)}\right) \\
&= \exp\left(\frac{-\varepsilon^3 \cdot M/(nN)/2}{\gamma_2\varepsilon^{-1}((A' - A_1) + 1/3)}\right) \\
&\leq \exp\left(\frac{-\varepsilon^4 \cdot M/(nN)}{2\gamma_2}\right) \\
&\leq \exp\left(-\ln\left(\frac{Nh_{\max}}{\delta}\right)\right) \\
&= \frac{\delta}{Nh_{\max}}.
\end{aligned}$$

Note that $\varepsilon E' \leq \varepsilon \cdot \frac{M}{nN} \cdot (A' - A_1) \leq \varepsilon \cdot \frac{M}{nN} \cdot A$ and thus $\mathbb{E}(\sum_{i\in[n]} X_i^2) + \varepsilon E' \leq \frac{M}{nN} \cdot (-A' + A_1 + \varepsilon A) \leq \frac{M}{nN} \cdot (-A_2)$. □

Our main lemma thus changes to:

Lemma 3.33. *With probability at least $1 - \frac{\delta}{h_m}$ the weight classes W_q^2 for $q \geq q_{(M,N)}(4) := \log_2(\gamma_2^{-1}) := \log_2(\frac{2Nn\ln(Nh_m/\delta)}{M\varepsilon^4})$ and $q \leq q_{(M,N)}(1) := \log_2(\frac{n\delta}{Mh_m})$ have zero contribution to $\sum_B G^+(B)$, i.e., for any bucket B we have $\sum_{z_i\in B\setminus I_r} z_i \leq 0$ where $I_r = \{i \in [n] \mid z_i \in W_q, q \in [q_{(M,N)}(1), q_{(M,N)}(4)]\}$. Further, with failure probability at most $\exp(-m_1)$, for each $\log_2(\frac{8q_m\mu m_1 n}{\varepsilon^3 M}))$ =: $q_{(M,N)}(2) \leq q \leq q_{(M,N)}(3) := \log_2(\frac{Nn\varepsilon^2}{4Mm_1\sqrt{n}})$ there exists W_q^* such that $\sum_{i\in W_q^*} G(B_i) \geq (1-\varepsilon)^2\|W_q^2\|_2 \cdot \frac{M}{n}$. Thus it holds that:*

$$\begin{aligned}
q_{(M,N)}(2) - q_{(M,N)}(1) &= \log_2\left(\frac{8q_m m_1 h_m}{\varepsilon^3\delta}\right) \\
q_{(M,N)}(3) - q_{(M,N)}(2) &= \log_2\left(\frac{N\varepsilon^5}{32m_1\mu q_m\sqrt{n}}\right) =: \log_2(b) \\
q_{(M,N)}(4) - q_{(M,N)}(3) &= \log_2\left(\frac{8\ln(Nh_m/\delta\sqrt{n})}{\varepsilon^6}\right).
\end{aligned}$$

If $N = M$ then we set $q_{(M,N)}(3) = q_{(M,N)}(4) = \infty$. If $M = n$ then we set set $q_{(M,N)}(1) = q_{(M,N)}(2) = 0$. We set $q_h(i) = q_{(M_h,N_h)}(i)$ for $i \in \{1,2,3,4\}$ and $Q_h = [q_h(2), q_h(3)]$ to be the well-approximated weight classes and $Q_h' = [q_h(1), q_h(4)]$ to be the relevant weight classes at level h. Note that $q_{(M,N)}(1)$ and $q_{(M,N)}(2)$ stay the same as before.

Heavy hitters The important changes to note here are that we need to replace Lemma 3.10 with an appropriate lemma for the ℓ_2-leverage scores and there is an additional factor of $\frac{1}{\sqrt{n}}$.

Lemma 3.34. *Let $Y_3 = \{i \mid u_i^{(2)} \geq \gamma_3\}$ where $\gamma_3 = \frac{\varepsilon^3}{8q_m\mu m_1}$. Further, for $j \in Y_3$ let $\mathcal{C}_j = \{B \mid \sum_{i\in B\setminus\{j\}} u_i \geq \varepsilon\gamma_3/\sqrt{n}\}$. Let $Y_3' = \{j \in Y_3 \mid j \in \mathcal{B}_0\setminus\mathcal{C}_j\}$ If $N_0 \geq \max\{\frac{2d^2\mu}{\gamma_3\varepsilon^2}, \frac{2\ln(\kappa^{-1})d^2}{\gamma_3\varepsilon\kappa}\}$ for $\kappa \in (0, 1/2)$, then with probability*

$1-2\kappa$, *it holds that* $\sum_{j\in Y_3\setminus Y_3'} u_j^{(2)} \leq 2\varepsilon\mu^{-1}$.

We denote by $\mathcal{E}_2$ the event that the event described in Lemma 3.34 holds. By Lemma 3.34, $\mathcal{E}_2$ holds with probability at least $1-\delta$ for an appropriate $N = N_0^{(2)}\mathcal{O}(\frac{d^2 q_m \mu^2 m_1\sqrt{n}}{\delta\varepsilon^6})$. For any entry $z_p \in H$ we have $z_p \geq \gamma_3$ and thus by Lemma 3.9, we have $p \in Y_3$ and for any entry $p \notin Y_4$ we have $z_p < \gamma_3 \cdot \gamma_4$.

Lemma 3.35. *Assume $\mathcal{E}_2$ holds. Then for any $z_i \in H$ we have $G(B_i) \geq (1-\varepsilon)z_i$. Further for it holds that $\sum_{j\in Y_3\setminus Y_3'} z_j^2 \leq 2\varepsilon\mu^{-1}\|z\|_2^2$.*

The proofs of Lemma 3.34 and Lemma 3.35 are similar to the proofs of Lemma 3.10 and Lemma 3.11.

Contraction bounds for a single point

Lemma 3.36. *Assume that $\mathcal{E}_2$ holds. Denote by z_i' the i-th row of $SX\beta$ for $i \in n'$. Then with failure probability at most $(2h_m + 2q_m)e^{-m_1}$ it holds that*

$$\sum_{i\in n', z_i'\geq 0} w_i z_i' \geq (1-12\varepsilon)G^+(X\beta).$$

Here we loose additional factors of ε for the following reason: assume that for some $z_i > 0$ we have $\|B_i\|_1 \geq (1-6\varepsilon)z_i$ then it holds that $\|B_i\|_1^2 \geq (1-12\varepsilon)z_i^2$.

Dilation bounds Here we have to cope with the additional factor of $\sqrt{n}$. Recall that if we choose M_i solving the equation $q_{(M_{i-1},N)}(3) = q_{(M_i,N)}(2)$ then it holds that

$$q_{(i+2)}(1) - q_i(4) = q_{i+1}(3) - q_{i+1}(2) - (q_{i+2}(2) - q_{i+2}(1)) - (q_i(4) - q_i(3)).$$

We now have

$$q_{i+1}(3) - q_{i+1}(2) = \log_2\left(\frac{N\varepsilon^5}{32\sqrt{n}m_1\mu q_m}\right)$$

and

$$(q_{i+2}(2) - q_{i+2}(1)) + (q_i(4) - q_i(3)) = \log_2\left(\frac{64m_1\ln(Nh_m/\delta)q_m h_m}{\sqrt{n}\varepsilon^9\delta}\right).$$

Further there is a change in Lemma 3.20 as we have to deal with possible overhead coming from the square function. We set $R = \{i | 2^{-q_h(1)} > z_i > 2^{-q_h(4)}\}$ to be the set of relevant (positive) elements and $W_R = \{z_i \mid i \in R\}$.

Lemma 3.37. *If $\sum_{i=1}^n z_i \leq 0$ and for all $i \leq h_m - 1$ it holds that $q_{(M_iN_i)}(4) < q_{(M_{i+k}N_{i+k})}(1)$ and $N_0 \geq N_0'$, then the expected contribution of any weight class Y_1' is at most $k \cdot \|W_R\|_2^2$.*

Proof. Fix a level h and a bucket B at level h. Recall that $\sum_{i\in R} z_i \leq A_2 = A - A_1 = \sum_{i,z_i\geq 2^{-q_h(4)}} z_i$. Note that by Lemma 3.32 we have that $\sum_{i\in Y_1'\cap B} \leq \frac{p_h(-A_2)}{N}$. Let Z_i be the random variable where $Z_i = z_i$ if $i \in R$ is assigned to B and 0 otherwise. Then the expected value of $Z = \max\{0, \sum_{i=1}^n Z_i\}$ is $\frac{p_h}{N} \cdot A_2$. Thus it holds that

$$\begin{aligned}\mathbb{E}(\max\{G(B),0\}^2) \leq \mathbb{E}\left(Z - \frac{p_h(-A_2)}{N}^2\right) &\leq \mathbb{E}\left((Z-\mathbb{E}(Z))^2\right)\\ &= \mathrm{Var}(Z) \leq \|W_R\|_2^2.\end{aligned}$$

□

Lemma 3.21 and Lemma 3.22 can be adapted as follows:

Lemma 3.38. *If we choose $N_i = N := \max\{N_0^{(2)}, \frac{\sqrt{32q_m\mu}m_1 n^{0.75}}{\varepsilon^{2.5}}\}$ for all $i \in [h_m]$ and M_i solving the equation $q_{(M_{i-1},N)}(3) = q_{(M_i,N)}(2)$ then the expected contribution of any weight class W_q is at most $2\|W_q\|_1$.*

Proof. The proof uses a different idea as before: since N is large enough, we only need 2 levels. More precisely we want to achieve $M_2 = N$. By our choice of M_2 this means

$$\log_2\left(\frac{8q_m\mu m_1 n}{\varepsilon^3 N}\right) = q_{(N,N)}(2) = q_{(n,N)}(3) = \log_2\left(\frac{Nn\varepsilon^2}{4n\sqrt{n}}\right)$$

or equivalently

$$N = \sqrt{\frac{32q_m\mu m_1 n^{1.5}}{\varepsilon^5}} = \frac{\sqrt{32q_m\mu}m_1 n^{0.75}}{\varepsilon^{2.5}}.$$

□

Lemma 3.39. *If for some $k \in \mathbb{N}$ we choose $N_i = N \geq \frac{32m_1\mu q_m}{\varepsilon^5} \cdot \left(\frac{64m_1^2\ln(n)q_m h_m\sqrt{n}}{\varepsilon^9\delta}\right)^{1/(k-1)}$ for all $i \in [h_m]$ and M_i solving the equation $q_{(M_{i-1},N)}(3) = q_{(M_i,N)}(2)$ then the expected contribution of any weight class W_q is at most $k\|W_q\|_1$.*

The proof is the same as for Lemma 3.22.

Net argument

Lemma 3.40. *For any $v, v' \in \mathbb{R}^n$ with $\|v - v'\|_1 \leq \varepsilon$ it holds that $|f_2(v) - f_2(v')| \leq (f_1(v) + \varepsilon)\varepsilon$.*

Proof. We have that $(\ell^2)'(v) = \frac{e^v}{e^v+1} \cdot \ell(v) \leq \ell(v)$. Further since $\ell'(v) \leq 1$ we have that for any $\nu \in [0,1]$ it holds that $|\ell(v + \nu(v'-v)) - \ell(v)| \leq (\ell(v)+\varepsilon)\varepsilon$. Thus we get that

$$|f_2(v) - f_2(v')| \leq \frac{1}{n}\cdot\sum_{i=1}^n |\ell(v_i)^2 - \ell(v_i')^2| \leq \frac{1}{n}\cdot\sum_{i=1}^n (\ell(v)+\varepsilon)\varepsilon = (f_1(v)+\varepsilon)\varepsilon$$

which proves the lemma. □

Lemma 3.41. *Assume that for $\beta \in \mathbb{R}^d$ it holds that $|f_2(X'\beta) - f_2(X\beta)| \leq \varepsilon$. Then for any $\beta' \in \mathbb{R}^d$ with $\|X\beta - X\beta'\|_1 \leq \varepsilon/(b^{h_m} h_m)$ it holds that $|f_2(X\beta') - f_2(X'\beta')| \leq \varepsilon + 2(f_1(X\beta') + \varepsilon)\varepsilon$.*

Proof. It holds that $\|X'(\beta - \beta')\|_1 = \|SX(\beta - \beta')\|_1 \leq b^{h_m} h_m \|X(\beta - \beta')\|_1 \leq \varepsilon$ since for each $i \in [n]$ there are at most h_m columns j such that $S_{ij} \neq 0$ and each entry of S is bounded by b^{h_m}. Thus, by the triangle inequality and applying Lemma 3.40 yields

$$\begin{aligned}|f_2(X\beta') - f_2(X'\beta')| &\leq |f_2(X'\beta') - f_2(X'\beta)| + |f_2(X'\beta) - f_2(X\beta)| + |f_2(X\beta) - f_2(X\beta')| \\ &\leq (f_1(X\beta') + \varepsilon)\varepsilon b^{-h_m} + \varepsilon + (f_1(X\beta') + \varepsilon)\varepsilon \leq \varepsilon + 2(f_1(X\beta') + \varepsilon)\varepsilon.\end{aligned}$$

□

Combining Lemma 3.17 and Lemma 3.41 we get:

Lemma 3.42. *There exists a net $\mathcal{N} \subset \mathbb{R}^d$ with $|\mathcal{N}| = \exp(\mathcal{O}(d \ln(n)))$ such that if $|f_1(X'\beta) - f_1(X\beta)| \leq \varepsilon$ holds for any $\beta \in \mathcal{N}$ then for any $\beta' \in \mathbb{R}^d$ with $\|X\beta'\|_1 \leq n\mu$ it holds that $|f_2(X'\beta') - f_2(X\beta')| \leq \varepsilon(f_2(X\beta') + f_1(X\beta'))$.*

Proof of Theorem 3 The proof of Theorem 3 works as the proof of Theorem 1, replacing the old lemmas with the new ones.

3.8 Lower bound

We note that the increased sketching dimension in terms of $\sqrt{n}$ comes from the inter norm inequality $\|x\|_1 \leq \sqrt{n}\|x\|_2$. Lemma 3.43 shows that there is no way to get around a factor of $\sqrt{n}$ using the `CountMin`-sketch. The proof gives an example where $\sqrt{n}$ is attained even for obtaining a superconstant (in μ) approximation. It does not rule out the existence of some other method that allows a lower sketching dimension. For example `Count`-sketch is known to work for ℓ_1 and ℓ_2 norms simultaneously within polylogarithmic size (Clarkson and Woodruff, 2015). But we stress that the standard sketches from the literature do not work for asymmetric functions since they confuse the signs of contributions leading to unbounded errors for our objective function or even for plain logistic regression, see (Munteanu et al., 2021).

The following lemma shows that there is no way to get around a factor of $\sqrt{n}$ using the `CountMin`-sketch. It constructs an input where $\sqrt{n}$ is attained even for obtaining a superconstant (in μ) approximation.

Lemma 3.43. *There exists a μ-complex data example X where our sketch with $o(\sqrt{n})$ rows fails to approximate f. Specifically, if $\lambda = 1$ it holds for the optimizer $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} f(SX\beta)$ that $f(X\tilde{\beta}) = \omega(\ln(\mu)^2) \cdot \min_{\beta \in \mathbb{R}^d} f(X\beta)$.*

Proof. Fix $\mu > 10$ and consider the following data

$$x_0 = (\sqrt{n})$$
$$x_i = (-1) \text{ for } i \in \left[1, n - \frac{n}{\mu}\right]$$
$$x_i = (1) \text{ for } i > n - \frac{n}{\mu}$$

As the example is 1-dimensional we only need to check the ratio for $\beta = 1$ and $\beta = -1$ in order to compute μ as multiplying with a scalar does dot not change the ratio between the sum of all positive points and the sum of all negative points. Also note that the ratio is inverted for $\beta = -1$ thus if the ratio is positive for $\beta = 1$ we do not need to check it for $\beta = -1$. Note that for $\beta = 1$ and $z = X\beta$ it holds that $\sum_{z_i>0} z_i = \sqrt{n} + \frac{n}{\mu}$ and $\sum_{z_i<0} |z_i| = n(1 - \frac{1}{\mu}) \geq \sqrt{n} + \frac{n}{\mu}$ if n is sufficiently large. We thus have

$$\mu_1(X) = \frac{n\left(1 - \frac{1}{\mu}\right)}{\sqrt{n} + \frac{n}{\mu}} \leq \frac{n}{\frac{n}{\mu}} = \mu$$

Further we have that $\sum_{z_i>0} z_i^2 = n + \frac{n}{\mu} \leq 2n$ and $\sum_{z_i<0} |z_i| = n(1 - \frac{1}{\mu}) \approx n$. Consequently we get that

$$\mu_2(X) = \frac{n + \frac{n}{\mu}}{n\left(1 - \frac{1}{\mu}\right)} \leq 2 < \mu$$

Since $d = 1$ this proves that our our example is 2μ-complex. Note that the following four facts hold for any level h:

- If for some c we have that $p_h \leq 1/b$ then with probability $1/b$ row x_0 is not sampled at level h. In particular this implies that x_0 is only present at level 0 with high probability, i.e. probability at least $\sum_{h=1}^{h_m} p_h \leq \frac{2}{b}$;
- If x_0 is in a bucket with $3\sqrt{n} \geq 2\sqrt{n}/(1 - \frac{2}{\mu})$ elements then with high probability $G(B_0) \leq 0$;
- If $\frac{N_h}{n} \ll p_h \ll 1$ then with high probability $G(B) < 0$ for any bucket at level h since the $\frac{\mu-1}{\mu} \cdot n \gg \frac{n}{\mu}$ negative elements cancel all positive rows;
- If $h = h_m$ then roughly $\frac{\mu-1}{\mu} \cdot N_u$ are -1 and $\frac{N_u}{\mu}$ are 1.

All of these follow from the Chernoff bounds using Lemma 2.14. Thus if $N_0 \ll \sqrt{n}/3$ then $X' = SX$ mimics the instance $X \setminus \{x_0\}$, i.e. the instance X with point x_0 removed, as x_0 is only appearing at level 0 where it is canceled by the other points. More precisely X' consists of roughly $n' - \frac{n'}{\mu}$ copies of the point -1 and $\frac{n'}{\mu}$ copies of the point 1. After multiplying with the weights we are back to roughly $n - \frac{n}{\mu}$ times the point -1 and $\frac{n}{\mu}$ times the point 1. To keep the presentation simple we only consider the instance $X' = -(X \setminus \{x_0\})$. The proof works the same for other sketched instances that we obtain using the above facts. Consider the

function

$$nf_1(X'r) = (n - \frac{n}{\mu}) \cdot \ell(-r) + \frac{n}{\mu} \cdot \ell(r) = n \cdot \ell(-r) + \frac{nr}{\mu}.$$

Thus, we have $f_1(X'r) = \ell(-r) + \frac{r}{\mu}$. Using that $\ell(r) < r + 1$ for all $r > 0$ we get

$$\begin{aligned} nf(X'r) &= nf_1(X'r) + \sum_{i=1}^{n}(x_i - f_1(X'r))^2 \\ &\leq n \cdot \ell(-r) + \frac{nr}{\mu} + \frac{n(r+1)^2}{\mu} + (n - \frac{n}{\mu}) \cdot \ell(-2r) \\ &\leq 2n \cdot \ell(-r) + \frac{nr}{\mu} + \frac{n(r+1)^2}{\mu}. \end{aligned}$$

Using that $\ell(-r) \leq e^{-r}$ it holds that

$$f(X'r) \leq 2e^{-r} + \frac{r}{\mu} + \frac{(r+1)^2}{\mu}.$$

Taking the derivative we get

$$f'(X'r) \leq -2e^{-r} + \frac{1}{\mu} + \frac{2(r+1)}{\mu}.$$

which is 0 if and only if $r = -\ln(\frac{2r+3}{\mu}) + \ln(2) = \Omega(\ln(\mu))$. This implies that for $\tilde{r} = \operatorname{argmin}_{r \in \mathbb{R}} f(Xr)$ we have that $\tilde{r} = \Omega(\ln(\mu))$. Now consider our original loss function $f(X\tilde{r})$. Here we have that

$$\begin{aligned} nf(Xr) = nf_1(Xr) + \sum_{i=1}^{n}(x_i - f_1(Xr))^2 &\geq n \cdot \ell(-r) + \frac{nr}{\mu} + (\sqrt{n} \cdot r)^2/2 \\ &\geq n \cdot r^2/2. \end{aligned}$$

In particular we have that $f(X\tilde{r}) = \Omega(\ln(\mu)^2)$. However for r^* minimizing $f(Xr)$ we have that $nf(Xr) \leq f(0) = \ln(2) = O(1)$. □

4 ℓ_p-leverage score sampling for p-probit regression

In this section we present and analyze a sampling algorithm to construct an $(1+\varepsilon)$-coreset for p-probit regression.

4.1 Setting and notations

Recall that we are given a data matrix $X \in \mathbb{R}^{n \times d}$ with rows $x_i \in \mathbb{R}^d$ for $i \in [n]$. Labels will again be omitted as pointed out in Subsection 2.6. Our goal is to find a weighted ε-coreset (X', w) (see Definition 2.15) for the following target function:

$$f_w(X\beta) = \sum_{i=1}^{n} -\ln(\Phi_p(-x_i\beta)) \cdot w_i.$$

where

$$\Phi_p(x) = \frac{p^{1-1/p}}{2\Gamma(1/p)} \int_{-\infty}^{x} \exp(-|t|^p/p)\, dt, x \in \mathbb{R}, p > 0.$$

is the p-generalized normal distribution. We omit the subscript whenever the weights are uniform, i.e., $w_i = 1$ for all $i \in [n]$. Moreover, to simplify notations we define the individual loss function

$$g(r) = -\ln(\Phi_p(-r)). \tag{16}$$

4.2 The algorithm

4.2.1 High level description

Before getting into the details we outline the Algorithm:

1. We make a first pass to sketch the data for the purpose of estimating their individual importance.
2. We make another pass to subsample the data proportional to their importance to obtain a coreset.
3. We solve the reduced problem on the coreset using a standard algorithm for convex optimization.

The first two steps are covered by Algorithm 2. This approach implements the sensitivity sampling framework (see Section 2.5). Recall that the importance measure that it builds upon is called sensitivity, which measures the worst case contribution of each input point to the objective function. For efficiency reasons we first compute a sketch of the data in one pass. In the second pass, the sketch is used to approximate the ℓ_p leverage scores, which upper bound the sensitivities of the input points. Hereby, we pass them one-by-one to a reservoir sampler to obtain the coreset. Finally, we can solve the original problem approximately using gradient descent or other standard methods for convex optimization (see Bubeck, 2015) on the resulting coreset.

For a more detailed description see the following pseudo code or the proof of Theorem 4.

4.2.2 Pseudo code

Algorithm 2 Coreset algorithm for p-generalized probit regression.

Input: data $X \in \mathbb{R}^{n \times d}$, number of rows k.;
Output: coreset $C = (X', w) \in \mathbb{R}^{k \times d}$ with k rows.;
Initialize sketch $X'' = \mathbf{0} \in \mathbb{R}^{n' \times d}$, (where $n' = O(d^2)$ for $p \leq 2$ or $n' = O(n^{1-\frac{2}{p}} \log n \cdot \text{poly}(d))$ for $p > 2$);
for $i = 1 \dots n$ **do**
    Draw a random number $B_i \in [n']$; ▷ hash to bucket B_i
    Draw a random number $\sigma_i \in \{-1, 1\}$; ▷ random sign
    if $p \neq 2$ **then**
        Draw a random number $\lambda_i \sim \exp(1)$; ▷ ℓ_p embedding
        $\sigma_i = \sigma_i / \lambda_i^{1/p}$.
    $X''_{B_i} = X''_{B_i} + \sigma_i \cdot x_i$. ▷ sketch
Compute the QR-decomposition of $X'' = QR$.; ▷ well-conditioned basis
Initialize coreset $X' = \mathbf{0} \in \mathbb{R}^{k \times d}$ ▷ coreset points
Initialize weights $w = 0 \in \mathbb{R}^k$; ▷ coreset weights
Initialize k independent weighted reservoir samplers S_j, sampling row X'_j, for each $j \in [k]$;
Initialize $G = I \in \mathbb{R}^{d \times d}$; ▷ Identity matrix
if $p = 2$ and $\ln n < d$ **then**
    Draw $G \in \mathbb{R}^{d \times \ln n}$ with $G_{ij} \sim N(0, \frac{1}{\ln n})$; ▷ JL-embedding
for $i = 1 \dots n$ **do**
    Compute $q_i = \|x_i(R^{-1}G)\|_p^p$; ▷ ℓ_p-leverage score approximation
    for $j = 1 \dots k$ **do**
        Feed $s_i = q_i + 1/n$ to S_j; ▷ unnormalized sampling probabilities
        if S_j samples x_i **then**
            $w_j = 1/(k \cdot s_i)$; ▷ unnormalized weights
            $X'_j = x_i$; ▷ save row identity in the coreset
$w = w \cdot \sum_{i=1}^n s_i$; ▷ normalize weights
return $C = (X', w)$;

4.3 Analysis

In the subsection we will analyze the target function and the algorithm.

4.3.1 Outline of the analysis

Our analysis is structured as follows:

We first analyze our loss function.

1) We then look at the tail behavior of the negative logarithm of p-generalized normal distribution and prove that it behaves non-asymptotically for all r roughly like r^p on the positive part of the reals and like $\exp(-|r|^p)$ on the negative part.

2) We then prove important properties of the individual loss function g itself that we need to bound sensitivities and VC-dimension

3) Next we look at the range space of the function space $\mathcal{F} = \{g_{x_i,w} \mid w \in \mathbb{R}_{\geq 0}, i \in [n]\}$ where $g_{x,w}(\beta) = wg(x\beta)$. We show that, when limiting the number of the weights, we can bound both the sensitivities and the VC-dimension and thus are able to use the sensitivity framework described in Section 2.5.

4) Then we show how to approximate the ℓ_p-leverage scores which are used as bounds for the sensitivities.

5) Last we combine everything to prove our main result.

4.3.2 Tails of the p-generalized normal distribution

For the normal distribution $\Phi_2(r) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{r} \exp(\frac{-|t|^2}{2})$ Gordon (1941) proved that for any $r \geq 0$ it holds that

$$\frac{r}{r^2+1} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-r^2}{2}\right) \leq \Phi_2(-r) \leq \frac{1}{r} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp\left(\frac{-r^2}{2}\right)$$

or equivalently

$$\left(1 - \frac{1}{r^2+1}\right) \cdot \frac{1}{r} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp(\frac{-r^2}{2}) \leq \Phi_2(-r) \leq \frac{1}{r} \cdot \frac{1}{\sqrt{2\pi}} \cdot \exp(\frac{-r^2}{2}).$$

In this section we generalize the analysis to the p-generalized normal distribution. We will present two proofs to show a similar result for general p. More precisely we show that

$$\frac{|\Phi_p(-r) - r^{-(p-1)} \exp\left(\frac{-r^p}{p}\right)|}{\exp\left(\frac{-r^p}{p}\right)} \in O(r^{-p})$$

for $r \geq 0$. The first proof gives a slightly weaker result but helps us understanding the integral appearing in in the generalized normal distribution. The second proof is less intuitive but gives a tighter result.

In this subsection we always assume that $r \geq 0$. For our first approach we consider the functions $f_1(r) = r^{p-1} \int_r^\infty \exp(-|t|^p/p)\, dt$ and the function $h_1(r) = \frac{\exp(-|r|^p/p) - f_1(r)}{\exp(-|r|^p/p)}$. Note that we have that

$$f_1(r) = (1 - h_1(r)) \exp(-|r|^p/p)$$

and that

$$\begin{aligned} h_1(r) &= \exp(|r|^p/p)(\exp(-|r|^p/p) - f_1(r)) \\ &= \exp(|r|^p/p)\left(\int_r^\infty t^{p-1}\exp(-t^p/p)\,dt - r^{p-1}\int_r^\infty \exp(-|t|^p/p)\,dt\right) \\ &= \int_r^\infty (t^{p-1} - r^{p-1})\exp(-(t^p - r^p)/p)\,dt \\ &= \int_0^\infty ((r+t)^{p-1} - r^{p-1})\exp(-((r+t)^p - r^p)/p)\,dt. \end{aligned}$$

Lemma 4.1. *The following claims hold:*

1 For $p = 1$ we have that $h_1(r) = 0$ for all r;

2 For $p = 2$ it holds that h_1 is a monotonically decreasing function;

3 It holds that $h_1(r) \leq (p-1)r^{-p}$.

To prove this Lemma we need the following Lemma:

Lemma 4.2. *For any $a, b \in \mathbb{R}_{\geq 0}$ and $p \geq 1$ it holds that $(a+b)^p \geq a^p + pa^{p-1}b$.*

Proof. First assume that $a \geq b$. Using Newton's generalized binomial theorem we get that

$$(a+b)^p = a^p + pa^{p-1}b + \frac{p(p-1)}{2!}a^{p-2}b^2 + \cdots.$$

If p is an integer all terms are positive and the Lemma follows. If p is not an integer then at least the first three terms are positive. More precisely the first $\lceil p \rceil$ terms are positive. Note that for $k \geq \lfloor p \rfloor$ the sequence $a_k = \frac{|(p)_k|}{k!} \cdot a^{p-k}b^k$, where $(p)_k = p(p-1)\cdots(p-k+1)$, is monotonically decreasing as $a_{k+1} = \frac{|p-k|}{k+1} \cdot \frac{b}{a} \cdot a_k$ and both $\frac{|p-k|}{k+1}$ and $\frac{b}{a}$ are less than 1 for $k \geq \lfloor p \rfloor$. Thus the sum $\sum_{k=\lfloor p \rfloor} \frac{(p)_k}{k!} \cdot a^{p-k}b^k$ is greater or equal to zero as the first term is positive, the sign of the terms is alternating and the absolute value of the terms is decreasing. We conclude that $(a+b)^p \geq a^p + pa^{p-1}b$.

Now assume that $b \geq a$. Then by the same argumentation as above we have that $(a+b)^p \geq b^p + pb^{p-1}a$. If $p \geq 2$ then we have that $b^{p-1}a \geq a^{p-1}b$ and thus $b^p + pb^{p-1}a \geq a^p + pa^{p-1}b$. To show the inequality for $p \in (1,2)$ note that $b^p + pb^{p-1}a \geq a^p + pa^{p-1}b$ is equivalent to

$$b^p - a^p \geq p(a^{p-1}b - b^{p-1}a).$$

The last inequality follows by

$$b^p - a^p = \int_a^b pt^{p-1}\,dt \geq (b-a)pa^{p-1} = p(ba^{p-1} - a^p) \geq p(ba^{p-1} - b^{p-1}a).$$

□

Proof. 1) Note that for $p = 1$ and $r \geq 0$ we have that $f_1(r) = \int_r^\infty \exp(-|t|)\,dt = \exp(-r)$ which implies that $h_1(r) = 0$ for any $r \geq 0$.

2) For $p = 2$ and $r \geq 0$ we have that

$$h_1(r) = \frac{\exp(-r^2/2) - f_1(2)}{\exp(-r^2/2)} = \int_0^\infty ((r+t) - r)\exp(-((r+t)^2 - r^2)/2)\,dt = \int_0^\infty t\exp(-(t^2/2 + rt)\,dt$$

which is monotonically decreasing in r.

3)For any $p \geq 1$ we have that $(t+r)^p - r^p \geq r^p + ptr^{p-1} - r^p = ptr^{p-1}$ by Lemma 4.2. Further recall that for any function $f : \mathbb{R} \to \mathbb{R}$ the derivative $\frac{d}{dr}\int_r^\infty f(t)\,dt$ equals $\lim_{t\to\infty} f(t) - f(r)$ if the integral is finite. We conclude that

$$\begin{aligned}
h_1(r) &= \int_0^\infty ((r+t)^{p-1} - r^{p-1})\exp(-((r+t)^p - r^p)/p)\,dt \\
&= -\frac{d}{dr}\int_0^\infty \exp(-((r+t)^p - r^p)/p)\,dt \\
&\geq -\frac{d}{dr}\int_0^\infty \exp(-tr^{p-1})\,dt \\
&= -\frac{d}{dr}\left(-\frac{1}{r^{p-1}}\cdot\exp(-tr^{p-1})\right)\Big|_0^\infty \\
&= -\frac{d}{dr}\frac{1}{r^{p-1}} = (p-1)r^{-p}.
\end{aligned}$$

□

Now using that $f_1(r) = (1 - h_1(r))\exp(-|r|^p/p)$ we get:

Corollary 4.3. *It holds that* $\left(1 - \frac{p-1}{r^p}\right)\cdot\frac{p^{1-1/p}}{2\Gamma(1/p)}\exp(-|r|^p/p) \leq \Phi_p(-r) \leq \frac{1}{r}\cdot\frac{p^{1-1/p}}{2\Gamma(1/p)}\exp(-|r|^p/p)$.

For our second approach we consider $h(r) := \exp(|r|^p/p)\int_r^\infty \exp(-|t|^p/p)\,dt$. This approach is similar to the approach of Gordon (1941) for the case $p = 2$, i.e., the standard normal distribution.

Lemma 4.4. *The following holds for any $r > 0$:*

$$h'(r) = r^{p-1}h(r) - 1; \tag{17}$$

$$h''(r) = (p-1)r^{p-2}h(r) + r^{p-1}h'(r); \tag{18}$$

$$h''(r) = \frac{r^p + p - 1}{r}h'(r) + \frac{p-1}{r}; \tag{19}$$

$$h'''(r) = \left(1 + \frac{p}{r^p + p - 1} + \frac{p-2}{r^p}\right)r^{p-1}h''(r) - \frac{(p-1)pr^{p-2}}{r^p + p - 1}; \tag{20}$$

$$h(r) > 0; \tag{21}$$

$$h'(r) < 0; \tag{22}$$

$$h(r) < \frac{1}{r^{p-1}} \tag{23}$$

Further if $r \geq 1$ or if $p \geq 2$ and $r > 0$ then it holds that

$$h''(r) \geq 0; \tag{24}$$

$$h(r) \geq \frac{r}{r^p + p - 1}. \tag{25}$$

Proof. Equations (17) and (18) can be derived by a direct calculation of the derivatives. Note that (17) is equivalent to

$$h(r) = \frac{h'(r) + 1}{r^{p-1}} \tag{26}$$

Equation (19) follows by substitution of (26) in (18). Equation (18) is equivalent to

$$h(r) = \frac{h''(r)}{(p-1)r^{p-2}} - \frac{r}{p-1}h'(r) \tag{27}$$

To get (20) we first note that by (17) and then (27) it holds

$$\begin{aligned}
\frac{p-1}{r^2}h'(r) &= \frac{p-1}{r^2}(r^{p-1}h(r) - 1) \\
&= \frac{p-1}{r^2}r^{p-1}h(r) - \frac{p-1}{r^2} \\
&= \frac{h''(r)}{r} - r^{p-2}h'(r) - \frac{p-1}{r^2}.
\end{aligned} \tag{28}$$

Further note that (19) is equivalent to

$$h'(r) = \frac{rh''(r)}{r^p + p - 1} - \frac{p-1}{r^p + p - 1} \tag{29}$$

Taking the derivative of (19) and using the equations (28) and (29) we get

$$\begin{aligned}
h'''(r) &= (p-1)r^{p-2}h'(r) - \frac{p-1}{r^2}h'(r) + r^{p-1}h''(r) \\
&\quad + \frac{p-1}{r}h''(r) - \frac{p-1}{r^2} \\
&\overset{(28)}{=} pr^{p-2}h'(r) + r^{p-1}h''(r) + \frac{p-2}{r}h''(r) \\
&\overset{(29)}{=} pr^{p-2} \cdot \left(\frac{r}{r^p + p - 1}h''(r) - \frac{p-1}{r^p + p - 1} \right) \\
&\quad + r^{p-1}h''(r) + \frac{p-2}{r}h''(r) \\
&= \left(1 + \frac{p}{r^p + p - 1} + \frac{p-2}{r^p} \right) r^{p-1}h''(r) \\
&\quad - \frac{(p-1)pr^{p-2}}{r^p + p - 1}.
\end{aligned}$$

Equation (21) follows since all terms appearing in $h(r)$ are positive.

For (22) we note that

$$\begin{aligned} r^{p-1}h(r) &= \exp(r^p/p)\int_r^\infty r^{p-1}\exp(-|t|^p/p)\,dt \\ &< \exp(r^p/p)\int_r^\infty \frac{p}{p}t^{p-1}\exp(-|t|^p/p)\,dt \\ &= \exp(r^p/p)\cdot\exp(-r^p/p) = 1 \end{aligned} \tag{30}$$

and thus (23) follows from dividing by r^{p-1} and (22) also follows from (30) using Equation (17).

Next we prove (24): For $r \geq 1$ it holds that $\left(1 + \frac{p}{r^p+p-1} + \frac{p-2}{r^p}\right) r^{p-1} > 0$. Now let $r_0 \geq 1$. Assume for the sake of contradiction that $h''(r_0) < 0$. Then using (20) we also get $h'''(r_0) \leq \left(1 + \frac{p}{r^p+p-1} + \frac{p-2}{r^p}\right) r^{p-1}h''(r_0) < 0$. Thus we have $h''(r) < h''(r_0)$ for all $r > r_0$. Consequently h' is also strictly decreasing by a rate of at least $h''(r_0)$ starting at r_0. This implies that there exists $r' > r_0$ with $h(r') < 0$, which contradicts (21) and thus (24) follows. Lastly (25) follows by substitution of (17) in (18) and using (24). □

4.3.3 Properties of g

In this section we will determine useful properties of g where $g(r) = -\ln(\Phi_p(-r))$. Recall from Section 2.8.1 that

$$\begin{aligned} g'(r) &= \frac{\varphi_p(r)}{1-\Phi(r)} = \frac{\exp(-|r|^p/p)}{\int_r^\infty \exp(-|t|^p/p)\,dt} \\ &= \frac{1}{\exp(|r|^p/p)\int_r^\infty \exp(-|t|^p/p)\,dt} > 0. \end{aligned}$$

For $r \in \mathbb{R}$ with $r \geq 0$ we can omit the absolute value bars. Note that $\frac{1}{g'(r)} = \exp(|r|^p/p)\int_r^\infty \exp(-|t|^p/p)\,dt = h(r)$. Our aim is to characterize the tail behavior of the p-generalized normal distribution.

Lemma 4.5. *The function g is convex and strictly increasing. Further for any $r \geq 0$ we have*

$$g'(r) \geq r^{p-1},$$

for any $r \geq 1$ we have

$$g'(r) \leq r^{p-1} + \frac{p-1}{r}.$$

and there exists a constant $c_1 > 0$ such that

$$g(r) \geq c_1 e^{-2|r|^p/p}$$

for any $r < 0$.

Proof. First note that $g = -\ln(\Phi_p(-r))$ is strictly increasing since $\Phi(-r) \in (0,1)$ is strictly decreasing for increasing r and $-\ln(t)$ is strictly increasing for decreasing t. Next consider $r \geq 0$. Then $g''(r) = \left(\frac{1}{h(r)}\right)' = -\frac{h'(r)}{h(r)^2} > 0$ by (22) thus g is convex on $[0,\infty)$. For $r < 0$ we have derived in Section 2.8.1 that

$$g''(r) = g'(r)(g'(r) - \operatorname{sgn}(r)|r|^{p-1}).$$

For $r < 0$ all terms are positive. Thus $g(r)$ is convex for all $r \in \mathbb{R}$. The bounds for g' follow immediately by the bounds for h from Lemma 4.4 (23) and (25).

Now, for $r < -1$ using the Taylor series of $-\ln(t)$ at $t = 1$, the normalizing constant

$$C_p = \int_{-\infty}^{\infty} \exp(-|t|^p/p)\, dt = \frac{2\Gamma(1/p)}{p^{1-1/p}}$$

and Equation (25) we have

$$\begin{aligned} g(r) &= -\ln\left(1 - C_p^{-1}\int_{-r}^{\infty} \exp(-|t|^p/p)\, dt\right) \\ &\geq C_p^{-1}\int_{-r}^{\infty} \exp(-|t|^p/p)\, dt \\ &\geq C_p^{-1}\exp(-(-r)^p/p)\cdot\frac{-r}{(-r)^p + p - 1} \geq \frac{\exp(-2(-r)^p/p)}{pC_p}. \end{aligned}$$

For any $r \in [-1,0]$ we have $g(r) \geq g(-1) \geq g(-1)\exp(-2(-r)^p/p)$. Thus for $c_1 = \min\{g(-1), 1/(pC_p)\}$ we have $g(r) \geq c_1 \exp(-(-r)^p/p)$. □

These properties can be used to prove the following lemma:

Lemma 4.6. *Set $G_p^+(r) = \frac{r^p}{p}$ if $r \geq 0$ and $G_p^+(r) = 0$ if $r < 0$. There exists $c_2 > 0$ depending only on p such that for any $\varepsilon \in (0, e^{-1})$ and any $r \in \mathbb{R}$ it holds that*

$$G_p^+(r) \leq g(r) \leq (1+\varepsilon)G_p^+(r) + c_2 \ln\left(\frac{p}{\varepsilon}\right). \tag{31}$$

Proof. For $r < 0$ we have $g(r) > 0 = G_p^+(r)$. For $r \geq 0$ by using Lemma 4.5 we get

$$\begin{aligned} g(r) \geq g(0) + \int_0^r g'(t)\, dt &\geq g(0) + \int_0^r t^{p-1}\, dt \\ &= g(0) + G_p^+(r) \geq G_p^+(r). \end{aligned}$$

For the second inequality we split the domain of g into three parts: First since g is monotonically increasing

for any r, we have $g(r) \leq g(1)$ for $r \in (-\infty, 1]$. For $r \geq 1$, by using Lemma 4.5, it holds that

$$\begin{aligned} g(r) &\leq g(1) + \int_1^r t^{p-1} + \frac{p-1}{t}\, dt \\ &= g(1) + G_p^+(r) - \frac{1}{p} + (p-1)\ln(r). \end{aligned} \tag{32}$$

Now consider $r \in (1, r_0]$ where $r_0 = \frac{p^3}{\varepsilon^3}$. Then we have

$$g(r) \leq g(1) + G_p^+(r) + (p-1)\ln(r_0) = g(1) + G_p^+(r) + 3(p-1)\ln\left(\frac{p}{\varepsilon}\right)$$

Our last step is to show that for $r > r_0$ it holds that $(p-1)\ln(r) \leq \varepsilon G_p^+(r)$. We assume without loss of generality that $\varepsilon^{-1} \geq 2$. Now the equation

$$\varepsilon G_p^+(r) = \varepsilon \frac{r^p}{p} \geq (p-1)\ln(r)$$

is equivalent to

$$\exp\left(\frac{\varepsilon r^p}{p^2 - p}\right) \geq r.$$

Note that $r^p \geq r$ holds since $r \geq r_0 > 1$ and thus we get for any $r = ar_0$ with $a \geq 1$ that

$$\exp\left(\frac{\varepsilon r^p}{p^2 - p}\right) \geq \exp\left(\frac{\varepsilon r}{p^2}\right) \geq \exp\left(\frac{\varepsilon a r_0}{p^2}\right) \geq \exp\left(\frac{ap}{\varepsilon^2}\right) \geq \exp\left(2a \cdot \frac{p}{\varepsilon}\right) \geq ar_0 = r.$$

The last inequality follows from the fact that $e^{2az} \geq az^3$ always holds in our case where $z \geq 2$ and $a \geq 1$. Consequently it holds for any $r \in [r_0, \infty)$ that

$$g(r) \leq g(1) + G_p^+(r) + (p-1)\ln(r) \leq g(1) + (1+\varepsilon)G_p^+(r).$$

Combining all three inequalities we note that for any $r \in \mathbb{R}$ it holds that

$$\begin{aligned} g(r) &\leq g(1) + (1+\varepsilon)G_p^+(r) + (p-1)\ln\left(\frac{p^3}{\varepsilon^3}\right) \\ &= (1+\varepsilon)G_p^+(r) + \left(\frac{g(1)}{\ln(p/\varepsilon)} + 3(p-1)\right)\ln\left(\frac{p}{\varepsilon}\right) \\ &\leq (1+\varepsilon)G_p^+(r) + c_2 \ln\left(\frac{p}{\varepsilon}\right) \end{aligned}$$

where $c_2 := (g(1) + 3(p-1)) \geq (\frac{g(1)}{\ln(p/\varepsilon)} + 3(p-1))$ holds, since $\varepsilon^{-1} \geq e$ and $p \geq 1$. □

For the p probit loss we can get a similar result as for the logistic loss in Lemma 3.3:

Lemma 4.7. *Assume $X \in \mathbb{R}^{n\times d}$ is μ-complex. Then we have for any $\beta \in \mathbb{R}^d$ that*

$$f(X\beta) = \Omega\left(\frac{n}{\mu}\left(1+\ln(\mu)\right)\right).$$

Proof. Let $z = X\beta$. For $r \leq 0$ we have $g(r) \geq c_1 e^{-2|r|^p/p}$ by Lemma 4.5. For $r \geq 0$ we have $g(r) = g(0) + \int_0^r g'(t)\,dt$. Recall that

$$g'(t) = \frac{1}{h(t)} \geq t^{p-1}$$

and thus

$$g(r) \geq g(0) + \int_0^r t^{p-1}\,dt = g(0) + \frac{r^p}{p}. \tag{33}$$

Set $z_- = \frac{1}{n}\sum_{z_i\leq 0}|z_i|^p$ and $z_+ = \frac{1}{n}\sum_{z_i\geq 0}|z_i|^p \geq \frac{z_-}{\mu}$. We set $z^- \in \mathbb{R}^n$ to be the vector with $z_i^- = z_i$ if $z_i < 0$ and $z_i^- = 0$ else. Using convexity of e^{-r} we can apply Jensens inequality to conclude that

$$\begin{aligned}
f(X\beta) &= \sum_{i=1}^n g(z_i) \\
&\geq \sum_{i=1}^n \min\{g(z_i), g(0)\} \;+\; \sum_{z_i\geq 0}\int_0^{z_i} t^{p-1}\,dt \\
&\geq \sum_{i=1}^n ce^{-2|z_i^-|^p/p} + \frac{1}{p}\sum_{z_i\geq 0} z_i^p \\
&\geq nc_1 e^{-2(z_-)/p} + \frac{nz_+}{p} \\
&\geq nc_1 e^{-2(z_-)/p} + \frac{nz_-}{\mu p}.
\end{aligned}$$

Taking the derivative of $\ell(r) = nc_1 e^{-(r)/p} + \frac{nr}{\mu p}$, i.e. $\ell'(r) = \frac{n}{p}(-c_1 e^{-2(r)/p} + \frac{1}{\mu})$ which is 0 if $\frac{r}{p} = \ln(c_1\mu)/2$. Thus it holds that

$$f(X\beta) \geq \ell(z_-) \geq \frac{n}{2\mu}(1+\ln(c_1\mu))$$

which is exactly what we needed to show. □

Using similar arguments as in the previous lemma we can further show the following:

Lemma 4.8. *Let $\beta^* \in \mathbb{R}^d$ be the minimizer of $\min_{\beta\in\mathbb{R}^d} f(X\beta)$. Then there exists some constants c_0 such that the following hold:*

1) It holds that $\|X\beta^*\|_p^p \leq pn\ln(c_0\mu)$;

2) For any $\beta \in \mathbb{R}^d$ with $\|X\beta\|_p^p \geq pn\ln(2c_0\mu) + \varepsilon n\mu/2$ it holds that $f(X\beta) \geq f(X\beta^*)(1+\varepsilon)$.

Proof. For $r \leq 0$ we have that $g'(r) \leq \varphi_p(r)/2 \leq c_0 \exp(-|r|^p/p)$ for some constant c_0 as $1 - \Phi_p(r) \leq 1/2$. For $r \geq 0$ we have that $g'(r) \geq r^{p-1}$ by Lemma 4.5.

1) This follows by using the same argumentation as in the proof of Lemma 4.7: We set $\beta = \beta^*/\|X\beta^*\|_p$ and $z = X\beta$. Now consider the function $f_1(r) = f(X\beta r^{1/p}) = \sum_{i=1}^n g(x_i \beta r^{1/p})$. We have that

$$\begin{aligned}
f_1'(r) &= \sum_{i=1}^n \frac{z_i r^{1/p-1}}{p} \cdot g'(z_i r^{1/p}) \\
&= \sum_{z_i \geq 0} \frac{z_i r^{1/p-1}}{p} \cdot g'(z_i r^{1/p}) \;+\; \sum_{z_i < 0} \frac{z_i r^{1/p-1}}{p} \cdot g'(z_i r^{1/p}) \\
&\geq \sum_{z_i \geq 0} \frac{z_i r^{1/p-1}}{p} \cdot (z_i r)^{p-1} \;-\; r^{1/p-1} \sum_{z_i < 0} c_0 |z_i| \exp(-|z_i r|^p/p) \\
&= \sum_{z_i \geq 0} \frac{z_i^p}{p} \;-\; r^{1/p-1} \sum_{z_i < 0} c_0 |z_i| \exp(-|z_i|^p r/p).
\end{aligned}$$

Set $z_- = \|z\|_p^p/n$. Observe that $\sum_{z_i<0} c_0 |z_i| \exp(-|z_i|^p r/p) \leq c_0 z_- \exp(-z_- r/p) \leq \frac{c_0}{n} \cdot \exp(-r/np)$. Using the convexity of the exponential function as in previous proof we get that for $r \geq 1$ it holds that

$$f_1'(r) \geq \|z^+\|_p^p - c_0 z_- n \exp(-z_- r/p) \geq \frac{1}{\mu} - c_0 \exp(-r/pn).$$

Consequently for $r > r_m := pn \ln(c_0 \mu)$ we have that $f_1'(r) > 0$ and thus, as for $X\beta r = X\beta^*$ it holds that $f_1'(r) = 0$, it must hold that

$$\|X\beta^*\|_p^p \leq \|X\beta r_m^{1/p}\|_p^p = r_m = 1 \cdot pn \ln(c_0 \mu).$$

2) We are using the same argumentation as in 1) but here we consider any β with $\|\beta\|_p^p = 1$. Define $f_1(r)$ as before. Using the equation from part 1) we get that for $r \geq pn \ln(2c_0\mu)$ it holds that $f_1'(r) \geq \frac{1}{2\mu}$. Since $f(X\beta^*) \leq f(0) = n \ln(2) < n$ we conclude that if $r \geq pn \ln(2c_0\mu) + \varepsilon n \mu/2$ it holds that

$$f(X\beta r) \geq f(X\beta r_m) + \int_{r_m}^r f_1'(t) dt \geq f(X\beta^*)(1+\varepsilon).$$

□

Consequently the optimum β cannot be large and any β with $f(X\beta) \leq f(X\beta^*)(1+\varepsilon)$ must be close to β^*:

Corollary 4.9. *For any $\beta \in \mathbb{R}^d$ with $\|X\beta - X\beta^*\|_p^p \geq 2pn \ln(2c_0\mu) + \varepsilon n\mu/2$ it holds that $f(X\beta) \geq f(X\beta^*)(1+\varepsilon)$.*

4.3.4 Bounding the VC-Dimension

In order to apply the sensitivity framework we need to bound both, the VC-dimension of the appropriate range space as well as the sensitivities. In this subsection we look at the VC-dimension. We do not know any bounds on the VC-dimension of the range space of the function space $\mathcal{F} = \{g_{x,w} \mid w \in \mathbb{R}_{\geq 0}, x \in \mathbb{R}^d\}$ where $g_{x,w}(\beta) = wg(x\beta)$. However if we limit the number of weights w allowed then we are able to bound the VC-dimension. Thus order to bound the VC-dimension of the range space induced by the weighted set of functions we reduce the number of distinct weights considered. We first round all sensitivities to their closest power of 2. The new total sensitivity S' is at most twice the old sensitivity S. Next we increase all sensitivities smaller than $\frac{S}{n}$ to $\frac{S}{n}$. The new sensitivity is at most $S' + n \cdot S/n = 3S$. The next step is to split the data into *high sensitivity* points and *low sensitivity* points.

Lemma 4.10. *Let I_1 be the index set of all data points with $s_i > s_0 := \frac{\mu S c \ln(p\varepsilon^{-1})}{\varepsilon n}$ for some constant $c \in \mathbb{R}_{>0}$. Then for all $\beta \in \mathbb{R}^d$ it holds that*

$$\sum_{i \in I_1} G_p^+(x_i\beta) \leq \sum_{i \in I_1} g(x_i\beta) \leq (1+\varepsilon) \sum_{i \in I_1} G_p^+(x_i\beta) \ + \ \varepsilon \cdot \frac{n}{\mu}.$$

Proof. We set $c = c_2$ as in Lemma 4.6. Note that there are at most $\frac{S}{s_0} = \frac{\varepsilon n}{c \ln(p\varepsilon^{-1})\mu}$ points in I_1. Thus the lemma follows by applying Lemma 4.6 to each point in I_1. □

As a consequence we get the following corollary:

Corollary 4.11. *Let $I_2 = [n] \setminus I_1$. Further let $(X', w) \in \mathbb{R}^{n' \times d} \times \mathbb{R}^{n'}$ with rows $x'_i = x_{\pi(i)}$ for some mapping $\pi : [n'] \to [n]$. We set $I'_1 = \{i \in [n'] \mid \pi(i) \in I_1\}$ and similarly $I'_2 = \{i \in [n'] \mid \pi(i) \in I_2\}$. Further define $\tilde{f}_w(X'\beta) = \sum_{i \in I'_2} w_i g(x'_i\beta) \ + \ \sum_{i \in I'_1} w_i G_p^+(x'_i\beta)$ and by $\tilde{f}(X\beta) = \sum_{i \in I_2} g(x_i\beta) \ + \ \sum_{i \in I_1} G_p^+(x_i\beta)$. Assume that for all $\beta \in \mathbb{R}^d$ it holds*

$$|\tilde{f}_w(X'\beta) - \tilde{f}(X\beta)| \leq \varepsilon \tilde{f}(X\beta) \tag{34}$$

and $\sum_{i \in I'_1} w_i \leq \frac{2S}{s_0}$. Further assume that $\varepsilon \leq \frac{1}{4}$. Then (X', w) is a 7ε-coreset for the original f.

Proof of Corollary 4.11. Observe that by triangle inequality

$$|f_w(X'\beta) - f(X\beta)| \leq |f_w(X'\beta) - \tilde{f}_w(X'\beta)| + |\tilde{f}_w(X'\beta) - \tilde{f}(X\beta)| + |\tilde{f}(X\beta) - f(X\beta)| \tag{35}$$

By Lemma 4.10 it holds that

$$\begin{aligned} \tilde{f}(X\beta) \leq f(X\beta) &\leq \tilde{f}(X\beta) + \varepsilon \sum_{i \in I_1} G_p^+(x_i\beta) + \varepsilon \cdot \frac{n}{\mu} \\ &\leq \tilde{f}(X\beta) + 2\varepsilon f(X\beta) \end{aligned}$$

We thus have that

$$|\tilde{f}(X\beta) - f(X\beta)| \leq 2\varepsilon f(X\beta).$$

Analogously to Lemma 4.10, using the bounded size of $\sum_{i\in I_1'} w_i$ and the assumption (34) one can show that

$$\begin{aligned}
\tilde{f}_w(X'\beta) \leq f_w(X'\beta) &\leq \tilde{f}_w(X'\beta) + \varepsilon \sum_{i\in I_1'} w_i G_p^+(x_i'\beta) + \sum_{i\in I_1'} w_i \cdot c_2 \ln\left(\frac{p}{\varepsilon}\right) \\
&\leq \tilde{f}_w(X'\beta) + \varepsilon\tilde{f}_w(X'\beta) + \frac{2S}{s_0} \cdot c_2 \ln\left(\frac{p}{\varepsilon}\right) \\
&\overset{(34)}{\leq} \tilde{f}_w(X'\beta) + \varepsilon(1+\varepsilon)\tilde{f}(X\beta) + 2\varepsilon\frac{n}{\mu} \\
&\leq \tilde{f}_w(X'\beta) + 2\varepsilon\tilde{f}(X\beta) + 2\varepsilon f(X\beta) \\
&\leq \tilde{f}_w(X'\beta) + 4\varepsilon f(X\beta)
\end{aligned}$$

and thus we have

$$|f_w(X'\beta) - \tilde{f}_w(X'\beta)| \leq 4\varepsilon f(X\beta).$$

Now combining everything into Equation (35) yields

$$\begin{aligned}
|f_w(X'\beta) - f(X\beta)| &\leq |f_w(X'\beta) - \tilde{f}_w(X'\beta)| + |\tilde{f}_w(X'\beta) - \tilde{f}(X\beta)| + |\tilde{f}(X\beta) - f(X\beta)| \\
&\leq 4\varepsilon f(X\beta) + \varepsilon\tilde{f}(X\beta) + 2\varepsilon f(X\beta) \\
&\leq 7\varepsilon f(X\beta)
\end{aligned}$$

and thus (X', w) is a 7ε-coreset. □

Before we continue showing that for the set of functions that we consider, the VC-dimension is not too large, we show that the assumption made in Corollary 4.11 that $\sum_{i\in I_1'} w_i \leq \frac{2S}{s_0}$ is reasonable, i.e., that it holds with high probability in our context:

Lemma 4.12. *Assume, as in the context of Proposition 2.23, that for R with $|R| = k$ where each element of R is sampled i.i.d. with probability $p_j = \frac{s_j}{S}$ from $\mathcal{F}$ and $w_i = \frac{S}{s_j|R|} = \frac{1}{kp_j}$ denotes the weight of a function $f_i \in R$ that corresponds to $f_j \in \mathcal{F}$. Then with probability at least $1 - \frac{1}{k}$ it holds that $\sum_{i\in I_1'} w_i \leq \frac{2S}{s_0}$.*

Proof of Lemma 4.12. Let $x_{\pi(i)}$ be the ith element of R. We set $Z_i = w_{\pi(i)}$ if $\pi(i) \in I_1$ and $Z_i = 0$ otherwise. Then $\lambda = \mathbb{E}(Z_i) = \sum_{j\in I_1} p_j w_j = \sum_{j\in I_1} p_j \frac{1}{kp_j} = \frac{|I_1|}{k}$. Recall from Lemma 4.10 that $|I_1| \leq \frac{S}{s_0}$ and for $j \in I_1$ we have $s_j > s_0$. For the variance it follows

$$\mathbb{E}[(Z_i - \lambda)^2] = \mathbb{E}[Z_i^2] - \mathbb{E}[Z_i]^2 \leq \mathbb{E}[Z_i^2] = \sum_{j\in I_1} p_j w_j^2 = \sum_{j\in I_1} \frac{1}{k^2 p_j} \leq \sum_{j\in I_1} \frac{S}{k^2 s_0} = |I_1| \cdot \frac{S}{k^2 s_0} \leq \frac{S^2}{s_0^2 k^2}.$$

Thus by independence of the Z_i the variance of $Z = \sum_{i=1}^{k} Z_i$ is bounded by $\frac{S^2}{s_0^2 k}$. Now applying Chebyshev's inequality yields

$$P\left(Z \geq 2 \cdot \frac{S}{s_0}\right) \leq P\left(Z - \mathbb{E}(Z) \geq \frac{S}{s_0}\right) \leq \frac{\mathrm{Var}(Z)}{S/s_0} \leq \frac{S^2/(s_0^2 k)}{S^2/s_0^2} = \frac{1}{k}.$$

□

By the technical Corollary 4.11 our goal of obtaining a coreset for f reduces to obtaining a coreset for the substitute function

$$\tilde{f}(X\beta) = \sum_{i \in [n] \setminus I_1} g(x_i\beta) + \sum_{i \in I_1} G_p^+(x_i\beta).$$

To this end we set $\mathcal{F}_1 = \{w_i G_i^+ \mid i \in I_1\}$ where $G_i^+(\beta) = G_p^+(x_i\beta)$ and $\mathcal{F}_2 = \{w_i g_i \mid i \in I_2 = [n] \setminus I_1\}$ where $g_i(\beta) = g(x_i\beta)$. Further we set $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$ and show that the VC-dimension of $\mathcal{F}$ can be bounded as desired:

Lemma 4.13. *For the VC-dimension Δ of $\mathfrak{R}_{\mathcal{F}}$ we have*

$$\Delta \leq (d+1)\left(\log_2\left(\mu c \varepsilon^{-2}\right) + 2\right) = O(d \log(\mu/\varepsilon)).$$

Proof. First note that for any $G \subseteq \mathcal{F}_1$, $\beta \in \mathbb{R}^d$ and $r \in \mathbb{R}$ it holds that $\mathrm{range}_{\mathcal{F}_1}(\beta, r) \cap G = \mathrm{range}_G(\beta, r)$. We show that the VC-dimension of $\mathfrak{R}_{\mathcal{F}_1}$ is at most $d+1$. Indeed, it holds that $\mathrm{range}_{\mathcal{F}_1}(\beta, r) = \mathcal{F}_1$ if $r \leq 0$ since all weights are positive and G_p^+ is also positive. Otherwise we have that

$$\begin{aligned}
\mathrm{range}_{\mathcal{F}_1}(\beta, r) &= \{w_i G_i^+ \in \mathcal{F}_1 \mid w_i G_i^+(\beta) \geq r\} \\
&= \{w_i G_i^+ \in \mathcal{F}_1 \mid w_i (x_i\beta)^p / p \geq r \wedge x_i\beta > 0\} \\
&= \left\{w_i G_i^+ \in \mathcal{F}_1 \mid x_i\beta \geq \left(\frac{pr}{w_i}\right)^{1/p}\right\}.
\end{aligned}$$

We conclude that for $G \subseteq \mathcal{F}_1$ it holds that

$$\begin{aligned}
|\{G \cap R \mid R \in \mathrm{ranges}(\mathcal{F}_1)\}| &= |\{\mathrm{range}_G(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{>0}\} \cup \{\mathrm{range}_G(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\leq 0}\}| \\
&= \left|\left\{\left\{w_i G_i^+ \in G \mid x_i\beta \geq (pr/w_i)^{1/p}\right\} \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\right\} \cup \{G\}\right| \\
&\leq |\{\{w_i G_i^+ \in G \mid x_i\beta - s \geq 0\} \mid \beta \in \mathbb{R}^d, s \in \mathbb{R}\}|
\end{aligned}$$

which corresponds to a set of affine hyperplane classifiers $\beta \mapsto \mathbf{1}_{x_i\beta - s \geq 0}$, which have VC-dimension $d+1$ (Kearns and Vazirani, 1994). Thus, the induced range space $\mathfrak{R}_{\mathcal{F}_1}$ has VC-dimension at most $d+1$.
Next consider $\mathcal{F}_2$. Note that g is a strictly monotonic and thus also invertible function. First fix a weight

$v \in \mathbb{R}_{>0}$ and let $\mathcal{F}_v = \{w_i g_i \mid w_i = v\}$. We have

$$\begin{aligned}\operatorname{range}_{\mathcal{F}_v}(\beta, r) &= \{w_i g_i \in \mathcal{F}_v \mid w_i g_i(\beta) \geq r\} \\ &= \left\{w_i g_i \in \mathcal{F}_v \mid x_i\beta \geq g^{-1}\left(\frac{r}{v}\right)\right\}\end{aligned}$$

which corresponds to a set of points shattered by the affine hyperplane classifier $\beta \mapsto \mathbf{1}_{x_i\beta - g^{-1}\left(\frac{r}{v}\right)\geq 0}$ and thus the VC-dimension of the induced range space $\mathfrak{R}_{\mathcal{F}_v}$ is at most $d+1$. Let W be the set of all weights for functions in $\mathcal{F}_2$. Since all weights are powers of 2, and we have $\frac{S}{n} \leq v \leq \frac{\mu S c \ln(p\varepsilon^{-1})}{\varepsilon n}$ it holds that$|W| \leq \log_2(\frac{\mu c \ln(p\varepsilon^{-1})}{\varepsilon}) \leq \left(\log_2\left(\mu c \varepsilon^{-2}\right) + 2\right)$. Now we claim that the VC-dimension of $\mathfrak{R}_{\mathcal{F}}$ is at most $(|W|+1)(d+1)$ as $\mathcal{F} = \mathcal{F}_1 \cup \bigcup_{v\in W} \mathcal{F}_v$. Assume for the sake of contradiction that there exists $G \subset \mathcal{F}$ such that $|G| > (|W|+1)(d+1)$ and G is shattered by the ranges of $\mathcal{F}$. Then by the pigeonhole principle $G' = G \cap \mathcal{F}' > d+1$ for some $\mathcal{F}' \in \{F_1\} \cup \bigcup_{v\in W}\{\mathcal{F}_v\}$. But due to the pairwise disjointness of all of $\mathcal{F}_1$ and $\mathcal{F}_v$, G' must be shattered by the ranges of $\mathcal{F}'$, which contradicts that their VC-dimension is bounded by $d+1$, cf. Lemma 11 in (Munteanu et al., 2018). □

4.3.5 Bounding the Sensitivities

In this subsection we bound the sensitivities. Therefore recall that ℓ_p-leverage scores of X are defined by $u_j = \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{|x_j\beta|^p}{\sum_{i=1}^n |x_i\beta|^p}$, cf. (Dasgupta et al., 2009). Also recall that the supremum is attained by some $\beta \in \mathbb{R}^d$ since

$$\sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{|x_j\beta|^p}{\sum_{i=1}^n |x_i\beta|^p} = \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{\|\beta\|_2^p \cdot |x_j\beta/\|\beta\|_2|^p}{\|\beta\|_2^p \cdot \sum_{i=1}^n |x_i\beta/\|\beta\|_2|^p} = \sup_{\beta\in\mathbb{R}^d, \|\beta\|_2=1} \frac{|x_j\beta|^p}{\sum_{i=1}^n |x_i\beta|^p}$$

and $\{\beta \in \mathbb{R}^d \mid \|\beta\|_2 = 1\}$ is a compact set. We also note that $u_p \leq 1$ always holds. The ℓ_p-leverage scores can be used to bound the sensitivities:

Lemma 4.14. *There is a constant c_s such that the sensitivity ζ_i of $x_i, i \in [n]$ for $\tilde{f}$ is bounded by*

$$\zeta_i \leq c_s \mu \left(\frac{1}{n} + u_i\right)$$

Proof of Lemma 4.14. First note that by Lemma 4.7 it holds that $f(X\beta) \geq \frac{n}{\mu}$ and thus by Lemma 4.10 $\tilde{f}(X\beta) \geq f(X\beta) - \varepsilon \sum_{i\in I_1}(x_i\beta) - \varepsilon \cdot \frac{n}{\mu} \geq \frac{f(X\beta)}{2} \geq \frac{n}{2\mu}$ holds for small enough $\varepsilon \leq 1/4$. Thus for any β with $x_i\beta \leq 1$ we have $\frac{(x_i\beta)}{\tilde{f}(X\beta)} \leq \frac{g(x_i\beta)}{\tilde{f}(X\beta)} \leq \frac{g(1)}{n/2\mu} = 2g(1)\frac{\mu}{n}$.

Further for β with $x_i\beta > 1$ it holds that $g(x_i\beta) \leq c_3(x_i\beta)^p$ for some constant $c_3 \leq 2g(1)+1$ since by

Lemma 3.1 and using $\frac{p-1}{t} \leq p-1$ for $t \geq 1$ it holds that

$$\begin{aligned}
g(x_i\beta) &= g(1) + \int_1^{x_i\beta} g'(t)\,dt \\
&\leq g(1) + \int_1^{x_i\beta} t^{p-1} + \frac{p-1}{t}\,dt \\
&\leq g(1) + \int_1^{x_i\beta} t^{p-1} + p - 1\,dt \\
&\leq g(1) + \int_1^{x_i\beta} t^{p-1} + (p-1)t^{p-1}\,dt \\
&\leq g(1) + \int_1^{x_i\beta} pt^{p-1}\,dt \\
&= g(1) + (x_i\beta)^p - 1 \leq (g(1)+1)(x_i\beta)^p.
\end{aligned}$$

Also note that by definition of μ it holds that

$$\frac{1}{\sum_{x_j\beta>0} |x_j\beta|^p(1+\mu)} \leq \frac{1}{\sum_{x_j\beta>0} |x_j\beta|^p + \sum_{x_j\beta<0} |x_j\beta|^p} = \frac{1}{\sum_{j=1}^n |x_j\beta|^p}$$

and thus

$$\frac{1}{\sum_{x_j\beta>0} |x_j\beta|^p} \leq \frac{1+\mu}{\sum_{j=1}^n |x_j\beta|^p}.$$

Now setting $c_3 = 2g(1)+1$ and using $\tilde{f}(X\beta) \geq \sum_{x_j\beta>0} \frac{|x_j\beta|^p}{p} = \sum_{j=1}^n (x_j\beta)$ we get

$$\frac{(x_i\beta)}{\tilde{f}(X\beta)} \leq \frac{g(x_i\beta)}{\tilde{f}(X\beta)} \leq \frac{c_3}{1/p} \cdot \frac{|x_i\beta|^p}{\sum_{x_j\beta>0} |x_j\beta|^p} \leq pc_3(1+\mu)u_i \leq 2pc_3\mu u_i := c_s\mu u_i.$$

Combining both bounds gives us the bound for ζ_i. □

4.3.6 Well Conditioned Bases and Approximate Leverage Scores

In order to approximate the leverage scores we will need well conditioned bases:

An (α, β, p)-well-conditioned basis V is a basis that preserves the norm of each vector well, in the sense that its entry-wise p norm $\|V\|_p \leq \alpha$ and for all $z \in \mathbb{R}^d$: $\|z\|_q \leq \beta\|Vz\|_p$, where q denotes the dual norm to p, i.e., $\frac{1}{p} + \frac{1}{q} = 1$, see Definition 2.6. We will first state the properties of the (α, β, p)-*well-conditioned basis* and then we describe how to compute the basis.

Lemma 4.15. *Let V be an* (α, β, p)-well-conditioned basis *for the column space of X. Then it holds for all $i \in [n]$ that $u_i \leq \beta^p \|v_i\|_p^p$. As a direct consequence we have $\sum_{i=1}^n u_i \leq \beta^p \|V\|_p^p \leq (\alpha\beta)^p = d^{O(p)}$.*

Proof. We have by a change of basis

$$u_i = \sup_{z \in \mathbb{R}^d \setminus \{0\}} \frac{|(Xz)_i|^p}{\|Xz\|_p^p} = \sup_{z \in \mathbb{R}^d \setminus \{0\}} \frac{|(Vz)_i|^p}{\|Vz\|_p^p}.$$

Now assume that z attains the value $\sup_{z\in\mathbb{R}^d\setminus\{0\}} \frac{|(Vz)_i|^p}{\|Vz\|_p^p}$. Then we get by using Hölder's inequality and the properties of V that

$$u_i = \frac{|(Vz)_i|^p}{\|Vz\|_p^p} \leq \frac{\beta^p|(Vz)_i|^p}{\|z\|_q^p} \leq \frac{\beta^p\|v_i\|_p^p\|z\|_q^p}{\|z\|_q^p} = \beta^p\|v_i\|_p^p.$$

□

An (α,β,p)-*well-conditioned basis* can be computed using sketching techniques.

Lemma 4.16. *(Woodruff and Zhang, 2013; Clarkson and Woodruff, 2017) There exists a random embedding matrix $\Pi\in\mathbb{R}^{n'\times n}$ and $\gamma = O(d\log(d))$ such that*

$$\forall\beta\in\mathbb{R}^d:\ \frac{1}{\gamma^{1/p}}\|X\beta\|_p \leq \|\Pi X\beta\|_q \leq \gamma^{1/p}\|X\beta\|_p$$

holds with constant probability, where

$$(q,n') = \begin{cases} (2, O(d^2)) & \text{if } p\in[1,2] \\ (\infty, O(n^{1-\frac{2}{p}}\log n(d\log d)^{1+\frac{2}{p}} + d^{5+4p})) & \text{if } p\in(2,\infty). \end{cases}$$

For $p=2$ we have $\gamma=2$. Further ΠX can be computed in $O(\mathrm{nnz}(X))$ time.

The sketching matrix Π can be constructed as follows: First let $D\in\mathbb{R}^{n\times n}$ be the diagonal matrix with $D_{ii}=1$ or $D_{ii}=-1$ each with probability $1/2$. Further let $h:[n]\to[n']$ be a random map where h hashes each entry of $[n]$ to one of n' buckets uniformly at random. Set $\Psi\in\mathbb{R}^{n'\times n}$ to be the matrix where $\Psi_{h(i)i}=1$ and $\Psi_{ji}=0$ if $j\neq h(i)$. For $p=2$ it suffices to take $\Pi=\Psi D$ (Clarkson and Woodruff, 2017). Otherwise if $p\neq 2$ let E be a diagonal matrix with $E_{ii}=1/\lambda_i^{1/p}$ where $\lambda_i\sim\exp(1)$ is drawn from a standard exponential distribution and set $\Pi=\Psi DE$ (Woodruff and Zhang, 2013).

Lemma 4.17. *Munteanu (2018) If Π satisfies Lemma 4.16 and $\Pi X = QR$ is the QR-decomposition of ΠX then $V = XR^{-1}$ is an (α,β,p)*-well-conditioned basis *for the columnspace of X, where for $\gamma = O(d\log(d))$ we have*

$$(\alpha,\beta) = \begin{cases} (\sqrt{2d},\sqrt{2}), & \text{for } p=2 \\ (d\gamma^{1/p},\gamma^{1/p}), & \text{for } p\in[1,2) \\ (d\gamma^{1/p},d\gamma^{1/p}), & \text{for } p\in(2,\infty). \end{cases}$$

Proof. We are going to use the fact that $Q=\Pi XR$ is an orthonormal basis. Let e_i for $i\in[d]$ denote the ith

standard basis vector. We define $(R^{-1})^{(i)}$ to be the ith column of R^{-1}. We have

$$\begin{aligned}\|V\|_p = \left\|XR^{-1}\right\|_p = \left\|X\sum_{i=1}^{d}(R^{-1})^{(i)}e_i^T\right\|_p = \left\|\sum_{i=1}^{d}X(R^{-1})^{(i)}e_i^T\right\|_p \\ \leq \sum_{i=1}^{d}\left\|X(R^{-1})^{(i)}e_i^T\right\|_p = \sum_{i=1}^{d}\left\|X(R^{-1})^{(i)}\right\|_p \end{aligned} \tag{36}$$

Now suppose $p \in (2, \infty)$.

$$\begin{aligned}(36) \leq \gamma^{1/p}\sum_{i=1}^{d}\left\|\Pi X(R^{-1})^{(i)}\right\|_\infty \leq \gamma^{1/p}\sqrt{d}\left(\sum_{i=1}^{d}\left\|\Pi X(R^{-1})^{(i)}\right\|_\infty^2\right)^{\frac{1}{2}} \\ \leq \gamma^{1/p}\sqrt{d}\left(\sum_{i=1}^{d}\left\|\Pi X(R^{-1})^{(i)}\right\|_2^2\right)^{\frac{1}{2}} \leq \gamma^{1/p}\sqrt{d}\left(\sum_{i=1}^{d}\underbrace{\left\|Q^{(i)}\right\|_2^2}_{=1}\right)^{\frac{1}{2}} = \gamma^{1/p}d\end{aligned}$$

For arbitrary $z \in \mathbb{R}^d$ it holds that

$$\|z\|_q \leq \sqrt{d}\,\|z\|_2 = \sqrt{d}\,\|Qz\|_2 = \sqrt{d}\left\|\Pi XR^{-1}z\right\|_2 \leq d\left\|\Pi XR^{-1}z\right\|_\infty \leq d\gamma^{1/p}\|Vz\|_p.$$

Consequently V is $(\gamma^{1/p}d, \gamma^{1/p}d, p)$-well-conditioned.

Next suppose $p \in [1, 2)$. Again we bound

$$\begin{aligned}(36) \leq \gamma^{1/p}\sum_{i=1}^{d}\left\|\Pi X(R^{-1})^{(i)}\right\|_2 \leq \gamma^{1/p}\sqrt{d}\left(\sum_{i=1}^{d}\left\|\Pi X(R^{-1})^{(i)}\right\|_2^2\right)^{\frac{1}{2}} \\ \leq \gamma^{1/p}\sqrt{d}\left(\sum_{i=1}^{d}\underbrace{\left\|Q^{(i)}\right\|_2^2}_{=1}\right)^{\frac{1}{2}} = \gamma^{1/p}d\end{aligned}$$

Also, since $p \leq 2$, the dual norm satisfies $q \geq 2$. Fix an arbitrary $z \in \mathbb{R}^d$. It follows that

$$\|z\|_q \leq \|z\|_2 = \|Qz\|_2 = \left\|\Pi XR^{-1}z\right\|_2 \leq \gamma^{1/p}\|Vz\|_p.$$

It follows that V is even $(\gamma^{1/p}d, \gamma^{1/p}, p)$-well-conditioned in this case.

Finally suppose $p = 2$, where the entry-wise matrix norm is the Frobenius norm $\|\cdot\|_F$. We have

$$\begin{aligned}\|V\|_F^2 = \sum_{i=1}^{d}\left\|V^{(i)}\right\|_2^2 = \sum_{i=1}^{d}\left\|(XR^{-1})^{(i)}\right\|_2^2 \leq \sum_{i=1}^{d}2\left\|(\Pi XR^{-1})^{(i)}\right\|_2^2 \\ = 2\sum_{i=1}^{d}\left\|Q^{(i)}\right\|_2^2 = 2d.\end{aligned}$$

Thus we have $\|V\|_F = \sqrt{2d}$. Since $p = q = 2$ we have for any $\beta \in \mathbb{R}^d$ that

$$\|z\|_q = \|z\|_2 = \|Qz\|_2 = \left\|\Pi X R^{-1} z\right\|_2 \leq \sqrt{2}\,\|Vz\|_p\,.$$

Consequently V is a $(\sqrt{2d}, \sqrt{2}, p)$-well-conditioned in this case. □

4.4 Main Results

Our main result shows that if μ is small, then there exists a small coreset C. In fact, the size of C does not depend on n at all and it can be computed efficiently in two passes over the data:

Theorem 4. *If $X \in \mathbb{R}^{n\times d}$ is μ-complex for any fixed $p \in [1,\infty)$ then with constant probability we can compute an ε-coreset $C = (X', w)$ for p-probit regression of size $k = O(\frac{S}{\varepsilon^2}(d\ln(\varepsilon^{-1}\mu)\ln S))$ in two passes over the data, where*

$$S = \begin{cases} O\left(\mu d\right), & \text{for } p = 2 \\ O\left(\mu d^p (d\log d)^2\right), & \text{for } p \in [1,2) \\ O\left(\mu d^{2p} (d\log d)^2\right), & \text{for } p \in (2,\infty). \end{cases}$$

Algorithm 2 runs in $O(\mathrm{nnz}(X)d + \mathrm{poly}(d))$ time for $p \in [1,2]$ and in $O(\mathrm{nnz}(X)d + \mathrm{poly}(d)n^{1-\frac{2}{p}}\log n)$ time for $p > 2$, were nnz *denotes the number of non-zeros.*

Proof. We first describe Algorithm 2 and its running time: First we apply our sketching matrix Π from Lemma 4.16 to compute ΠX in time $O(\mathrm{nnz}(X))$ in one pass over the data. The number of rows is $n' = O(d^2)$ for $p \in [1,2]$ and $n' = O(n^{1-\frac{2}{p}}\log n\,\mathrm{poly}(d))$ for $p > 2$. Then we calculate the QR-decomposition of $\Pi X = QR$ in time $O(d^4)$ respectively in time $O(n^{1-\frac{2}{p}}\log n\,\mathrm{poly}(d))$ depending on the value of p, which is faster than $O(nd^2)$ without sketching. In a second pass over the data we compute the row norms $\|v_i\|_p^p = \|x_i R^{-1}\|_p^p$ used in our sampling probabilities. We set $s_i = \frac{1}{n} + \beta^p \|v_i\|_p^p$. By Lemma 2.7, $S_0 = 1 + (\alpha\beta)^p$ is an upper bound for $\sum_{i=1}^n s_i$. Next we set $s_i' = \max\{2^{\lceil \log_2(s_i)\rceil}, \frac{S_0}{n}\}$, i.e. we round s_i to the next power of 2 such that $S' = \sum_{i=1}^n s_i' \leq 2S_0 + n\cdot\frac{S_0}{n} = 3S_0$. As we calculate those values, we can feed the point x_i augmented with the corresponding sampling weight s_i' directly to k independent copies of a weighted reservoir sampler (Chao, 1982). The latter is an online algorithm and updates its sample in constant time. The second pass takes $O(\mathrm{nnz}(X)d + \mathrm{poly}(d))$ time for $p \in [1,2]$, respectively $O(\mathrm{nnz}(X)d + \mathrm{poly}(d)\,n^{1-\frac{2}{p}}\log n)$ for $p > 2$. Lemma 4.14 yields $S = \sum_{i\in[n]} c_s \mu(1/n + u_i) = O(\mu\sum_{i\in[n]} u_i)$ and by Lemma 2.7 we have that $\sum_{i\in[n]} u_i \leq (\alpha\beta)^p = d^{O(p)}$ where the values of α, β are detailed in Lemma 4.17. Using Lemmas 4.13, 4.14, and 2.7 to bound the parameters of Proposition 2.23 we get for the substitute function $\tilde{f}$ that $\forall\beta\colon |\tilde{f}_w(X'\beta) - \tilde{f}(X\beta)| \leq \varepsilon\tilde{f}(X\beta)$ with probability at least $1-\delta$. By Corollary 4.11 and Lemma 4.12 this implies with high probability that (X', w) is a 7ε-coreset for f. Folding the constant into ε completes the proof. □

In the case $p = 2$, which is of special importance since it corresponds to the standard probit regression model, we have the following improvements:

Corollary 4.18. *Consider the setting of Theorem 4, for $p = 2$. The running time can be reduced to $O(\mathrm{nnz}(X) \log n + \mathrm{poly}(d))$. Moreover there exists a single pass online algorithm that runs in time $O(nd^2 + \mathrm{poly}(d))$ and computes a coreset of size*

$$O\left(\frac{\mu d^2 \ln(\|X\|_2)}{\varepsilon^2} \ln(\varepsilon^{-1}\mu) \ln(\mu d \ln(\|X\|_2))\right),$$

where $\|X\|_2$ denotes the largest singular value of X.

Proof of Corollary 4.18. If $p = 2$ and $d = \omega(\ln n)$, we can use a Johnson–Lindenstrauss transform, i.e., a matrix $G \in \mathbb{R}^{d \times m}$ where $m = O(\ln(n))$ and whose entries are i.i.d. $G_{ij} \sim N(0, \frac{1}{m})$ (Johnson and Lindenstrauss, 1984) to compute a $\frac{1}{2}$-approximation to the row norms: We have $\|v_i'\|_2^2 := \|x_i(R^{-1}G)\|_2^2 \geq \|x_i R^{-1}\|_2^2/2$ for all $i \in [n]$ simultaneously with constant probability. The running time reduces to $O(\mathrm{nnz}(X) \ln(n) + \mathrm{poly}(d))$. The online algorithm is obtained by running the online ℓ_2 leverage score algorithm of Chhaya et al. (2020) that recently extended the previous work of Cohen et al. (2020). Each row update takes $O(d^2)$ time except for at most $O(d)$ updates that take $O(d^3)$ time, implying $O(nd^2 + \mathrm{poly}(d))$ total running time. The slightly increased coreset size results from an increase of the total sensitivity by at most $\log(\|X\|_2)$ due to the online procedure (Chhaya et al., 2020). □

The coreset of Theorem 4 or Corollary 4.18 can then be used to compute a $(1 + \varepsilon)$-approximation for the optimal maximum likelihood estimator for β by Corollary 2.16.

5 Reducing the width of two layer ReLU networks

In this section we focus on reducing the width of two layer ReLU networks that is necessary to get an arbitrarily small training error and thus also an arbitrarily small number of missclassifications. We first state the setting, the initialization consisting of coupled random Gaussian vectors at the first layer with alternating labels at the second layer, which allows us to reduce the width of the network and its motivation. Then we state our main assumption, that the separation margin γ is bounded. We follow up with some intuition for the separation margin as well as some examples where it can be computed explicitly. Next we prove some lower bounds. Then we go into the analysis for the upper bound. Last we finish showing that our analysis is in some sense tight.

5.1 Setting and notations

We consider a set of data points $x_1, \dots, x_n \in \mathbb{R}^d$ with $\|x_i\|_2 = 1$ and labels $y_1, \dots, y_n \in \{-1, 1\}$. The two layer network is parameterized by $m \in \mathbb{N}, a \in \mathbb{R}^m$ and $W \in \mathbb{R}^{m \times d}$ as follows: we set the output function

$$f(x, W, a) = \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \phi\left(\langle w_s, x \rangle\right),$$

closely comparable to Ji and Telgarsky (2020). In the output function, $\phi(v) = \max\{0, v\}$ denotes the ReLU function for $v \in \mathbb{R}$. To simplify notation we set $f_i(W) = f(x_i, W, a)$. Further we set $\ell(v) = \ln(1 + \exp(-v))$ to be the logistic loss function. We use a random initialization W_0, a_0 given in Definition 5.1. Our goal is to minimize the empirical risk of W given by

$$R(W) = \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i f_i(W)\right).$$

To accomplish this, we use a standard gradient descent algorithm. More precisely for $t \geq 0$ we set

$$W_{t+1} = W_t - \eta \nabla R(W_t)$$

for some step size η. Further, it holds that

$$\nabla R(W) = \frac{1}{n} \sum_{i=1}^{n} y_i \nabla f_i(W) \ell'\left(y_i f_i(W)\right).$$

Moreover, we use the following notation

$$f_i^{(t)}(W) := \langle \nabla f_i(W_t), W \rangle$$

and

$$R^{(t)}(W) := \sum_{i=1}^{n} \ell\left(y_i f_i^{(t)}(W)\right).$$

Note that $\frac{\partial f_i(W)}{\partial w_s} = \frac{1}{\sqrt{m}} a_s \mathbf{1}[\langle w_s, x_i\rangle > 0] x_i$. In particular the gradient is independent of $\|w_s\|_2$, which will be crucial in our improved analysis.

5.2 The initialization and its motivation

Definition 5.1 (Coupled Initialization)**.** *We initialize the network weights as follows:*

- *For each $r = 2i - 1$, we choose w_r to be a random Gaussian vector drawn from $\mathcal{N}(0, I)$.*
- *For each $r = 2i - 1$, we sample a_r from $\{-1, +1\}$ uniformly at random.*
- *For each $r = 2i$, we choose $w_r = w_{r-1}$.*
- *For each $r = 2i$, we choose $a_r = -a_{r-1}$.*

We note this coupled initialization appeared before in Daniely (2020) for analyzing well-spread random inputs on the sphere. The initialization is chosen in such a way as to ensure that for each of the n input points, the initial value of the network is always 0. This is crucial for our analysis, and is precisely what allows us to use arbitrarily large weight vectors. Indeed, a large number of normalized weight vectors were there precisely to ensure that the initial value of the network is small in previous works. One might worry that our initialization causes the weights to be *dependent*. Indeed, each weight vector occurs exactly twice in the hidden layer. We are able to show that this dependence does not cause problems for our analysis. In particular, the separation margin in the NTK-induced feature space required for convergence in previous work can be shown to still hold, since such analyses are loose enough to accommodate such dependencies. Now, we have a similar initialization as in previous work, but since we no longer have normalized weight vectors, we can show that we can change the learning rate of gradient descent from that in previous work and it no longer needs to be balanced with the initial value, since the latter is 0. This ultimately allows for us to use a smaller width (i.e., value of m) in our analyses. For $r \in [m]$, we have

$$\frac{\partial f(W, x, a)}{\partial w_r} = a_r x \mathbf{1}_{w_r^\top x \geq 0}. \tag{37}$$

In particular scaling w_r large enough will guarantee that $\mathbf{1}_{w_r^\top x \geq 0}$ and thus $\frac{\partial f(W,x,a)}{\partial w_r}$ does not change during the gradient descent which is one of the main arguments in the analysis. The other important point about the initialization is that if the width is large enough then there is a direction improving all predictions by the separation margin γ. This is not necessarily exactly the same direction as the gradient but its existence ensures that the gradient will continue improve the loss of the network. The fact that $\frac{\partial f(W,x,a)}{\partial w_r}$ does not

change will guarantee that this directions always exists even after the weight vectors changed by applying the gradient descent.

5.3 Outline of the analysis

The analysis part of this chapter is built up as follows:

1) We first describe our main assumption that the separation margin is bounded and prove some properties regarding the separation margin.

2) We then give some examples of instances where we can bound the separation margin.

3) We then prove our lower bounds on the width of two layer ReLU network needed to converge to an arbitrarily good solution using one of the examples we give.

4) Next we prove the upper bound for the width.

5) Last we prove a tightness result for our analysis and show how a different analysis improves the upper bound in the two dimensional case.

The first two points here can be seen as a motivation and explanation of the setting we are working in and with and the remaining points are the theoretical analysis of our main results.

5.4 Main assumption and examples

5.4.1 Main assumption

Here, we define the parameter $\gamma > 0$ which was also used in Ji and Telgarsky (2020). Intuitively, γ determines the separation margin of the NTK. Let $B = B^d = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq 1\}$ be the unit ball in d dimensions. We set $\mathcal{F}_B$ to be the set of functions f mapping from $\mathrm{dom}(f) = \mathbb{R}^d$ to $\mathrm{range}(f) = B$. Let $\mu_{\mathcal{N}}$ denote the Gaussian measure on $\mathbb{R}^d$, specified by the Gaussian density with respect to the Lebesgue measure on $\mathbb{R}^d$.

Definition 5.2. *Given a data set $(X, Y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ and a map $\bar{v} \in \mathcal{F}_B$ we set*

$$\gamma_{\bar{v}} = \gamma_{\bar{v}}(X, Y) := \min_{i\in[n]} y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0] \; d\mu_{\mathcal{N}}(z).$$

We say that $\bar{v}$ is optimal if $\gamma_{\bar{v}} = \gamma(X, Y) := \max_{\bar{v}'\in\mathcal{F}_B} \gamma_{\bar{v}'}$.

We note that $\max_{\bar{v}'\in\mathcal{F}_B} \gamma_{\bar{v}'}$ always exists since $\mathcal{F}_B$ is a set of bounded functions on a compact subset of $\mathbb{R}^d$. We make the following assumption, which is also used in Ji and Telgarsky (2020):

Assumption 5.1. *It holds that $\gamma = \gamma(X, Y) > 0$.*

Before we prove our main results we show some properties of $\bar{v}$ to develop a better understanding of our assumption. The following lemma shows that the integral can be viewed as a finite sum over certain cones in $\mathbb{R}^d$. Given $U \subseteq \{1, 2, \dots, n\} = [n]$ we define the cone

$$C(U) := \{x \in \mathbb{R}^d \mid \langle x, x_i \rangle > 0 \text{ if and only if } i \in U\}.$$

Note that $C(\emptyset) = \{x \in \mathbb{R}^d \mid \langle x, x_i \rangle \leq 0 \text{ for all } i \in [n]\}$ and that $\mathbb{R}^d = \dot{\bigcup}_{U \subseteq [n]} C(U)$ as for any point $x \in \mathbb{R}^d$ there is exactly one set $U = U_x = \{i \in [n] \mid \langle x, x_i \rangle > 0\}$ such that $x \in C(U)$. Further we set $P(U)$ to be the probability that a random Gaussian is an element of $C(U)$ and P_U to be the probability measure of random Gaussians $z \sim \mathcal{N}(0, I)$ restricted to the event that $z \in C(U)$. The following lemma shows that we do not have to consider each mapping in $\mathcal{F}_B$ but it suffices to focus on a specific subset. More precisely we can assume that $\bar{v}$ is constant on the cones $C(U)$. In particular this means we can assume $\bar{v}(z) = \bar{v}(cz)$ for any $z \in \mathbb{R}^d$ and scalar $c > 0$ and that $\bar{v}$ is locally constant.

Lemma 5.3. *Let $\bar{v} \in \mathcal{F}_B$. Then there exists $\bar{v}'$ such that $\gamma_{\bar{v}'} = \gamma_{\bar{v}}$ and $\bar{v}'$ is constant on $C(U)$ for any $U \subseteq [n]$.*

Proof. Observe that for any distinct $U, U' \subseteq [n]$ the cones $C(U)$ and $C(U')$ are disjoint since for any $x \in \mathbb{R}^d$ the cone $C(U_x)$ containing x is given by $U_x = \{i \in [n] \mid \langle x, x_i \rangle > 0\}$. Further we have that $\bigcup_{U \subseteq [n]} C(U) = \mathbb{R}^d$ since any $x \in \mathbb{R}^d$ is included in some $C(U_x)$. Thus for any $i \in [n]$ we have

$$\begin{aligned} y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) &= y_i \sum_{U \subseteq [n]} P(U) \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; dP_U(z) \\ &= y_i \sum_{U \subseteq [n], i \in U} P(U) \int \langle \bar{v}(z), x_i \rangle \; dP_U(z) \\ &= y_i \sum_{U \subseteq [n], i \in U} P(U) \; \langle x_i, \int \bar{v}(z) \; dP_U(z) \rangle. \end{aligned}$$

Hence defining $\bar{v}'(x) = P(U_x) \int \bar{v}(z) \; dP_{U_x}(z)$ satisfies

$$y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) = y_i \int \langle \bar{v}'(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \;\; \mu_{\mathcal{N}}(z)$$

and since $\|\bar{v}(z)\|_2 \leq 1$ it follows that $\|\bar{v}'(z)\|_2 \leq 1$ for all $z \in \mathbb{R}^d$. □

Next we give an idea how the dimension d can impact γ. We show that in the simple case, where $\mathbb{R}^d$ can be divided into orthogonal subspaces, such that each data point x_i is an element of one of the subspaces, there is a helpful connection between a mapping $\bar{v} \in \mathcal{F}_B$ and the mapping that $\bar{v}$ induces on the subspaces.

Lemma 5.4. *Assume there exist orthogonal subspaces $V_1, \dots V_s$ of $\mathbb{R}^d$ with $\mathbb{R}^d = \bigoplus_{j \leq s} V_j$ such that for each $i \in [n]$ there exists $j \in [s]$ such that $x_i \in V_j$. Then the following two statements hold:*

Part 1. *Assume that for each $j \in [s]$ there exists $\gamma_j > 0$ and $\bar{v}_j \in \mathcal{F}_B$ such that for all $x_i \in V_j$ we have*

$$y_i \int \langle \bar{v}_j(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) \geq \gamma_j.$$

Then for each $\rho \in \mathbb{R}^s$ with $\|\rho\|_2 = 1$ there exists $\bar{v} \in \mathcal{F}_B$ with

$$\min_{i \in [n]} y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) \geq \min_{j \in [s]} \rho_j \gamma_j.$$

Part 2. *Assume that $\bar{v}$ maximizes the term*

$$\gamma^* = \min_{i \in [n]} y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z),$$

and that $\gamma^* > 0$. Given any vector $z \in \mathbb{R}^d$ we denote by $p_j(z) \in V_j$ the projection of z onto V_j. Let $\rho'_j = \max_{z \in \mathbb{R}^d} \|p_j(\bar{v}(z))\|_2$. Then for all $j \in [s]$ the mapping $\bar{v}_j(z) = \frac{p_j(\bar{v}(z))}{\rho'_j}$ maximizes

$$\gamma_j = \min_{x_i \in V_j} y_i \int \langle \bar{v}_j(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z)$$

and it holds that $\|\bar{v}_j(z)\|_2 \leq 1$ for all $z \in \mathbb{R}^d$. In other words if $\bar{v}$ is optimal for (X, Y) then $\bar{v}_j$ is optimal for (X_j, Y_j) where $X_j = \{x_i \in V_j \mid i \in [n]\}$ with the corresponding labels, i.e., $y_{x_i} = y_i$.

Proof. **Part 1.**

Since applying the projection p_j onto V_j to any point $z \in \mathbb{R}^d$ does not change the scalar product of z and $x_i \in V_j$, i.e., $\langle x_i, z \rangle = \langle x_i, p_j(z) \rangle$, we can assume that for all $z \in \mathbb{R}^d$ we have $\bar{v}_j(z) \in V_j$. Let $z \in \mathbb{R}^d$. We define $\bar{v}(z) := \sum_{j=1}^s \rho_j \bar{v}_j(z)$. Then by orthogonality

$$\|\bar{v}(z)\|_2^2 = \sum_{j=1}^s \rho_j^2 \|\bar{v}_j(z)\|_2^2 \leq \sum_{j=1}^s \rho_j^2 \cdot 1 = 1.$$

Thus it holds that $\bar{v} \in \mathcal{F}_B$. Further we have $\langle x_i, \bar{v}(z) \rangle = \sum_{k=1}^s \rho_k \langle x_i, \bar{v}_k(z) \rangle = \rho_j \langle x_i, \bar{v}_j(z) \rangle$ for $x_i \in V_j$ again by orthogonality it holds that

$$y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) = \rho_j y_i \int \langle \bar{v}_j(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) \geq \; \rho_j \gamma_j.$$

Part 2.

For the sake of contradiction assume that there are $k \leq s$ and $\bar{v}_k^* \in \mathcal{F}_B$ such that

$$\gamma_k^* = \min_{x_i \in V_k} y_i \int \langle \bar{v}_k(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \; d\mu_{\mathcal{N}}(z) = \gamma_k + \varepsilon$$

for some $\varepsilon > 0$. Using **Part 1.** we can construct a new mapping $\bar{v}' \in \mathcal{F}_B$ by using the mappings $\bar{v}_j$ defined in

the lemma for $j \neq k$, and exchange $\bar{v}_k$ by $\bar{v}_k^*$. Also as in **Part 1** let $\rho_j = \rho'_j + \varepsilon'$ for $j \neq k$ and $\rho_k = \rho'_k - 2\frac{s\varepsilon'}{\rho'_k}$ with $\varepsilon' = \min\{\frac{\rho'^2_k}{4s}, \frac{\rho'^2_k \varepsilon}{4(\gamma_k+\varepsilon)s}\}$. Then we have

$$2s + s\varepsilon' + 4s^2 \frac{\varepsilon'}{\rho'^2_k} \leq 4s.$$

Subtracting $4s$ and multiplying with ε' gives us

$$2s\varepsilon' + s\varepsilon'^2 - 4s\varepsilon' + 4\left(\frac{s\varepsilon'}{\rho'_k}\right)^2 \leq 0.$$

Hence it holds that

$$\begin{aligned}\sum_{j=1}^{s} \rho_j^2 \leq & \left(\sum_{j\neq k}(\rho'^2_j + 2\varepsilon' + \varepsilon'^2)\right) + \rho'^2_k - 4s\varepsilon' + 4\left(\frac{s\varepsilon'}{\rho'_k}\right)^2 \\ \leq & \left(\sum_{j=1}^{s} \rho'^2_j\right) + 2s\varepsilon' + s\varepsilon'^2 - 4s\varepsilon' + 4\left(\frac{s\varepsilon'}{\rho'_k}\right)^2 \leq \sum_{j=1}^{s} \rho'^2_j \leq 1.\end{aligned}$$

For any $x_i \in V_j$ with $j \neq k$ we have by orthogonality as in **Part 1.**

$$\begin{aligned}y_i \int \langle \bar{v}'(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) &= \rho_j y_i \int \langle \bar{v}_j(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) \\ &= (\rho'_j + \varepsilon')y_i \int \langle \bar{v}_j(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z).\end{aligned}$$

Further we have

$$\begin{aligned}\min_{x_i \in V_k} y_i \int \langle \bar{v}'(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) &= \rho_k \gamma_k^* \\ &= (\rho'_k - 2\frac{s}{\rho_k}\varepsilon')(\gamma_k + \varepsilon) \\ &\geq \rho'_k\gamma_k - \frac{2s}{\rho'_k} \cdot \frac{\rho'^2_k \varepsilon}{4(\gamma_k+\varepsilon)s}(\gamma_k + \varepsilon) + \rho'_k\varepsilon \\ &= \rho'_k\gamma_k + \frac{\rho'_k\varepsilon}{2}.\end{aligned}$$

We conclude again by orthogonality that

$$\begin{aligned}y_i \int \langle \bar{v}'(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) &= \rho_j y_i \int \langle \bar{v}'_j(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) \\ &> \min_j \rho'_j\gamma_j \\ &= \gamma^*\end{aligned}$$

and thus $\bar{v}'$ contradicts the maximizing choice of $\bar{v}$. □

As a direct consequence we get that the problem of finding an optimal $\bar{v}$ for the whole data set can be reduced to finding an optimal $\bar{v}_j$ for each subspace.

Corollary 5.5. *Assume there exist orthogonal subspaces $V_1, \dots V_s$ of $\mathbb{R}^d$ with $\mathbb{R}^d = \bigoplus_{j \leq s} V_j$ such that for each $i \in [n]$ there exists $j \in [s]$ with $x_i \in V_j$. For $j \in [s]$ let (X_j, Y_j) denote the data set consisting of all data points (x_i, y_i) where $x_i \in V_j$. Then $\bar{v}$ is optimal for (X, Y) if and only if for all $j \in [s]$ the mapping $\bar{v}_j$ defined in Lemma 5.4 is optimal for (X_j, Y_j) and $\gamma_{\bar{v}} = \sum_{j \in [s]} \gamma_j \rho_j^*$ where $\rho^* = \operatorname{argmax}_{\rho \in \mathcal{S}^{s-1}} \min_{j \in [s]} \rho_j \gamma_j$.*

Proof. One direction follows immediately by Lemma 5.4 2) the other direction is a direct consequence of the formula given in Lemma 5.4 1). □

The following bound for γ simplifies calculations in some cases of interest. It also gives us a natural candidate for an optimal $\bar{v} \in \mathcal{F}_B$. Given an instance (X, Y) recall that $U_z = \{i \in [n] \mid \langle z, x_i \rangle > 0\}$. We set

$$\bar{v}_0(z) = \frac{\sum_{i \in [n] \cap U_z} x_i y_i}{\| \sum_{i \in [n] \cap U_z} x_i y_i \|_2}. \tag{38}$$

We note that $\bar{v}_0(z)$ is not optimal in general but if instances have certain symmetry properties, then $\bar{v}_0(z)$ is optimal.

Lemma 5.6. *For any subset $S \subseteq [n]$ it holds that*

$$\gamma \leq \sum_{U \subseteq [n]} P(U) \frac{1}{|S|} \left\| \sum_{i \in S \cap U} x_i y_i \right\|_2$$

Proof. By Lemma 5.3 there exists an optimal $\bar{v}$ that is constant on $C(U)$ for all $U \subseteq [n]$. For $x \in U$ let $z_U = \bar{v}(x)$. Then by using an averaging argument and the Cauchy–Schwarz inequality we get

$$\begin{aligned}
\gamma &\leq \frac{1}{|S|} \sum_{i \in S} y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] \, d\mu_{\mathcal{N}}(z) \\
&= \frac{1}{|S|} \sum_{i \in S} y_i \sum_{U \subseteq [n], i \in U} P(U) \langle x_i, z_U \rangle \\
&= \frac{1}{|S|} \sum_{U \subseteq [n]} P(U) \langle \sum_{i \in S \cap U} y_i x_i, z_U \rangle \\
&\leq \sum_{U \subseteq [n]} P(U) \frac{1}{|S|} \left\| \sum_{i \in S \cap U} x_i y_i \right\|_2 .
\end{aligned}$$

□

Finally we give an idea of how two points and their distance impacts the cones and their hitting probabilities.

Lemma 5.7. *Let $x_1, x_2 \in \mathcal{S}^{d-1}$ be two points with $\langle x_1, x_2\rangle > 0$ and $\|x_1 - x_2\|_2 = b > 0$. Set $V_1' = \{x \in \mathbb{R}^d \mid \langle x_1, x\rangle > 0 \geq \langle x_2, x\rangle\}$. Then for a random Gaussian z we have $z \in V_1'$ with probability $P(V_1')$ where $\frac{b}{7} \leq P(V_1') \leq \frac{b}{5}$. Further for any z with $\|z\|_2 = 1$ it holds that $|\langle x_1, z\rangle - \langle x_2, z\rangle| \leq b$.*

Proof. We define $V_1 = \{x \in \mathbb{R} \mid \langle x_1, x\rangle > 0\}$. Then $P(V_1) = \frac{1}{2}$ since for a random Gaussian z it holds that $\langle x_1, z\rangle > 0$ with probability $\frac{1}{2}$. Since the space spanned by x_1 and x_2 is 2-dimensional, we can assume that x_1 and x_2 are on the unit circle and that $x_1 = (1, 0)$ and $x_2 = (\cos(\varphi), \sin(\varphi))$ for $\varphi \leq \frac{\pi}{2}$. Note that $P(V_1')$ is given by $\frac{b'}{2\pi}$ where $b' = \varphi$ is the length of the arc connecting x_1 and x_2 on the circle. Since b is the Euclidean distance and thus the shortest distance between x_1 and x_2 we have $b \leq b'$. Further it holds that

$$h(\varphi) := \frac{b'}{b} = \frac{\varphi}{\sqrt{(1-\cos(\varphi))^2 + \sin(\varphi)^2}} = \frac{\varphi}{\sqrt{2 - 2\cos(\varphi)}}.$$

Then $h'(\varphi)$ is positive on $(0, \frac{\pi}{2}]$, so $h(\varphi)$ is monotonously non-decreasing, and thus $h(\varphi) \leq h(\frac{\pi}{2}) = \frac{(\pi/2)}{\sqrt{2}} = \frac{\pi}{\sqrt{8}}$ and $b' \leq b \cdot \frac{\pi}{\sqrt{8}}$. Consequently for $P(V_1') = \frac{b'}{2\pi}$ we have that

$$\frac{b}{7} \leq \frac{b}{2\pi} \leq P(V_1') \leq \frac{b}{2\pi} \cdot \frac{\pi}{\sqrt{8}} \leq \frac{b}{5}.$$

For the second part we note that for any z with $\|z\|_2 = 1$ we get

$$|\langle z, x_1\rangle - \langle z, x_2\rangle| = |\langle z, x_1 - x_2\rangle| \leq \|z\|_2 \|x_1 - x_2\|_2 = 1 \cdot b$$

by using the Cauchy–Schwarz inequality. □

5.4.2 Example 1: orthonormal unit vectors

Let us start with a simple example first: let $e_i \in \mathbb{R}^d$ be the i-th unit vector. Let $n = 2d$, $x_i = e_i$ for $i \leq d$ and $x_i = -e_{i-d}$ otherwise with arbitrary labels. First consider the instance (X_i, Y_i) created by the points x_i and x_{i+d} for $i \leq d$. Then we note that $\bar{v}_i$ sending any point z with $\langle z, e_i\rangle > 0$ to $e_i y_i$ and any other point to $-e_i y_{i+d}$ is optimal since it holds that $\gamma_i = \gamma_{\bar{v}_i}(X_i, Y_i) = \int 1 \cdot \mathbf{1}[\langle x_i, z\rangle > 0]\, d\mu_{\mathcal{N}}(z) = \frac{1}{2}$. Since the subspaces $V_i = \operatorname{span}\{e_i\}$ are orthogonal we can apply Corollary 5.5 with vector $\rho = (\frac{1}{\sqrt{d}})^d$. Thus the optimal γ for our instance is $\frac{1}{2\sqrt{d}}$.

5.4.3 Example 2: Two differently labeled points at distance b

The next example is a set of two points $x_1, x_2 \in \mathbb{R}^d$ with $y_1 = 1 = -y_2$ and $\langle x_1, x_2\rangle > 0$. Let $U_1 = \{1\}, U_2 = \{2\}, U = \{1, 2\}$ and $V_1 = \{x \in \mathbb{R} \mid \langle x_1, x\rangle > 0\}$. Then $P(U) = P(V_1) - P(U_1) \geq \frac{1}{2} - \frac{b}{5}$ by Lemma 5.7 and $P(U_1) = P(U_2) = P(V_1) - P(U) \leq \frac{1}{2} - (\frac{1}{2} - \frac{b}{5}) = \frac{b}{5}$. For an illustration see Figure 3.

By Lemma 5.3 we can assume that there exists an optimal $\bar{v}$ which is constant on $C(U)$ and constant on $C(U_i)$ for $i \in \{1, 2\}$, i.e., that $\bar{v}(z) = z' \in B$ for all $z \in C(U)$ and $\bar{v}(z) = z'' \in B$ for all $z \in C(U_1)$.

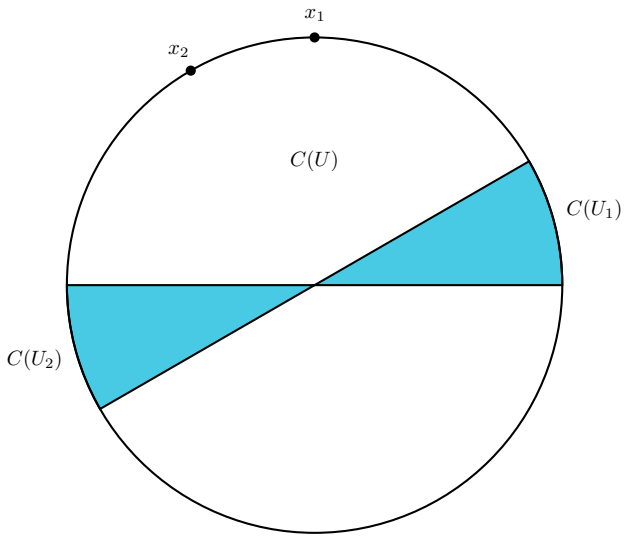

Figure 3: a) Two points x_1 and x_2 on the sphere. $C(U)$ is the cone consisting of vectors having positive scalar product with both points. The cone $C(U_i)$ consists of vectors having positive scalar product with x_i but negative scalar product with the other point. b) The probability $P(U_i)$ of a random Gaussian being in the cone $C(U_i)$ is exactly the length of the shortest arc on the circle (which is close to the Euclidean distance) connecting the points, divided by 2π.

By Lemma 5.7 we have $|\langle x_1, z'\rangle - \langle x_2, z'\rangle| \leq b$. Consequently since x_1 and x_2 have different labels there exists at least one $i \in \{1, 2\}$ with $\langle z', x_i\rangle y_i \leq b/2$ since $\langle z', x_1\rangle \geq b/2$ implies $-\langle z', x_2\rangle \leq -\langle z', x_1\rangle + |\langle x_1, z'\rangle - \langle x_2, z'\rangle| \leq -b/2 + b = b/2$. Then by Lemma 5.3 we have

$$\begin{aligned} y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) &\leq P(U) \cdot \langle z', x_i\rangle + P(U_i) \cdot \langle z'', x_i\rangle \\ &\leq \frac{1}{2} \cdot \frac{b}{2} + \frac{b}{5} \cdot 1 \\ &\leq \frac{b}{2}. \end{aligned}$$

5.4.4 Example 3: Constant labels

Let X be any data set and let Y be the all 1s vector. Then for $\bar{v}(z) = \frac{z}{\|z\|_2}$ it holds that

$$y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) = y_i \int \left\langle \frac{z}{\|z\|_2}, x_i \right\rangle \mathbf{1}[\langle x_i, z\rangle > 0]\ d\mu_{\mathcal{N}}(z) \overset{*}{=} \Omega\left(\frac{1}{\sqrt{d}}\right).$$

Thus we have $\gamma(X, Y) = \Omega\left(\frac{1}{\sqrt{d}}\right)$. We note that $*$ is a well-known fact, see Blum et al. (2020). Since we consider only a fixed x_i, we can assume that $y_i x_i$ equals the first standard basis vector e_1. We are interested in the expected projection of a uniformly random unit vector $\frac{z}{\|z\|_2}$ in the same halfspace as e_1.

We give a short proof for completeness: note that $\frac{z}{\|z\|_2} = (z_1, \ldots, z_d)/\sqrt{\sum_{i=1}^d z_i^2}$ with $z_i \sim \mathcal{N}(0,1)$, is a uniformly random unit vector u. By Jensen's inequality we have $\mathbb{E}[\sqrt{\sum_{i=1}^d z_i^2}] \leq \sqrt{\mathbb{E}[\sum_{i=1}^d z_i^2]} = \sqrt{\sum_{i=1}^d \mathbb{E}[z_i^2]} = \sqrt{d}$. Thus, with probability at least 3/4 it holds that $\sqrt{\sum_{i=1}^d g_i^2} \leq 4\sqrt{d}$, by a Markov bound. Also, $|z_i| \geq \sqrt{2} \cdot \mathrm{erf}^{-1}(1/2)$ holds with probability at least 1/2, since the right hand side is the median of the half-normal distribution, i.e., the distribution of $|z_i|$, where $z_i \sim \mathcal{N}(0,1)$. Here erf denotes the the Gauss error function.

By a union bound over the two events it follows with probability at least $1 - \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$ that

$$|u_i| = |z_i| / \sqrt{\sum_{i=1}^d z_i^2} \geq \sqrt{2} \cdot \mathrm{erf}^{-1}(1/2)/(4\sqrt{d}).$$

Consequently $\mathbb{E}[|u_i|] \geq \frac{1}{4} \cdot \sqrt{2} \cdot \mathrm{erf}^{-1}(1/2)/(4\sqrt{d}) = \Omega(1/\sqrt{d})$ and thus

$$y_i \int \left\langle \frac{z}{\|z\|_2}, x_i \right\rangle \mathbf{1}[\langle x_i, z\rangle > 0] \; d\mu_{\mathcal{N}}(z) = \frac{1}{2}\mathbb{E}[|u_i|] = \Omega(1/\sqrt{d}).$$

5.4.5 Example 4: The hypercube

In the following example we use x_i for the i-th coordinate of $x \in \mathbb{R}^d$ rather than for the i-th data point. We consider the hypercube $X = \{-\frac{1}{\sqrt{d}}, +\frac{1}{\sqrt{d}}\}^d$ with different labelings. Given $x \in X$ we set $S_x = \{i \in [d] \mid x_i = -\frac{1}{\sqrt{d}}\}$ and $\sigma(x) = |S_i|$.

Majority labels First we consider the data set $X' = X \setminus \{x \in X \mid \sigma(x) = \frac{d}{2}\}$ and assign $y_x = -1$ if $\sigma(x) > \frac{d}{2}$ and $y_x = -1$ if $\sigma(x) < \frac{d}{2}$. Note that $d - 2\sigma(x) < 0$ holds if and only if $y_x = -1$. Let $x_c \in X$ be the constant vector that has all coordinates equal to $1/\sqrt{d}$. Now, if we fix $\bar{v}(z) = x_c$ for any z, then for all $x \in X'$ we have that

$$y_x \int \langle \bar{v}(z), x\rangle \mathbf{1}[\langle x, z\rangle > 0] \; d\mu_{\mathcal{N}}(z) = \frac{y_x}{2} \cdot \frac{d - 2\sigma(x)}{d} \geq \frac{1}{2} \cdot \frac{1}{d}.$$

Hence it follows that $\gamma(X', Y) \geq \frac{1}{2d}$

Parity labels Second we consider the case where $y_x = (-1)^{\sigma(x)}$. Then we get the following bounds for γ:

Lemma 5.8. *Consider the hypercube with parity labels.*

1) If d is odd, then $\gamma = 0$.

2) If d is even, then $\gamma > 0$.

Proof. 1): First note that the set $Z = \{z \in \mathbb{R}^d \mid \exists x \in X \text{ with } \langle x, z\rangle = 0\}$ is a null set with respect to the Gaussian measure $\mu_{\mathcal{N}}$. Fix any coordinate $i \leq d$. W.l.o.g. let $i \neq 1$. Given $x \in M := \{\frac{1}{\sqrt{d}}\} \times \{-\frac{1}{\sqrt{d}}, \frac{1}{\sqrt{d}}\}^{d-2}$ consider the set $S(x) = \{(\frac{1}{\sqrt{d}}, x), (-\frac{1}{\sqrt{d}}, x), (\frac{1}{\sqrt{d}}, -x), (-\frac{1}{\sqrt{d}}, -x)\}$. Note that X is the disjoint union $X =$

$\dot{\bigcup}_{x\in M} S(x)$. Further since $d-1$ is even, it holds that $y_{(\frac{1}{\sqrt{d}},x)} = y_{(\frac{1}{\sqrt{d}},-x)} = -y_{(-\frac{1}{\sqrt{d}},-x)} = -y_{(-\frac{1}{\sqrt{d}},x)}$. Let $z \in Z$ and let $U_z = \{x' \in X \mid \langle z, x'\rangle > 0\}$. W.l.o.g. let $\langle z, (\frac{1}{\sqrt{d}}, x)\rangle > 0$. Then we have $\langle z, x'\rangle > 0$ for exactly one $x' \in \{(-\frac{1}{\sqrt{d}}, x), (\frac{1}{\sqrt{d}}, -x)\}$ and $\langle z, (-\frac{1}{\sqrt{d}}, -x)\rangle < 0$. Now since $y_{(\frac{1}{\sqrt{d}},x)}(\frac{1}{\sqrt{d}}, x)_i = -y_{(\frac{1}{\sqrt{d}},x)}(-1, x)_i = -y_{(\frac{1}{\sqrt{d}},x)}(1, -x)_i$ we conclude that for all $x \in M$ it holds that

$$\sum_{x'\in S(x)\cap U_z} (x'y_{x'})_i = \frac{1}{\sqrt{d}} + \left(-\frac{1}{\sqrt{d}}\right) = 0$$

and thus we get

$$\sum_{x\in X\cap U_z} (xy_x)_i = \sum_{x\in M} \sum_{x'\in S(x)\cap U_z} (xy_x)_i = 0.$$

Thus by Corollary 5.6 it holds that $\gamma = 0$.

2): Consider the set M comprising the middle points of the edges, i.e., $M = \{x \in \{-\frac{1}{\sqrt{d}}, 0, \frac{1}{\sqrt{d}}\}^d \mid x_i = 0 \text{ for exactly one coordinate } i \in [d]\}$. Observe that for any $x \in X$ and $z \in M$ the dot product $d \cdot \langle x, z\rangle$ is an odd integer and thus $|\langle x, z\rangle| \geq 1/d$. Hence, for the cone $C(U_z)$ containing z we have $P(U_z) > 0$.

Now fix $z \in M$ and let $i \in [d]$ be the coordinate with $z_i = 0$. Recall $\sigma(z) = |\{k \in [d] \mid z_k = -\frac{1}{\sqrt{d}}\}|$ and set $\bar{v}(z) = e_i \cdot \sigma(z) \cdot (-1)^{d/2+1}$. Let $j \in [d]$ be any coordinate other than i and consider the pairs $\{v, w\} \subset X$ where $v \in X$ with $v_j = z_j$, $\langle v, z\rangle > 1/d$ and $w = v - 2v_j e_j$. We denote the union of all those pairs by V'. The points v and w have the same entry at coordinate i but different labels. Hence it holds that $\sum_{(v,w)\in V'} v_i y_v + w_i y_w = 0$.

Next consider the set of remaining vectors with $\langle v, z\rangle > 0$ which is given by $V = \{x \in X \mid x_j = z_j \text{ and } \langle x, z\rangle = 1/d\}$. For all $x \in V$ with $x_i = \frac{1}{\sqrt{d}}$ it holds that $\sigma(x) = \sigma(z) - (\frac{d}{2} - 1) = \sigma(z) \cdot (-1)^{d/2+1}$ since the projection of x to $\mathbb{R}^{d-1}$ that results from removing the i-th entry of x, has Hamming distance $(\frac{d}{2} - 1)$ to z projected to $\mathbb{R}^{d-1}$, and vice versa for all $x \in V$ with $x_i = -1/\sqrt{d}$ we have that $\sigma(x) = \sigma(z) \cdot (-1)^{d/2}$. Hence for $x \in V$ it holds that $y_x \bar{v}(z) = e_i \cdot \sigma(z) \cdot (-1)^{d/2+1} = e_i \cdot \operatorname{sgn}(x_i)$ and thus we have

$$\begin{aligned}
\sum_{x\in X\cap U_z} y_x \langle x, \bar{v}(z)\rangle &= \sum_{x\in V} y_x \langle x, \bar{v}(z)\rangle \;+\; \sum_{(v,w)\in V'} y_v \langle v, \bar{v}(z)\rangle + y_w \langle w, \bar{v}(z)\rangle \\
&= \sum_{x\in V} \operatorname{sgn}(x_i) \langle x, e_i\rangle \;+\; 0 \\
&= \sum_{x\in V} \frac{1}{\sqrt{d}} = 2\binom{d-1}{d/2-1}\frac{1}{\sqrt{d}}
\end{aligned}$$

since the number of elements $x \in V$ with $x_i = 1/\sqrt{d}$ is the same as the number of elements $x' \in V$ with $x'_i = -1/\sqrt{d}$. More specifically, it equals the number of points with Hamming distance $(\frac{d}{2} - 1)$ to the projection of z onto $\mathbb{R}^{d-1}$, which is $\binom{d-1}{d/2-1}$ since the i-th coordinate is fixed and we need to choose $d/2 - 1$ coordinates that differ from the remaining coordinates of z. Let $P > 0$ be the probability that a random Gaussian is in the

same cone $C(U)$ as z for some $z \in M$. Then by symmetry it holds that $\gamma_{\bar{v}} = P \cdot 2\binom{d-1}{d/2-1} \cdot \frac{1}{\sqrt{d}} \cdot \frac{1}{|X|} > 0$. $\square$

5.5 Lower bounds for log width

5.5.1 Example 5: Alternating points on a circle

Next consider the following set of n points for n divisible by 4:
$x_k = \left(\cos\left(\frac{2k\pi}{n}\right), \sin\left(\frac{2k\pi}{n}\right)\right)$ and $y_k = (-1)^k$. Intuitively, defining $\bar{v}$ to send $z \in \mathbb{R}^d$ to the closest point of our data set X multiplied by its label, gives us a natural candidate for $\bar{v}$. However, applying Lemma 5.3 gives us a better mapping that also follows from Equation (38), and which is optimal by Lemma 5.6:
Define the set $S = \{x \in \mathbb{R}^2 \mid \exists x_i \in X, \alpha \geq 0 \colon x = \alpha x_i\}$ Now, for any $z \in \mathbb{R}^d \setminus S$ there exists a unique i_z such that $z \in \text{Cone}(\{x_{i_z}, x_{i_z+1}\})$. We set $r_z = \frac{x_{i_z} y_{i_z} + x_{i_z+1} y_{i_z+1}}{\|x_{i_z} - x_{i_z+1}\|_2}$. We define the function $\bar{v} : \mathbb{R}^d \to \mathbb{R}^d$ by

$$\bar{v}(z) = \begin{cases} 0 & z \in S \\ (-1)^{n/4+1} r_z & \text{otherwise.} \end{cases}$$

Observe that for $i = i_z$ we have

$$\begin{aligned} r_z &= \left(\cos\left(\frac{2\pi}{2n} \cdot (i - \frac{n}{2} + 1)\right), \sin\left(\frac{2\pi}{2n} \cdot (i - \frac{n}{2} + 1)\right)\right) \\ &= (-1)^i \left(\sin\left(\frac{(i+1)2\pi}{2n}\right), -\cos\left(\frac{(i+1)2\pi}{2n}\right)\right). \end{aligned}$$

Figure 4 shows how $\bar{v}(z)$ is constructed for $n = 12$. We note that $\bar{v} = \bar{v}_0$ holds almost surely, which in particular implies the optimality of $\bar{v}$, cf. Equation (38). For computing γ we need the following lemma.

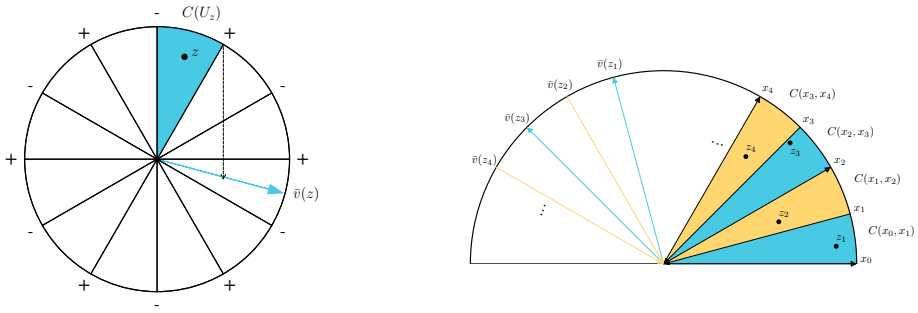

Figure 4: The left picture shows how $\bar{v}(z)$ is constructed: we subtract the vector x_3 which is labeled -1 from the vector x_2 which is labeled 1. We obtain r_z after rescaling to unit norm. Since $n/4 = 3$ is odd we have $\bar{v}(z) = r_z$. The right picture demonstrates the values of $\bar{v}(z)$ that are relevant for computing $y_i \int \langle \bar{v}(z), x_i \rangle \mathbf{1}[\langle x_i, z \rangle > 0] d\mu_{\mathcal{N}}(z)$ for the single point $x_i = (0, 1)$. Here we have $y_i \langle x_i, \bar{v}(z_j) \rangle = (-1)^{j-1} \cos\left(\frac{(2j-1)\pi}{2n}\right)$. The same argument can be repeated on the left side of the half circle.

We found the result in a post on `math.stackexchange.com` but could not find it in published literature

and so we reproduce the full proof from StEx (2011) for completeness of presentation.

Lemma 5.9 (StEx (2011))**.** *For any $a, b \in \mathbb{R}$ and $\tilde{n} \in \mathbb{N}$ it holds that*

$$\sum_{k=0}^{\tilde{n}-1} \cos(a+kb) = \frac{\cos(a+(\tilde{n}-1)b/2)\sin(\tilde{n}b/2)}{\sin(b/2)}.$$

Proof. We use $\mathbf{i}$ to denote the imaginary unit defined by the property $\mathbf{i}^2 = -1$. From Euler's identity we know that $\cos(a+kb) = \mathrm{Re}(e^{\mathbf{i}(a+kb)})$ and $\sin(a+kb) = \mathrm{Im}(e^{\mathbf{i}(a+kb)})$. Then

$$\begin{aligned}
\sum_{k=0}^{\tilde{n}-1} \cos(a+kb) &= \sum_{k=0}^{\tilde{n}-1} \mathrm{Re}\left(e^{\mathbf{i}(a+kb)}\right) \\
&= \mathrm{Re}\left(\sum_{k=0}^{\tilde{n}-1} e^{\mathbf{i}(a+kb)}\right) \\
&= \mathrm{Re}\left(e^{\mathbf{i}a}\sum_{k=0}^{\tilde{n}-1} (e^{\mathbf{i}b})^k\right) \\
&= \mathrm{Re}\left(e^{\mathbf{i}a}\frac{1-e^{\mathbf{i}b\tilde{n}}}{1-e^{\mathbf{i}b}}\right) \\
&= \mathrm{Re}\left(e^{\mathbf{i}a}\frac{e^{\mathbf{i}b\tilde{n}/2}(e^{-\mathbf{i}b\tilde{n}/2}-e^{\mathbf{i}b\tilde{n}/2})}{e^{\mathbf{i}b/2}(e^{-\mathbf{i}b/2}-e^{\mathbf{i}b/2})}\right) \\
&= \frac{\cos(a+(\tilde{n}-1)b/2)\sin(\tilde{n}b/2)}{\sin(b/2)}.
\end{aligned}$$

□

Lemma 5.10. *For all $i \in [n]$ it holds that*

$$y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0] d\mu_{\mathcal{N}}(z) = \Omega\left(\frac{1}{n}\right).$$

Proof. We set $n' = n/4$. Note that by symmetry the value of the given integral is the same for all $i \in [n]$. Thus it suffices to compute $y_i \int \langle \bar{v}(z), x_i\rangle \mathbf{1}[\langle x_i, z\rangle > 0] d\mu_{\mathcal{N}}(z) = \gamma$ for $x_i = (0,1)$, and note that $i = n/4$ for this special choice. See Figure 4 for an illustration of the following argument. For a fixed $z \in \mathbb{R}^2$ consider the cone $\mathrm{Cone}(\{x_{i_z}, x_{i_z+1}\} = \mathrm{Cone}(\{x_j, x_{j+1}\}) \subset \{x \in \mathbb{R}^2 \mid \langle x, (0,1)\rangle > 0\}$. Then $j \in [0, \frac{n}{2}-1]$ and $\langle \bar{v}(z), x_i\rangle = (-1)^{n/4+1}\langle r_z, x_i\rangle = (-1)^{n/4+1}(-1)^{j+1}\cos(\frac{(2j+1)2\pi}{2n})$ since $y_i = (-1)^{n/4}$. Further, for $j \le \frac{n}{4}-1$ it holds that

$$\langle \bar{v}(z), x_i\rangle = (-1)^{n/4}(-1)^j \cos\left(\frac{(2j+1)2\pi}{2n}\right) = y_i(-1)^j \cos\left(\frac{(2j+1)2\pi}{2n}\right),$$

and by using the symmetry of cos we get

$$(-1)^{n/4}(-1)^{(n/2)-j-1}\cos\left(\frac{(2(n/2)-2j-1)2\pi}{2n}\right) = y_i(-1)^{j+1}\left(-\cos\left(\frac{(2j+1)2\pi}{2n}\right)\right)$$
$$= y_i(-1)^j\cos\left(\frac{(2j+1)2\pi}{2n}\right).$$

Now assume w.l.o.g. that $n \geq 8$. Further we set $\tilde{n} = (n'-1)/2$ and $b = \frac{4\pi}{n} = \frac{4\pi}{4n'}$. By using Lemma 5.9 and the Taylor series expansion of $\cos(\cdot)$ and $\sin(\cdot)$ we get

$$\begin{aligned}
\gamma &= \frac{1}{n}\left(2\sum_{k=1}^{n'}\cos\left(\frac{(2k-1)\pi}{4n'}\right)(-1)^{k-1}\right) \\
&= \frac{2}{n}\left(\sum_{k=0}^{\lceil (n'-1)/2\rceil}\cos\left(\frac{(4k+1)\pi}{4n'}\right) - \sum_{k=0}^{\lfloor (n'-1)/2\rfloor}\cos\left(\frac{(4k+3)\pi}{4n'}\right)\right) \\
&\overset{*}{\geq} \frac{2}{n}\left(\frac{\cos(\pi/n+(\tilde{n}-1)b/2)\sin(\tilde{n}b/2)}{\sin(b/2)} - \frac{\cos(3\pi/n+(\tilde{n}-1)b/2)\sin(\tilde{n}b/2)}{\sin(b/2)}\right) \\
&= \frac{2}{n}\left(\frac{\overbrace{(\cos(\pi/n+(\tilde{n}-1)b/2)-\cos(3\pi/n+(\tilde{n}-1)b/2))}^{=\Theta(b)}\overbrace{\sin(\tilde{n}b/2)}^{=1-\Theta(b)}}{\sin(b/2)}\right) \\
&= \frac{2}{n}\left(\frac{\Theta(b)}{\Theta(b)}\right) = \frac{2}{n}\Theta(1) = \Omega(n^{-1}).
\end{aligned}$$

$*$ when n' is odd then we have an exact equality. □

5.5.2 Lower Bounds

Lemma 5.11. *If $m = o(n\log(n))$ then with constant probability over the random initialization of W_0 it holds for any weights $V \in \mathbb{R}^{m\times d}$ that $y_i\langle V, \nabla f_i(W_0)\rangle \leq 0$ for at least one $i \in [n]$.*

Proof. We set $x_{-i} := x_{n-i}$ for $i \geq 0$. Consider the set $\{x_i, x_{i+1}, x_{i+2}, x_{i+3}\}$ for i with $i \mod 4 = 0$. For any s let $A_{i,s}$ denote the event that

$$\mathbf{1}[\langle x_i, w_s\rangle > 0] = \mathbf{1}[\langle x_{i+1}, w_s\rangle > 0] = \mathbf{1}[\langle x_{i+2}, w_s\rangle > 0] = \mathbf{1}[\langle x_{i+3}, w_s\rangle > 0].$$

If there exists $i \in \{0, 4, \ldots, n-4\}$ such that for all $s \in [m]$ the event $A_{i,s}$ is true then at least one of the points $x_i, x_{i+1}, x_{i+2}, x_{i+3}$ is misclassified. To see this, note that there exists $\rho \in \mathbb{R}^4_{>0}$ such that $\rho_1 x_i + \rho_3 x_{i+2} - (\rho_2 x_{i+1} + \rho_4 x_{i+3}) = 0$ since the line connecting x_i and x_{i+3} crosses the line segment between

x_{i+2} and x_{i+4}. Now let $S = \{s \in [m] \mid \langle x_i, w_s\rangle > 0\}$. If the event $A_{i,s}$ is true for all $s \in [m]$ then it holds that

$$\begin{aligned}
0 &= \sum_{s\in[m],\langle x_i,w_s\rangle>0} \langle \rho_1 x_i + \rho_3 x_{i+2} - (\rho_2 x_{i+1} + \rho_4 x_{i+3}), w_s\rangle \\
&= \sum_{j=0}^{3} \sum_{s\in[m],\langle x_i,w_s\rangle>0} \rho_j y_{i+j} \langle x_{i+j}, w_s\rangle \\
&= \sum_{j=0}^{3} \rho_j \sum_{s\in[m],\langle x_{i+j},w_s\rangle>0} y_{i+j} \langle x_{i+j}, w_s\rangle
\end{aligned}$$

and since $\rho_j > 0$ it must hold $\sum_{s\in[m],\langle x_{i+j},w_s\rangle>0} y_{i+j}\langle x_{i+j}, w_s\rangle \le 0$ for at least one $j \in \{0,\dots,3\}$.

Note that $A_{i,s}$ is false with probability $2 \cdot \frac{3}{n}$, namely if $\frac{w_s}{\|w_s\|_2}$ is between the point $x_{i+n/4}$ and $x_{i+3+n/4}$ or between the points $x_{i-n/4}$ and $x_{i+3-n/4}$. We denote the union of these areas by Z_i. Further these areas are disjoint for different $i, i' \in \{0, 4, \dots n/4\}$. Now, as we have discussed above, we need at least one $A_{i,s}$ to be false for each i. This occurs only if for each i there exists at least one s such that $\frac{w_s}{\|w_s\|_2} \in Z_i$. Let T be the minimum number of trials needed to hit every one of the $n' := n/4$ regions Z_i. This is the coupon collector's problem for which it is known Erdős and Rényi (1961) that for arbitrary $c \in \mathbb{R}$ it holds that $\Pr[T < n' \log n' + cn'] = \exp(-\exp(-c))$ as $n' \to \infty$. Thus for sufficiently large n' and $c = -1$ we have

$$\Pr[T > n' \log n' - n'] > 1 - e^{-e} > 0.9.$$

□

Indeed we can show an even stronger result:

Lemma 5.12. *Let $\varepsilon \ge 0$. Any two-layer ReLU neural network with width $m < (1-\varepsilon)n/6 - 2$ misclassifies more than $\varepsilon n/3$ points of the alternating points on the circle example.*

Proof. Set $\mathcal{D} = \{x \in \mathbb{R}^2 \mid \|x\|_1 = 1\}$. Given parameters W and a consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x) = \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \varphi(\langle w_s, x\rangle)$. Note that the points $x_i' = \frac{x_i}{\|x_i\|_1} \in \mathcal{D}$ do not change their order along the ℓ_1 sphere and thus by definition of (x_i, y_i) have alternating labels. Also note that $f(x_i) > 0$ if and only if $f(x_i') > 0$. Further note that the restriction of f to $\mathcal{D}$ denoted $f_{|\mathcal{D}}$ is a piecewise linear function. More precisely the gradient $\frac{\partial f}{\partial x} = \frac{1}{\sqrt{m}} \sum_{s=1}^{m} a_s \mathbf{1}[\langle w_s, x\rangle > 0] w_s$ can only change at the points $(1,0), (0,1), (-1,0), (0,-1)$ and at points orthogonal to some w_s for $s \le m$. Since for each w_s there are exactly two points on $\mathcal{D}$ that are orthogonal to w_s this means the gradient changes at most $2m+4$ times. Now for i divisible by 3 consider the points x_i, x_{i+1}, x_{i+2}. If the gradient does not change in the interval induced by x_i and x_{i+2} then at least one of the three points is misclassified. Hence if $2m + 4 < (1-\varepsilon)\frac{n}{3}$ then *strictly* more than an $(\varepsilon/3)$-fraction of the n points is misclassified. □

5.6 Upper bound

We use the following initialization, see Definition 5.1: we set $m = 2m'$ for some natural number m'. Put $w_{s,0} = w_{s+m',0} = \beta w'_s$ where $w'_s \sim \mathcal{N}(0, I_d), \beta \in \mathbb{R}$ is an appropriate scaling factor to be defined later and $a_i = 1$ for $i < m'$ and $a_i = -1$ for $i \geq m'$. We note that to simplify notations the a_i are permuted compared to Definition 5.1, which does not make a difference. Further note that $\frac{\partial f}{\partial w_s} = \frac{\partial f}{\partial w'_s}$.

The goal of this section is to show our main theorem:

Theorem 5. *Given an error parameter $\varepsilon \in (0, 1/10)$ and any failure probability $\delta \in (0, 1/10)$, let $\rho = 2 \cdot \gamma^{-1} \cdot \ln(4/\varepsilon)$. Then if*

$$m = 2m' \geq 2\gamma^{-2} \cdot 8 \ln(2n/\delta),$$

$\beta = \frac{4 \cdot 2\rho^2 n\sqrt{m}}{5\varepsilon\delta}$ and $\eta = 1$ we have with probability at most $1 - 3\delta$ over the random initialization that $\frac{1}{T}\sum_{t=0}^{T-1} R(W_t) \leq \varepsilon$, where $T = \lceil 2\rho^2/\varepsilon \rceil$.

Before going in to the details of the proof of Theorem 5 we give a short outline: The idea of the proof is that at any time during the gradient descent there exists a good direction improving the prediction of all points. We first prove that this direction exists in the beginning and that the scalar product of the initial weight vectors and our data points is large enough so that the sign of the scalar product does not change at any time with high probability. We then state a helpful Lemma from (Ji and Telgarsky, 2020) which allows us to show the convergence to a good solution if the mentioned good direction exists. Last we prove the bounds on the number of steps needed to convergence to a good solution.

Our first lemma shows that with high probability there is a good separator at initialization, similar to Ji and Telgarsky (2020).

Lemma 5.13. *If $m' \geq \frac{8\ln(2n/\delta)}{\gamma^2}$ then there exists $U \in \mathbb{R}^{m \times d}$ with $\|u_s\|_2 \leq \frac{1}{\sqrt{m}}$ for all $s \leq m$, and $\|U\|_F \leq 1$, such that with probability at least $1 - \delta$ it holds simultaneously for all $i \leq n$ that*

$$y_i f_i^{(0)}(U) \geq \frac{\gamma}{2}$$

Proof. We define U by $u_s = \frac{a_s}{\sqrt{m}} \bar{v}(w_{s,0})$. Observe that

$$\mu_i = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)} \left[y_i \langle \bar{v}(w), x_i \rangle \right] \mathbf{1}\left[\langle x_i, w \rangle > 0 \right] \geq \gamma$$

by assumption. Further since $w_{s,0} = w_{s+m',0} = \beta w'_{s,0}$ and $a_s^2 = 1$, we have $a_s u_s = a_{s+m'} u_{s+m'}$ for $s \leq m'$. Also by Lemma 5.3 we can assume that $\bar{v}(w_{s,0}) = \bar{v}(w'_{s,0})$. Thus

$$y_i f_i^{(0)}(U) = \frac{1}{m'} \sum_{s=1}^{m'} y_i \langle \bar{v}(w_{s,0}), x_i \rangle \mathbf{1}\left[\langle x_i, w_{s,0} \rangle > 0 \right]$$

is the empirical mean of i.i.d. random variables supported on $[-1, +1]$ with mean μ_i. Therefore by Hoeffding's

inequality (Lemma 2.10), using $m' \geq \frac{8\ln(2n/\delta)}{\gamma^2}$ it holds that

$$\begin{aligned}\Pr\left[y_i f_i^{(0)}(U) \leq \frac{\gamma}{2}\right] &\leq \Pr\left[|y_i f_i^{(0)}(U) - \mu_i| \geq \frac{\mu_i}{2}\right] \\ &\leq 2\exp\left(-\frac{2\mu_i^2 m'^2/4}{m' \cdot 4}\right) \\ &\leq 2\exp\left(-\frac{\gamma^2 m'}{8}\right) \leq \frac{\delta}{n}\end{aligned}$$

Applying the union bound proves the lemma. □

Lemma 5.14. *With probability $1-\delta$ it holds that $|\langle x_i, w_{s,0}\rangle| > \frac{2\rho^2}{\varepsilon\sqrt{m}}$ for all $i \in [n]$ and $s \in [m]$*

Proof. By anti-concentration of the Gaussian distribution (Lemma 2.12), we have for any i

$$\begin{aligned}\Pr\left[|\langle x_i, w_{s,0}\rangle| \leq \frac{2\rho^2}{\varepsilon\sqrt{m}}\right] &= \Pr\left[|\langle x_i, w'_{s,0}\rangle| \leq \frac{2\rho^2}{\beta\varepsilon\sqrt{m}}\right] \\ &\leq \frac{2\rho^2}{\beta\varepsilon\sqrt{m}}\frac{4}{5} \\ &\leq \frac{\delta}{mn}.\end{aligned}$$

Thus applying the union bound proves the lemma. □

Lemma 5.15. *For all $i \in [n]$ it holds that $f_i(W_0) = 0$*

Proof. Since $a_s = -a_{s+m'}$ we have

$$f_i(W_0) = \sum_{s=1}^{m} \frac{1}{\sqrt{m}} a_s \varphi\left(\langle w_{s,0}, x_i\rangle\right) = \sum_{s=1}^{m'} \frac{1}{\sqrt{m}}(a_s + a_{s+m'})\varphi\left(\langle w_{s,0}, x_i\rangle\right) = 0.$$

□

Further we need the following lemma proved in Ji and Telgarsky (2020).

Lemma 5.16 (Lemma 2.6 in Ji and Telgarsky (2020))**.** *For any $t \geq 0$ and $\bar{W}$, if $\eta_t \leq 1$ then*

$$\eta_t R(W_t) \leq \|W_t - \bar{W}\|_F^2 - \|W_{t+1} - \bar{W}\|_F^2 + 2\eta_t R^{(t)}(\bar{W}).$$

Consequently, if we use a constant step size $\eta \leq 1$ for $0 \leq \tau < t$, then

$$\eta \sum_{\tau<t} R(W_\tau) \leq \eta \sum_{\tau<t} R(W_\tau) + \|W_t - \bar{W}\|_F^2 \leq \|W_0 - \bar{W}\|_F^2 + 2\eta \sum_{\tau<t} R^{(\tau)}(\bar{W}).$$

Now we are ready to prove the main theorem:

Proof of Theorem 5. With probability at least $1-2\delta$ there exists U as in Lemma 5.13 and also the statement of Lemma 5.14 holds. We set $\bar{W} = W_0 + \rho U$. First we show that for any $t < T$ and any $s \in [m]$ we have

$\|w_{s,t} - w_{s,0}\|_2 \leq \frac{2\rho^2}{\varepsilon\sqrt{m}}$. Observe that $|\ell'(v)| = |\frac{-e^{-v}}{1+e^{-v}}| \leq 1$ since $e^{-v} > 0$ for all $v \in \mathbb{R}$. Thus for any $t \geq 0$ we have

$$\|w_{s,t} - w_{s,0}\|_2 \leq \sum_{\tau<t} \frac{1}{n} \sum_{i=1}^{n} |\ell'(y_i f_i(W_\tau))| \left\| \frac{\partial f_i}{\partial w_{s,t}} \right\|_2 \leq \sum_{\tau<t} \frac{1}{n} \sum_{i=1}^{n} 1 \cdot \frac{1}{\sqrt{m}} \leq \frac{t}{\sqrt{m}}.$$

Consequently we have $\|w_{s,t} - w_{s,0}\|_2 \leq \frac{2\rho^2}{\varepsilon\sqrt{m}}$ for $t < T = \lceil \frac{2\rho^2}{\varepsilon} \rceil$.

Next we prove that for any $t < T$ we have $R^{(t)}(\bar{W}) < \varepsilon/4$. Since $\ln(1+r) \leq r$ for any r, the logistic loss satisfies $\ell(z) = \ln(1+\exp(-z)) \leq \exp(-z)$, and it is sufficient to prove that for any $1 \leq i \leq n$ we have

$$y_i \langle \nabla f_i(W_t), \bar{W} \rangle \geq \ln\left(\frac{\varepsilon}{4}\right).$$

Note that

$$\begin{aligned}
y_i \langle \nabla f_i(W_t), \bar{W} \rangle &= y_i \langle \nabla f_i(W_t), W_0 \rangle + y_i \rho \langle \nabla f_i(W_t), U \rangle \\
&= y_i \langle \nabla f_i(W_t), W_0 \rangle + y_i \langle \nabla f_i(W_0), W_0 \rangle - y_i \langle \nabla f_i(W_0), W_0 \rangle + y_i \rho \langle \nabla f_i(W_t), U \rangle \\
&= y_i \langle \nabla f_i(W_0), W_0 \rangle + y_i \langle \nabla f_i(W_t) - \nabla f_i(W_0), W_0 \rangle + y_i \rho \langle \nabla f_i(W_t), U \rangle.
\end{aligned}$$

For the first term we have $y_i \langle \nabla f_i(W_0), W_0 \rangle = y_i f_i(W_0) = 0$ by Lemma 5.15. For the second term we note that $|\langle x_i, w_{s,0} \rangle - \langle x_i, w_{s,t} \rangle| = |\langle x_i, w_{s,0} - w_{s,t} \rangle| \leq \|x_i\|_2 \|w_{s,0} - w_{s,t}\|_2 \leq \frac{2\rho^2}{\varepsilon\sqrt{m}}$. Thus $\mathbf{1}[\langle x_i, w_{s,0} \rangle > 0] \neq \mathbf{1}[\langle x_i, w_{s,t} \rangle > 0]$ can only hold if $|\langle x_i, w_{s,0} \rangle| \leq \frac{2\rho^2}{\varepsilon\sqrt{m}}$ which is false for all i, s by Lemma 5.14. Hence it holds that

$$\frac{\partial f_i}{\partial w_{s,t}} = \frac{1}{\sqrt{m}} a_s \mathbf{1}[\langle x_i, w_{s,t} \rangle > 0] x_i = \frac{1}{\sqrt{m}} a_s \mathbf{1}[\langle x_i, w_{s,0} \rangle > 0] x_i = \frac{\partial f_i}{\partial w_{s,0}}$$

and consequently $\nabla f_i(W_t) = \nabla f_i(W_0)$. It follows for the second term that

$$y_i \langle \nabla f_i(W_t) - \nabla f_i(W_0), W_0 \rangle = 0.$$

Moreover by Lemma 5.13 for the third term it follows

$$y_i \rho \langle \nabla f_i(W_t), U \rangle = y_i \rho \langle \nabla f_i(W_0), U \rangle \geq \rho \frac{\gamma}{2}.$$

Thus $y_i \langle \nabla f_i(W_t), \bar{W} \rangle \geq \rho \frac{\gamma}{2} \geq \ln(4/\varepsilon)$ since $\rho = 2\gamma^{-1} \cdot \ln(4/\varepsilon)$. Consequently it holds that $R^{(t)}(\bar{W}) < \varepsilon/4$.

Now using $T = \lceil \frac{2\rho^2}{\varepsilon} \rceil$ applying Lemma 5.16 with step size $\eta = 1$ gives us the desired result:

$$\begin{aligned}\frac{1}{T}\sum_{t<T} R(W_t) &\leq \frac{\|W_0 - \bar{W}\|_F^2}{T} + \frac{2}{T}\sum_{\tau<T} R^{(t)}(\bar{W}) \\ &= \frac{\|\rho U\|_F^2}{T} + \frac{2}{T}\sum_{\tau<T} R^{(t)}(\bar{W}) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &\leq \varepsilon.\end{aligned}$$

□

5.7 On the construction of U

5.7.1 Tightness of the construction of U

We note that for the construction of U used in the upper bound of Lemma 5.13 $m' \geq \frac{8\ln(2n/\delta)}{\gamma^2}$ is tight in the following sense: For $\bar{v} \in \mathcal{F}_B$, the natural estimator of γ is given by the empirical mean $\frac{1}{m}\sum_{s=1}^{m'} y_i \langle \bar{v}(w_{s,0}), x_i\rangle]\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right]$. The following lemma shows that using this estimator, the bound given in Lemma 5.13 is tight with respect to the squared dependence on γ up to a constant factor. In particular we need $m = \Omega(\gamma^{-2}\log(n))$ if we want to use the union bound over all data points.

Lemma 5.17. *Fix the choice of $u_s = \frac{a_s}{\sqrt{m}}\bar{v}(w_s)$ for $s \in [m]$. Then for each $\gamma_0 \in (0,1)$ there exists an instance (X,Y) and $\bar{v}(z) \in \mathcal{F}_B$, such that for each $i \in [n]$ it holds with probability at least $P_m = c\exp\left(-8m'\gamma^2/3\right)$ for an absolute constant $c > 0$ that*

$$y_i f_i^{(0)}(U) = \frac{1}{m'}\sum_{s=1}^{m'} y_i \langle \bar{v}(w_{s,0}), x_i\rangle]\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right] \leq 0.$$

Proof of Lemma 5.17. Consider Example 5.5.1. Recall that $\gamma(X,Y) = \Theta(1/n)$. Choose a sufficiently large n, divisible by 8, such that $\gamma(X,Y) \leq \gamma_0$. Note that the mapping $\bar{v}$ that we constructed, has a high variance since for any i, the probability that a random Gaussian z satisfies $\langle \bar{v}(z), x_i\rangle\mathbf{1}[\langle x_i, z\rangle > 0] \geq \frac{1}{\sqrt{2}}$ as well as the probability that $\langle \bar{v}(z), x_i\rangle\mathbf{1}[\langle x_i, z\rangle > 0] \leq -\frac{1}{\sqrt{2}}$ are equal to $\frac{1}{8}$. To see this, note that $|\langle \bar{v}(z), x_i\rangle| \geq \frac{1}{\sqrt{2}}$ if $\langle z, x_i\rangle < \frac{1}{\sqrt{2}}$ and in this case $\langle \bar{v}(z), x_i\rangle$ is negative with probability $\frac{1}{2}$. Thus the variance of $Z_s = y_i\langle \bar{v}(w_{s,0}), x_i\rangle]\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right]$ is at least $\frac{1}{\sqrt{2}^2}\cdot\frac{2}{8} = \frac{1}{8}$. Observe that the random variable $Z_s' = \frac{1}{2}(1 - Z_s)$ attains values in $[0,1]$. Further the expected value of Z_s' is $\frac{1}{2}(1-\gamma)$, and the variance is at least $\frac{1}{32}$. Now set $Z = \sum_{s=1}^{m'} Z_s'$ and note that $y_i f_i^{(0)}(U) = \frac{1}{m'}\sum_{s=1}^{m'} y_i\langle \bar{v}(w_{s,0}), x_i\rangle]\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right] \leq 0$ holds if and only if $Z \geq \frac{m'}{2} = \mathbb{E}(Z) + \frac{m'\gamma}{2}$. By Lemma 5.12 we know that $y_i f_i^{(0)}(U) = \frac{1}{m'}\sum_{s=1}^{m'} y_i\langle \bar{v}(w_{s,0}), x_i\rangle]\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right] \leq 0$ is true for at least one $i \in [n]$ if $m \leq \frac{n}{6} - 3$. Now choosing n large enough this implies we only need to show

the result for $m' \geq 200^2 \cdot 32$. Hence we can apply Lemma 2.13 to Z and get

$$\Pr[Z \geq \mathbb{E}(Z) + \frac{m'\gamma}{2}] \geq c \exp\left(-m'^2\gamma^2 / \left(\frac{4 \cdot 3m'}{32}\right)\right) = c \exp\left(-8m'\gamma^2/3\right)$$

for $\frac{m'\gamma}{2} \leq \frac{1}{100}\frac{m'}{32}$ or equivalently $\gamma \leq \frac{1}{1600}$ which holds if n is large enough. □

Thus we need that $m = \Omega(\frac{\ln(n/\delta)}{\gamma^2})$ for the given error probability if we construct U as in Lemma 5.13.

5.7.2 The two dimensional case (upper bound)

In the following we show how we can improve the construction of U in the special case of $d = 2$ such that

$$m = O\left(\gamma^{-1}\left(\ln(4n/\delta) + \ln(4/\varepsilon)\right)\right)$$

suffices for getting the same result as in Theorem 5. We note that the only place where we have a dependence on γ^{-2} is in Lemma 5.13. It thus suffices to replace it by the following lemma that improves the dependence to γ^{-1} in the special case of $d = 2$:

Lemma 5.18. *Let (X, Y) be an instance in $d = 2$ dimensions. Then there exists a constant $K > 1$ such that for $m \geq \frac{K\ln(n/\delta)}{\gamma}$ with probability $1 - 2\delta$ there exists $U \in \mathbb{R}^{m \times d}$ with $\|u_s\|_2 \leq \frac{1}{\sqrt{m}}$ for all $s \leq m$, and $\|U\|_F \leq 1$, such that*

$$y_i f_i^{(0)}(U) \geq \frac{\gamma}{4}$$

for all $i \leq n$.

Proof. The proof consists of three steps. The first step is to construct a net X' that consists only of 'large cones of positive volume' such that for each data point x there exists a point $x' \in X'$ whose distance from x on the circle is at most $b = \frac{\gamma}{4}$: Let $n' = \lceil 2\pi/b \rceil$ and consider the set

$$X'' = \{x \in \mathbb{R}^2 \mid x = (\cos(j/n'), \sin(j/n')), j \in \mathbb{N}\}.$$

Given $x \in X$ we define $g(x) \in \operatorname{argmin}_{x' \in X''} \|x - x'\|_2$ and $h(x) \in \operatorname{argmin}_{x' \in X'' \setminus \{g(x)\}} \|x - x'\|_2$, where ties are broken arbitrarily. We set $X' = \{g(x) \mid x \in X\} \cup \{h(x) \mid x \in X\}$. We note that the distance on the circle between two neighboring points in X' is a multiple of $\frac{2\pi}{n'}$. This implies that for any cone $C(V)$ between consecutive points in X' with $P(V) > 0$ we have $P(V) \geq 1/n' \geq b/7$ and $\|x - g(x)\|_2 \leq \frac{b}{2}$. Further note that there are at most $|X'| \leq 2n$ cones of this form.

The second step is to construct a separator $(u_s)_{s \leq m} \in \mathbb{R}^{m \times d}$: Let $\bar{v} \in \mathcal{F}_B$ be optimal for (X, Y), i.e., $\gamma = \gamma(X, Y) = \gamma_{\bar{v}}$. As in Lemma 5.3 construct $\bar{v}' \in \mathcal{F}_B$ with $\mathbb{E}[\langle \bar{v}'(z), x' \rangle \mid z \in C(V)] = \mathbb{E}[\langle \bar{v}(z), x' \rangle \mid z \in C(V)]$ where $\bar{v}'$ is constant for any cone of the form $C(V)$. Using the Chernoff bound (2.9) we get with failure

probability at most $2\exp(\frac{1}{8}\cdot\frac{b}{7}\cdot m') = 2\exp(\frac{1}{224}\cdot\gamma\cdot m')$ that the number n_V of points $w_{j,0}$ in $C(V)$ lies in the interval $[\frac{P(V)m}{2}, 2P(V)m]$. Now using $m' \geq 224\gamma^{-1}\log(\frac{2n}{\delta})$ and applying a union bound we get that this holds for all cones of the form $C(V)$ with failure probability at most 2δ. For $w_j \in C(V)$ we define $u_j = a_j\frac{\bar{v}'(w_j)}{\sqrt{m}}\cdot\frac{P(V)m}{2n_V}$. Since $n_V \in [\frac{P(V)m}{2}, 2P(V)m]$ it follows that $\|u_j\|_2 \leq \frac{\|\bar{v}'(w_j)\|_2}{\sqrt{m}} \leq \frac{1}{\sqrt{m}}$ and consequently $\|U\|_F \leq 1$. Moreover we have

$$\sum_{s\in[m], w_{s,0}\in C(V)} a_s u_s = P(V)m\cdot\frac{1}{2\sqrt{m}}\cdot\bar{v}'(V),$$

where we set $\bar{v}'(V)$ to be equal to $\bar{v}'(z)$, which is constant for any $z \in C(V)$.

The third step is to prove that U is a good separator for (X, Y): To this end, let $x \in X$ and $x' = g(x_i)$. If $x_i = x'$ then

$$\begin{aligned}
y_i f_i^{(0)}(U) &= y_i\frac{1}{\sqrt{m}}\sum_{s=1}^{m} a_s\langle u_s, x_i\rangle\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right] \\
&= y_i\frac{1}{\sqrt{m}}\sum_{V\subseteq X', x'\in V}\ \sum_{s\in[m], w_{s,0}\in C(V)} a_s\langle u_s, x_i\rangle \\
&= y_i\frac{1}{2m}\sum_{V\subseteq X', x'\in V} P(V)m\cdot\langle\bar{v}'(V), x_i\rangle \\
&= y_i\frac{1}{2}\mathbb{E}[\langle\bar{v}(z), x_i\rangle\mathbf{1}\left[\langle x_i, z\rangle > 0\right]] \\
&= y_i\frac{1}{2}\int\langle\bar{v}(z), x_i\rangle\mathbf{1}[\langle x_i, z\rangle > 0]d\mu_{\mathcal{N}}(z) \geq \frac{\gamma}{2}.
\end{aligned}$$

Otherwise if $x_i \neq x'$ then there is exactly one cone $C(V_1)$ with $z \in C(V_1)$ such that $\langle x', z\rangle < 0$ and $\langle x_i, z\rangle > 0$ and exactly one cone $C(V_2)$ with $z \in C(V_2)$ such that $\langle x', z\rangle > 0$ and $\langle x, z\rangle < 0$. Recall that $P(V_i) = \frac{1}{n'} \leq b$ for $i = 1, 2$. We set $M = \{V \subseteq [n'] \mid x' \in V, V \notin \{V_1, V_2\}\}$. Then it holds that

$$\begin{aligned}
&y_i f_i^{(0)}(U) = \frac{1}{\sqrt{m}}\sum_{s=1}^{m} y_i\langle u_s, x_i\rangle\mathbf{1}\left[\langle x_i, w_{s,0}\rangle > 0\right] \\
&\geq \frac{1}{\sqrt{m}}\left(\sum_{V\in M}\ \sum_{s\in[m], w_{s,0}\in C(V)} y_i\langle u_s, x_i\rangle\ -\ \sum_{s\in[m], w_{s,0}\in C(V_1)}|\langle u_s, x_i\rangle|\right) \\
&\geq \frac{1}{\sqrt{m}}\left(\sum_{V\in M}\ \sum_{w_{s,0}\in C(V)} y_i\langle u_s, x_i\rangle + \sum_{w_{s,0}\in C(V_2)}|\langle u_s, x_i\rangle| - \sum_{s\in[m], w_{s,0}\in C(V_2)}|\langle u_s, x_i\rangle| - \frac{1}{2\sqrt{m}}P(V_1)m\right) \\
&\geq \frac{1}{\sqrt{m}}\left(\frac{\sqrt{m}}{2}\mathbb{E}[y_i\langle\bar{v}(z), x_i\rangle\mathbf{1}\left[\langle x_i, z\rangle > 0\right]]\ -\ \frac{1}{2\sqrt{m}}P(V_2)m\ -\ \frac{1}{2\sqrt{m}}P(V_1)m\right) \\
&= \frac{1}{2}\left(\mathbb{E}[y_i\langle\bar{v}(z), x_i\rangle\mathbf{1}\left[\langle x_i, z\rangle > 0\right]] - 2b\right) \geq \frac{1}{2}\left(\gamma - \frac{\gamma}{2}\right) = \frac{\gamma}{4}.
\end{aligned}$$

□

6 Conclusion and open problems

We summarize the main results of the present manuscript and pose related open questions.

6.1 Sketching for logistic regression

We have introduced the to our knowledge first data oblivious sketch for logistic regression. The sketch consists of multiple levels and each level is a Count-Min sketch of a subsample of the data where the size of the subsample depends on the level. We show that the optimal solution computed on the sketch is a good approximation to the optimal solution on the original data. More precisely we can get a constant approximation with sketch size almost linear in the dimension d in almost linear running time (in n). Further we can get a $(2+\varepsilon)$-approximation in expectation with sketch size polynomial in the dimension $d, \ln(n)$ and ε in linear running time. Last we can get a $(1+\varepsilon)$-approximation with sketch size exponential in ε^{-1} in linear running time.

We then showed that our sketch can also be used for ℓ_1-regression as well as variance-based regularized logistic regression. For ℓ_1-regression the size of our sketch is close to linear in d and is constructed only by multiplying a random matrix to the orignal data matrix.

Our main contributions are not only the results themselves but also the in depth analysis. Here we first fix a parameter β and divide the vector $z = X\beta$ into several weight classes similarly as in (Clarkson and Woodruff, 2015). We showed that the contraction bounds hold by showing that for any weight class with a notable contribution there is a level that preserves the contribution. We then apply a standard net argument to get the contraction bound for all possible parameter vectors. Moreover to achieve the dilation bounds we show that for any weight class there are only few levels where the weight class can have a non zero contribution. To get the approximation ratio below 2 in expectation we then apply random shift by choosing the size of the first level at random.

There are mainly two possibilities for future research. The first one is to improve the analysis (and possibly the algorithm itself) further reducing sketch size, approximation and/or running time. In particular it might be possible to reduce the dependence of d here by using a different net argument. The second future research direction is to look at other target functions such as Poisson regression, p-probit regression, etc. to see how well the sketch performs and to summarize the set of functions our sketch performs well on. Note that it is likely that the more closely the target function is to the ℓ_1-norm the better the sketch will perform. More precisely if the elements with large contribution to the target function also have a large contribution to the ℓ_1 norm and the other elements behave uniform then it is likely that our sketch will work on those functions as well.

6.2 ℓ_p-leverage score sampling for probit regression

Using sensitivity sampling we constructed an ε-coreset for p-generalized sketching. More precisely we analyzed both, the tail behavior of p-generalized normal distribution in a similar manner as (Gordon, 1941) the standard normal distribution as well as the log-likelihood of the p-probit model, to show that the sensitivity scores with respect to the log-likelihood of the p-probit model are proportional to the ℓ_p-leverage scores plus $\frac{1}{n}$. Moreover by rounding the weights we were able to show the VC-dimension of the function space with rounded weights is bounded by $O(d\log(\mu/\varepsilon))$. The same technique can also be used to reduce the VC-dimension used in (Munteanu et al., 2018) from $O(d\log(n))$ to $O(d\log(\mu/\varepsilon))$ completely removing the dependency on n. Using a first pass over the data we used a sketching algorithm invented by (Woodruff and Zhang, 2013) to get a sketch of the data preserving the ℓ_p-norm up to some factor. In a second pass over the data, we computed an approximation to the ℓ_p-leverage scores and sampling probabilities and plug the into a reservoir sampler. Using the sensitivity framework (Braverman et al., 2021; Feldman, 2020) we proved that this yields an ε-coreset.

There are a couple of questions that remain open. Similar to logistic regression there is no upper bound of the VC dimension of the range space of the function space considered in the analysis, i.e. $\mathcal{F} = \{wg_x \mid x \in \mathbb{R}^d, w \in \mathbb{R}_{\geq 0}\}$ (where $g_x(\beta) = -\ln(\Phi_p(-x\beta))$, known. We get around this by limiting the number of weights however it would be interesting to know whether the VC dimension is d which would further improve our analysis. Moreover it might be possible to use a different net argument such as in (Musco et al., 2022) to get an even lower dependence on d in the size of our coreset. More generally it would be desirable to know whether it is possible to construct a coreset that works for multiple values of p. Finally one might also look at the problem from a Bayesian perspective as suggested in (Geppert et al., 2017). Here instead of optimizing at the negative log likelihood we look at the distribution of β's we get by the the likelihoods of β. Can we bound the Wasserstein distance between the original distribution and the distribution on the coreset?

6.3 Reducing the width of two layer ReLU networks

Finally we studied two layer ReLU networks. Here we analyzed the performance of gradient descent coupled initialization. For this initialization technique, which appeared before independently in (Daniely, 2020), we showed that using this initialization technique we can get the required width to get an error of less than ε down from $\tilde{O}(\gamma^{-8})$(Ji and Telgarsky (2020)) to $\tilde{O}(\gamma^{-2})$ for any instance with separation margin γ. We further gave some intuition to get a better understanding of the parameter γ by stating and proving some properties as well as giving examples where bounds for γ can be obtained. Using one of these examples we also gave a lower bound of $\Omega(\gamma^{-1})$ for the width of any two layer ReLU network achieving an error of less than ε. We further proved lower bounds of $\Omega(\gamma^{-1}\log(n))$ and $\Omega(\gamma^{-2}\log(n))$ for parts of our analysis.

There remains a gap of $\Omega(\gamma^{-1})$ and $\tilde{O}(\gamma^{-2})$ in the bound on the width. For the 2-dimensional case we have seen, that a width of $\tilde{O}(\gamma^{-1})$ suffices however it remains open whether this is true for any $d > 2$. Note

that the squared dependence on γ^{-1} comes from the Hoeffding bound. In order to get rid of it in the case $d = 2$ we used a different approach to proof the existence of a good direction on the NTK. As the structure of point sets in higher dimensional spaces becomes much more complex it is unclear whether a similar approach could work here. On the other hand in (Daniely, 2020) it was noticed that with higher dimension the upper bound on the width for random points on the sphere becomes even less, being $\tilde{O}(n/d)$, so it might even be possible that with higher dimensions the width necessary for convergence is even less than just $\tilde{O}(\gamma^{-1})$ if all subsets of the dataset of size d are of rank d.

Further it would be interesting to also consider networks of higher depths as done in (Seleznova and Kutyniok, 2022) to see if it is possible to get sparser networks in this setting as well by using coupled initialization.

7 Bibliography

Yuqing Ai, Wei Hu, Yi Li, and David P. Woodruff. New characterizations in turnstile streams with applications. In *31st Conference on Computational Complexity*, pages 20:1–20:22, 2016.

Nir Ailon and Edo Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, 42:615–630, 2009.

Nir Ailon and Supratim Shit. Efficient NTK using dimensionality reduction. *CoRR*, abs/2210.04807, 2022.

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. In *Advances in Neural Information Processing Systems*, pages 6155–6166, 2019a.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *Proceedings of the 36th International Conference on Machine Learning*, pages 242–252, 2019b.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Conference on Neural Information Processing Systems*, 2019c.

Noga Alon, László Babai, and Alon Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, 7(4):567–583, 1986.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, 58(1):137–147, 1999.

Alexandr Andoni, Khanh Do Ba, Piotr Indyk, and David P. Woodruff. Efficient sketches for earth-mover distance, with applications. In *50th Annual IEEE Symposium on Foundations of Computer Science*, pages 324–330, 2009.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 2019a.

Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *34th Conference on Computational Complexity*, 2019b.

Herman Auerbach. *On the area of convex curves with conjugate diameters*. PhD thesis, University of Lwów, 1930.

Ainesh Bakshi, Rajesh Jayaram, and David P. Woodruff. Learning two layer rectified neural networks in polynomial time. In *Conference on Learning Theory*, pages 195–268, 2019.

Jon Louis Bentley and James B. Saxe. Decomposable searching problems I: Static-to-dynamic transformation. *Journal of Algorithms*, 1(4):301–358, 1980.

Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Avrim Blum, John Hopcroft, and Ravi Kannan. *Foundations of Data Science*. Cambridge University Press, 2020.

Allan Borodin and Ran El-Yaniv. *Online computation and competitive analysis*. Cambridge University Press, 2005.

Jan van den Brand, Binghui Peng, Zhao Song, and Omri Weinstein. Training (overparametrized) neural networks in near-linear time. In *12th Innovations in Theoretical Computer Science Conference*, 2021.

Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. Efficient coreset constructions via sensitivity sampling. In *Asian Conference on Machine Learning*, volume 157, pages 948–963, 2021.

Bo Brinkman and Moses Charikar. On the impossibility of dimension reduction in l_1. *Journal of the ACM*, 52(5):766–788, 2005.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Sébastien Bubeck, Ronen Eldan, Yin Tat Lee, and Dan Mikulincer. Network size and size of the weights in memorization with two-layers neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pages 10835–10845, 2019.

Yuan Cao and Quanquan Gu. Generalization error bounds of gradient descent for learning over-parameterized deep ReLU networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3349–3356, 2020.

Min-Te Chao. A general purpose unequal probability sampling plan. *Biometrika*, 69(3):653–656, 1982.

Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theor. Comput. Sci.*, 312(1):3–15, 2004.

Lin Chen and Sheng Xu. Deep neural tangent kernel and Laplace kernel have the same RKHS. *arXiv preprint arXiv:2009.10683*, 2020.

Sitan Chen, Adam R Klivans, and Raghu Meka. Learning deep ReLU networks is fixed-parameter tractable. In *IEEE 62nd Annual Symposium on Foundations of Computer Science*, pages 696–707, 2022.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, pages 493–507, 1952.

Rachit Chhaya, Jayesh Choudhari, Anirban Dasgupta, and Supratim Shit. Streaming coresets for symmetric tensor factorization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1855–1865, 2020.

Kenneth L. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 257–266, 2005.

Kenneth L. Clarkson and David P. Woodruff. Sketching for M-estimators: A unified approach to robust regression. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 921–939, 2015.

Kenneth L. Clarkson and David P Woodruff. Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, 63(6):1–45, 2017.

Kenneth L. Clarkson, Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, Xiangrui Meng, and David P. Woodruff. The fast Cauchy transform and faster robust linear regression. *SIAM Journal on Computing*, 45(3):763–810, 2016.

Kenneth L. Clarkson, Ruosong Wang, and David P. Woodruff. Dimensionality reduction for Tukey regression. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1262–1271, 2019.

Michael B. Cohen, Cameron Musco, and Jakub Pachocki. Online row sampling. *Theory of Computing*, 16: 1–25, 2020.

R. Dennis Cook. Detection of influential observation in linear regression. *Technometrics*, 19(1):15–18, 1977.

Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

Amit Daniely. Neural networks learning and memorization with (almost) no over-parameterization. In *Advances in Neural Information Processing Systems*, 2020.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W. Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal of Computing*, 38(5):2060–2078, 2009.

Simon S. Du, Kangcheng Hou, Russ R. Salakhutdinov, Barnabas Poczos, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. *Advances in Neural Information Processing Systems*, 32, 2019a.

Simon S. Du, Jason D. Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, 2019b.

Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations*, 2019c.

John C. Duchi and Hongseok Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20:68:1–68:55, 2019.

Alex Dytso, Ronit Bustin, H. Vincent Poor, and Shlomo Shamai. Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5:1–40, 12 2018.

Paul Erdős and Alfréd Rényi. On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 6:215–220, 1961.

Dan Feldman. Core-sets: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(1), 2020.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing*, pages 569–578, 2011.

Dan Feldman, Melanie Schmidt, and Christian Sohler. Turning Big Data into tiny data: Constant-size coresets for k-means, PCA, and projective clustering. *SIAM Journal of Computing*, 49(3):601–657, 2020.

William Feller. Generalization of a probability limit theorem of Cramér. *Trans. Am. Math. Soc.*, 54:361–372, 1943.

Rong Ge, Jason D. Lee, and Tengyu Ma. Learning one-hidden-layer neural networks with landscape design. In *International Conference on Learning Representations*, 2018.

Leo N. Geppert, Katja Ickstadt, Alexander Munteanu, Jens Quedenfeld, and Christian Sohler. Random projections for Bayesian regression. *Statistics and Computing*, 27(1):79–101, 2017.

Felix Gessert, Wolfram Wingerath, Steffen Friedrich, and Norbert Ritter. NoSQL database systems: a survey and decision guidance. *Computer Science - Research and Development*, 32(3-4):353–365, 2017.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

Irwin R. Goodman and Samuel Kotz. Multivariate θ-generalized normal distributions. *Journal of Multivariate Analysis*, 3(2):204–219, 1973.

Robert D. Gordon. Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Jürgen Groß. *Linear regression*, volume 175. Springer Science & Business Media, 2003.

Joseph M. Hilbe. *Logistic regression models*. Chapman and Hall/CRC, 2009.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In *International Conference on Machine Learning*, pages 4542–4551, 2020.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems*, pages 4080–4088, 2016.

Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM*, 53(3):307–323, 2006.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: convergence and generalization in neural networks. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8580–8589, 2018.

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta mathematica*, 30(1):175–193, 1906.

Ziwei Ji and Matus Telgarsky. Polylogarithmic width suffices for gradient descent to achieve arbitrarily small test error with shallow ReLU networks. In *International Conference on Learning Representations*, 2020.

Norman L. Johnson, Samuel Kotz, and Narayanaswamy Balakrishnan. *Continuous univariate distributions, Volume 1*. Wiley & Sons, 2nd edition, 1994.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, 26(1):189–206, 1984.

S. Kalke and W.-D. Richter. Simulation of the p-generalized Gaussian distribution. *Journal of Statistical Computation and Simulation*, 83(4):641–667, 2013.

Kenji Kawaguchi and Jiaoyang Huang. Gradient descent finds global minima for generalizable deep neural networks of practical sizes. In *57th Annual Allerton Conference on Communication, Control, and Computing*, pages 92–99, 2019.

Michael J. Kearns and Umesh V. Vazirani. *An Introduction to Computational Learning Theory*. MIT Press, Cambridge, 1994.

Christian Kleiber and Samuel Kotz. *Statistical size distributions in economics and actuarial sciences*. John Wiley & Sons, 2003.

Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

Ilan Kremer, Noam Nisan, and Dana Ron. On randomized one-round communication complexity. *Computational Complexity*, 8(1):21–49, 1999.

Michael Langberg and Leonard J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 598–607, 2010.

Jason D. Lee, Ruoqi Shen, Zhao Song, Mengdi Wang, and Zheng Yu. Generalized leverage score sampling for neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021a.

Yi Li, Huy L. Nguyen, and David P. Woodruff. Turnstile streaming algorithms might as well be linear sketches. In *Symposium on Theory of Computing,*, pages 174–183, 2014.

Yi Li, David P. Woodruff, and Taisuke Yasuda. Exponentially improved dimensionality reduction for ℓ_1: Subspace embeddings and independence testing. In *Conference on Learning Theory*, pages 3111–3195, 2021b.

Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, 2018.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with ReLU activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification - simplified and strengthened. In *Advances in Neural Information Processing Systems*, pages 11643–11654, 2021.

Tung Mai, Alexander Munteanu, Cameron Musco, Anup B. Rao, Chris Schwiegelshohn, and David P. Woodruff. Optimal sketching bounds for sparse linear regression. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proc. of the 22nd Conference on Learning Theory*, 2009.

P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman & Hall, London, 1989.

Xiangrui Meng and Michael W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 91–100, 2013.

Alejandro Molina, Alexander Munteanu, and Kristian Kersting. Core dependency networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 3820–3827. AAAI Press, 2018.

Alexander Munteanu. *On large-scale probabilistic and statistical data analysis*. PhD thesis, Technische Universität Dortmund, 2018.

Alexander Munteanu. Coresets and sketches for regression problems on data streams and distributed data. In *Machine Learning under Resource Constraints, Volume 1 - Fundamentals*, pages 85–98. De Gruyter, 2023.

Alexander Munteanu and Chris Schwiegelshohn. Coresets-methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intelligenz*, 32(1):37–53, 2018.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David P. Woodruff. On coresets for logistic regression. In *Advances in Neural Information Processing Systems*, pages 6562–6571, 2018.

Alexander Munteanu, Simon Omlor, and David P. Woodruff. Oblivious sketching for logistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7861–7871, 2021.

Alexander Munteanu, Simon Omlor, and Christian Peters. p-Generalized probit regression and scalable maximum likelihood estimation via sketching and coresets. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pages 2073–2100, 2022a.

Alexander Munteanu, Simon Omlor, Zhao Song, and David Woodruff. Bounding the width of neural networks via coupled initialization - a worst case analysis. In *International Conference on Machine Learning*, pages 16083–16122, 2022b.

Alexander Munteanu, Simon Omlor, and David P. Woodruff. Almost linear constant-factor sketching for ℓ_1 and logistic regression. In *Proceedings of the 11th International Conference on Learning Representations*, 2023. to appear.

Cameron Musco, Christopher Musco, David P. Woodruff, and Taisuke Yasuda. Active linear regression for ℓ_p norms and beyond. In *63rd IEEE Annual Symposium on Foundations of Computer Science*, pages 744–753, 2022.

S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, 1(2), 2005.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Atsushi Nitanda, Geoffrey Chinot, and Taiji Suzuki. Gradient descent can learn less over-parameterized two-layer neural networks on classification problems. *arXiv preprint arXiv:1905.09870*, 2019.

Jacek Osiewalski and Mark F. J. Steel. Robust Bayesian inference in ℓ_q-spherical models. *Biometrika*, 80(2): 456–460, 1993.

Samet Oymak and Mahdi Soltanolkotabi. Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks. *IEEE Journal on Selected Areas in Information Theory*, 1 (1):84–105, 2020.

Jeff M. Phillips. Coresets and sketches. In *Handbook of Discrete and Computational Geometry*, pages 1269–1288. Chapman and Hall/CRC, 3rd edition, 2017.

Hossein Pishro-Nik. *Introduction to probability, statistics, and random processes*. Kappa Research LLC, 2014.

David Rohr. Data processing and online reconstruction. *arXiv preprint arXiv:1811.11485*, 2018.

Florin Rusu and Alin Dobra. Pseudo-random number generation for sketch-based estimations. *ACM Transactions on Database Systems*, 32(2):1–48, 2007.

Alireza Samadian, Kirk Pruhs, Benjamin Moseley, Sungjin Im, and Ryan R. Curtin. Unconditional coresets for regularized loss minimization. In *The 23rd International Conference on Artificial Intelligence and Statistics,*, pages 482–492, 2020.

Tamás Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 143–152, 2006.

Mariia Seleznova and Gitta Kutyniok. Neural tangent kernel beyond the infinite-width limit: Effects of depth and initialization. In *International Conference on Machine Learning*, pages 19522–19560, 2022.

Aisha Siddiqa, Ahmad Karim, and Abdullah Gani. Big Data storage technologies: A survey. *Frontiers of Information Technology & Electronic Engineering*, 18(8):1040–1070, 2017.

Fabian Sinz, Sebastian Gerwinn, and Matthias Bethge. Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820, 2009.

Christian Sohler and David P. Woodruff. Subspace embeddings for the ℓ_1-norm with applications. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 755–764, 2011.

Zhao Song and Xin Yang. Quadratic suffices for over-parametrization via matrix Chernoff bound. *arXiv preprint arXiv:1906.03593*, 2019.

Zhao Song, Shuo Yang, and Ruizhe Zhang. Does preprocessing help training over-parameterized neural networks? *Advances in Neural Information Processing Systems*, 2021.

StEx StEx. How can we sum up sin and cos series when the angles are in arithmetic progression? `https://math.stackexchange.com/questions/17966/`, 2011. Accessed: 2021-05-21.

Marco Stolpe, Kanishka Bhaduri, Kamalika Das, and Katharina Morik. Anomaly detection in vertically partitioned data by distributed core vector machines. In *Machine Learning and Knowledge Discovery in Databases - European Conference*, pages 321–336, 2013.

Mikhail F. Subbotin. On the law of frequency of error. *Matematicheskii Sbornik*, 31(2):296–301, 1923.

Elad Tolochinsky, Ibrahim Jubran, and Dan Feldman. Generic coreset for scalable learning of monotonic kernels: Logistic regression, sigmoid and more. In *39th International Conference on Machine Learning*, pages 21520–21547, 2022.

Murad Tukan, Alaa Maalouf, and Dan Feldman. Coresets for near-convex functions. In *Advances in Neural Information Processing Systems*, 2020.

Elad Verbin and Qin Zhang. Rademacher-sketch: A dimensionality-reducing embedding for sum-product norms, with an application to earth-mover distance. In *Automata, Languages, and Programming - 39th International Colloquium*, pages 834–845, 2012.

Ruosong Wang and David P. Woodruff. Tight bounds for ℓ_1 oblivious subspace embeddings. *ACM Transactions on Algorithms*, 18(1):8:1–8:32, 2022.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1-2):1–157, 2014.

David P. Woodruff. Problem set (+ solution), 2021. `http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall21/ps3.pdf`, Solution: `http://www.cs.cmu.edu/afs/cs/user/dwoodruf/www/teaching/15859-fall21/hw3Solutions.pdf`, Accessed: 5-18-2022.

David P Woodruff and Taisuke Yasuda. Online Lewis weight sampling. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 4622–4666, 2023.

David P. Woodruff and Qin Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In *Proceedings of the 26th Annual Conference on Learning Theory*, pages 546–567, 2013.

Yan Yan, Yi Xu, Lijun Zhang, Xiaoyu Wang, and Tianbao Yang. Stochastic optimization for non-convex inf-projection problems. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10660–10669, 2020.

Jiawei Zhang, Yushun Zhang, Mingyi Hong, Ruoyu Sun, and Zhi-Quan Luo. When expressivity meets trainability: Fewer than n neurons can work. In *Advances in Neural Information Processing Systems*, pages 9167–9180, 2021.

Kai Zhong, Zhao Song, and Inderjit S. Dhillon. Learning non-overlapping convolutional neural networks with multiple kernels. *arXiv preprint arXiv:1711.03440*, 2017a.

Kai Zhong, Zhao Song, Prateek Jain, Peter L. Bartlett, and Inderjit S. Dhillon. Recovery guarantees for one-hidden-layer neural networks. In *34th International Conference on Machine Learning*, 2017b.

Difan Zou and Quanquan Gu. An improved analysis of training over-parameterized deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2053–2062, 2019.