On large-scale probabilistic and statistical data analysis

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik

von

Alexander Munteanu

Dortmund
2018

Tag der mündlichen Prüfung:	25. Juli 2018
Dekan:	Prof. Dr. Gernot Fink
Gutachter:	Prof. Dr. Christian Sohler
	Prof. Dr. Katja Ickstadt

Abstract

In this manuscript we develop and apply modern algorithmic data reduction techniques to tackle scalability issues and enable statistical data analysis of massive data sets. Our algorithms follow a general scheme, where a reduction technique is applied to the large-scale data to obtain a small summary of sublinear size to which a classical algorithm is applied. The techniques for obtaining these summaries depend on the problem that we want to solve. The size of the summaries is usually parametrized by an approximation parameter, expressing the trade-off between efficiency and accuracy. In some cases the data can be reduced to a size that has no or only negligible dependency on n, the initial number of data items. However, for other problems it turns out that sublinear summaries do not exist in the worst case. In such situations, we exploit statistical or geometric relaxations to obtain useful sublinear summaries under certain mildness assumptions. We present, in particular, the data reduction methods called *coresets* and *subspace embeddings*, and several algorithmic techniques to construct these via random projections and sampling.

First, we consider the problem of Bayesian linear regression, where the aim is to approximate the posterior distribution. We show approximation results on the normal distribution defined over ℓ_2 spaces and generalize to ℓ_p spaces for $p \in [1, \infty)$. Specifically, we obtain $(1+\varepsilon)$-approximations via ℓ_2 subspace embeddings which reduce the data to only $O(\varepsilon^{-2}d^{O(1)})$ points. For ℓ_p, the results have approximation errors of $(4+2\varepsilon)$ and the final corsets are larger but still in $O(\varepsilon^{-2}d^{O(1)})$ independent of n.

Hereafter, we develop the first coresets for probabilistic graph models called *dependency networks* involving (generalized) linear regression models. We show that a coreset of size $O(\varepsilon^{-2}d \log d)$ exists in the case of normal distributions implying an approximation error of $(1+\varepsilon)$. For Poisson as well as logistic regression, we obtain the first linear lower bounds on any data reduction preserving their objectives up to large constant factors. A statistical relaxation provides intuition why coresets for the ℓ_2 problem yield a good approximation for Poisson regression anyway. To tackle the limitation in logistic regression, we introduce a novel complexity measure μ for compressing the data. We show for data, attaining a sufficiently small value of μ, that we can compute coresets of sublinear size $\tilde{O}(\varepsilon^{-2}\mu d^{3/2}\sqrt{n}\log^{O(1)} n)$. A recursive application of the algorithm yields first coresets of polylogarithmic size $\tilde{O}(\varepsilon^{-4}\mu^6 d^3 \log^{O(1)} n)$.

Finally, we develop the first $(1+\varepsilon)$-approximation algorithm for the smallest enclosing ball problem on probabilistic data. Combining reductions to deterministic geometric problems, random sampling techniques, and grid based coreset constructions, we obtain the approximation in roughly $\tilde{O}(\varepsilon^{-3}n + \varepsilon^{-O(d)})$ time on a summary of only $O(\varepsilon^{-O(d)})$ points.

Contents

1 Introduction and motivation

Social media, online retailers, streaming services and other online platforms have infiltrated everyone's everyday life for the past two decades. The ease of use and permanent availability make them very attractive. People are continuously using these technologies and leaving their user data and electronic fingerprints behind. Sensors in consumer electronics are continuously collecting physical measurements and usage data. Medical and financial records of individuals are gathered and stored digitally.

Big Data is now ubiquitous, and great value lies in understanding the data. Knowledge derived from the data can, for instance, be used in advertising to make offers more appealing to the individual, in medicine to assist the doctor in finding common or unusual patterns in disease diagnosis, or in physics to find rare or even unknown particles within a plethora of rather uninteresting observations. Modern statistics, machine learning, and artificial intelligence achieved considerable successes in recent years, and an ever-growing number of disciplines rely on them.

The area of statistics has brought up a number of classical results on the asymptotic distributions of data, like the laws of large numbers and central limit theorems. Massive data sets are highly desirable in this light, since they can provide extremely reliable information on the average behavior of large groups of individuals. However, with the advent of massive data sets, scalability has become crucial for any useful data analysis approach.

This manuscript contributes to the theoretical foundations of massive data analysis by developing methods to

- compress the data to a tractable size, and
- preserve its properties with respect to statistical models.

1.1 Data compression

Scalability is arguably the most important and central challenge in modern computational statistics, machine learning, and related optimization problems. Algorithms, whose running times are polynomial in the input size, might be regarded as efficient in a conventional sense. Nevertheless, the computations become tedious or even intractable when applied to massive data sets. Additional complications arise when data is stored on slow external memory devices. The data cannot be accessed in random order, but is limited to sequential streaming access, or the data might be distributed on physically separated devices. Computing devices are also limited in their available resources, like computational power, internal memory or communication capacities.

As a result, performing data reduction techniques in a preprocessing step to speed up a subsequent data analysis or optimization task has received considerable attention. This is commonly known as the *sketch and solve* paradigm stemming from the theory of approximation and streaming algorithms. If the data is reduced from n to $k \ll n$ points, obviously, this saves memory and communication requirements. Consequently, the running time of a subsequent computation task is also reduced from $T(n)$ to $T(k)$. It is highly desirable and often turns out possible, that k has no or only a low polylogarithmic dependency on n.

Compression techniques The main techniques that we consider, are random projections and coresets. There is extensive work on both of these methods. Both techniques often apply to streaming and distributed environments via standard approaches or little modifications.

Random projections can be used to reduce the dimension of vectors while maintaining the algebraic structure of their entire spanned vector space, up to little distortion. Data can usually be interpreted as a collection of vectors and random projections are thus a suitable technique to reduce the size of the data. They have successfully been applied to several problems in the area of numerical linear algebra.

Coresets are geometric summaries or subsamples of points designed to approximate the objective function of a problem for all candidate solutions. They have been introduced in the context of fast approximation algorithms for clustering and shape fitting problems. Coresets have subsequently been developed for a plethora of computational problems.

1.2 Statistical data analysis

In statistical data analysis, it is commonly assumed that observations are drawn from a fixed generating distribution, which is unknown or stems from a parametrized family of distributions, whose parameters are unknown. Additionally we may assume that there is to some amount a noisy error in the observations which is not observable or separable. The shape and amount of noise might be unknown. For example a scale might weigh up to deviations of $\pm 1g$ around the true mass, but human behavior might deviate and even change from its *typical* characteristic on a daily basis. Consider, for instance, the location of an individual. People move from home, to work, school or to a shopping mall.

The variety in behavior is thus modeled as a random variable Y. The *frequentist* approach to statistical inference is now to consider the most likely value of Y according to its distribution as the fixed ground truth. In several cases like linear regression with Gaussian error, this corresponds to the expectation $\mathbb{E}[Y]$ of Y. In *Bayesian statistics*, however, one assumes that the ground truth itself is again a random variable, depending on a given set of fixed observations. Additionally one may assume prior belief on how likely different models or explanations may be.

Preserving statistical properties Many important statistical data analysis tasks like inference in (generalized) linear regression models or probabilistic shape fitting, rely on numerical linear algebra, multivariate calculus or solving geometric problems on the data.

As discussed before, random projection and coreset techniques are developed to reduce the number of observations in a data set, and hereby introduce only little error on the underlying algebraic or geometric structure induced by the point set. The results in this manuscript show for several statistical data analysis problems that a suitable data compression also preserves their statistical model up to little errors. These properties imply that performing the analysis on the reduced data yields controllable errors and comparable results to the same analysis on the original massive data set, which might be computationally intractable.

1.3 Problems considered in this manuscript

The unifying algorithmic approach is sketched in the following scheme, where Π is a problem specific map, i.e., an algorithm that maps the large data set X to a significantly smaller data set $\Pi(X)$. By $f(\beta \mid X)$ we denote some function that depends on the data and that we want to approximate depending on an approximation parameter ε, for all or special choices of solutions β.

$$\begin{tikzcd}
X \arrow[r, "\Pi"] \arrow[d] & \Pi(X) \arrow[d] \\
f(\beta \mid X) \arrow[r, phantom, "\approx_\varepsilon"] & f(\beta \mid \Pi(X)).
\end{tikzcd}$$

Problems The functions f that we are going to approximate within this general approach are the objective functions resp. posterior distributions of

- (Bayesian) linear regression with normal as well as p-generalized normal error,
- Gaussian dependency networks,
- Poisson regression,
- logistic regression, and
- the smallest enclosing ball problem on probabilistic data points.

1.4 Outline and results

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

Chapter 2 Here we introduce general notation and give a review of important concepts from linear algebra. We continue with important statistical preliminaries on probability distributions, Bayesian linear regression, generalized linear regression models and related computational problems. Hereafter we present several methods to reduce the size of a given data set, while preserving its properties with respect to a given objective function. Namely, we will introduce coresets and subspace embeddings, and several algorithmic approaches to construct these via random projections and sampling techniques. Finally we introduce the notion of probabilistic data and define important problems in that area.

Chapter 3 In this chapter we consider the problem of Bayesian linear regression. We show approximation results on the normal posterior distribution defined over ℓ_2 spaces. Specifically, we show that any ε-subspace embedding applied as a data reduction technique, preserves the underlying normal posterior distribution up to little distortion depending on only ε-fractions of its defining location and covariance parameters. Under mild assumptions, this can be shown to be a $(1 + O(\varepsilon))$-approximation with respect to the second moment of the normal distribution. The embedding dimension depends on the method but can be as little as $O(\varepsilon^{-2}d)$, allowing for efficient Bayesian regression analysis where the memory and time do not depend on the massive initial size parameter n.

We further extend the method to Bayesian linear regression for p-generalized normal distributions defined over ℓ_p spaces, for $p \in [1, \infty)$. The previous relative approximation errors become as weak as $4 + 2\varepsilon$ and the intermediate embedding dimension is significantly larger, especially for values of $p > 2$ where they quickly approach linear size. However, the final sampling based coresets have size roughly $\tilde{O}(\varepsilon^{-2}d^{2p+3}\log^2 d)$, where $\tilde{O}(\cdot)$ hides polylogarithmic factors independent of n.

These results extend existing work on approximation algorithms for maximum likelihood estimation for ℓ_p-regression models to preserving the entire posterior distributions studied in the Bayesian setting.

Chapter 4 This chapter deals with coresets for graph structures involving different (generalized) linear regression models. First we show that a single coreset based on sampled ε-subspace embeddings of size $O(\varepsilon^{-2}d\log d)$ preserves all local conditional models simultaneously and thus preserves the total structure of Gaussian dependency networks up to a factor of $(1+\varepsilon)$.

Motivated by this result we study coresets for Poisson regression and show that unfortunately, no sublinear coresets exist for this generalized linear model. However, leveraging the statistical relaxation of the Poisson log-normal model for count data, we get an intuition why the same sampling based coresets for the ℓ_2 problem yield a good approximation for Poisson regression.

Finally, we study coresets for logistic regression. Our first contribution is again a linear lower bound on the size of a coreset in the worst case. To tackle this limitation, we introduce a novel complexity measure $\mu(X)$ and outline its natural statistical interpretation. Our analysis is parametrized with the value of μ, setting it into the light of beyond worst-case analysis. Specifically, we show for μ-complex data, attaining a sufficiently small value of μ that we can compute coresets of sublinear size $\tilde{O}(\varepsilon^{-2}\mu d^{3/2}\sqrt{n}\log^{O(1)} n)$. A recursive application of our main algorithm yields coresets of polylogarithmic size $\tilde{O}(\varepsilon^{-4}\mu^6 d^3\log^{O(1)} n)$. These are the first coresets for logistic regression of rigorously sublinear size.

Chapter 5 Here we develop the first $(1+\varepsilon)$-approximation algorithm for the smallest enclosing ball problem on probabilistic data, where every input point is a distribution over z locations in $\mathbb{R}^d$ for constant d. The approximation is achieved via two reductions to related 1-median problems on metric spaces; one of them being a complex space on realizations of random sets, exponential in the input size parameters. To overcome this blow-up we apply sampling of a constant number of elements for 1-median in metric spaces and since each realization can consist of up to n points, we employ coreset constructions to keep the realizations small. The algorithm runs in linear time for sampling the realizations. Solving the subsampled problem takes only constant time, though exponential in the dimension

d. This yields a total running time of roughly $\tilde{O}(\varepsilon^{-3}nz + \varepsilon^{-O(d)})$. This is still the only FPTAS for this problem to date.

Chapter 6 In this final chapter we conclude our work and propose interesting directions and challenging open questions for future research.

1.5 Publications

The present manuscript is based on the following publications. All authors contributed equally and are stated in alphabetical order.

- Chapter 3 is based on [62],

  L. N. Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, and C. Sohler. Random projections for Bayesian regression. *Statistics and Computing*, **27**(1):79–101, 2017

- Chapter 4 is based on [95],

  K. Kersting, A. Molina, and A. Munteanu. Core dependency networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 3820–3827, 2018

  and on [115],

  A. Munteanu, C. Schwiegelshohn, C. Sohler, and D. P. Woodruff. On coresets for logistic regression. *CoRR*, abs/1805.08571, 2018

- Chapter 5 is based on [56],

  D. Feldman, A. Munteanu, and C. Sohler. Smallest enclosing ball for probabilistic data. In *Proceedings of the 30th Annual Symposium on Computational Geometry (SoCG)*, pages 214–223, 2014

2 Preliminaries

2.1 General notation and brief review of linear algebra

General notation We introduce some notation for a concise presentation. We denote by $[n] = \{1, \dots, n\}$ the set of all positive integers up to n. For a random variable X and a probability measure λ, we write $X \sim \lambda$ to indicate that X is distributed according to λ. Let $\mathbb{E}_\lambda[X] = \int_\Omega x \, \mathrm{d}\lambda(x)$ be the expected value of $X \sim \lambda$ with respect to λ over the domain Ω. We skip the subscript in $\mathbb{E}[X]$ if the probability measure is clear from the context. We denote $\mathbb{V}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. In our definitions we will use the gamma function which is defined as $\Gamma \colon \mathbb{R} \setminus \{-n \mid n \in \mathbb{N}_0\} \to \mathbb{R}$, mapping $x \mapsto \int_0^\infty t^{x-1} e^{-t} \, \mathrm{d}t$. Given a subset $\mathcal{E}$ of a domain D, we denote by $\mathbb{1}_\mathcal{E} : D \to \{0, 1\}$ the indicator function such that $\mathbb{1}_\mathcal{E}(d) = 1$ if $d \in \mathcal{E}$ and $\mathbb{1}_\mathcal{E}(d) = 0$ otherwise. For real valued functions f and g, we write $f \propto g$ if there exists a constant $c \in \mathbb{R}$ such that $f = cg$ holds pointwise for f and g. We are going to use accuracy parameters ε and failure probability parameters η in our studies. Here and in the rest of the manuscript we assume $0 < \varepsilon, \eta \leq \frac{1}{2}$ unless stated otherwise.

Basics on linear algebra For a matrix $M \in \mathbb{R}^{n \times d}, d \leq n$, we denote its rows by M_i for $i \in [n]$ and its columns by $M^{(i)}$ for $i \in [d]$. Also we denote by $M^{\setminus i}$ the matrix formed by all columns of M except for column i, i.e., which are indexed by $[d] \setminus \{i\}$. We call the set $\{Mx \mid x \in \mathbb{R}^d\}$ the columnspace and $\{M^T x \mid x \in \mathbb{R}^n\}$ the rowspace of M. We denote the identity matrix in d-dimensions by $I_d \in \mathbb{R}^{d \times d}$, where $(I_d)_{ii} = 1$ for all $i \in [d]$ and $(I_d)_{ij} = 0$ otherwise. We say M is orthonormal or has orthonormal columns if $M^T M = I_d$. Given a vector $w \in \mathbb{R}^d$, we denote by $D_w = \operatorname{diag}(w)$ the diagonal matrix that carries the entries of w in canonical order, i.e., $(D_w)_{ii} = w_i$ for all $i \in [d]$ and $D_{ij} = 0$ otherwise. We sometimes call such a matrix D_w a *sampling and reweighting* matrix, if w was generated by taking a sample of elements and multiplying them by weights $w_i > 0$, or $w_i = 0$ if the element is not included in the sample. We denote the standard basis vectors for $\mathbb{R}^d$ as e_i, whose entries are $(e_i)_j = 0$ for all $j \neq i$, except for entry $(e_i)_i = 1$. Another special vector is the all one vector $\mathbf{1} = \sum_{i=1}^d e_i$. It is well known, cf. [65], that every matrix has a so-called *(thin) singular value decomposition* that is unique up to permutations of the columns and rows. We will use only the *thin* version defined below in this manuscript and will simply call it singular value decomposition, omitting the additional specification.

Definition 2.1.1 (Thin singular value decomposition, singular values, [65])**.** *Let $M \in \mathbb{R}^{n\times d}$ for $d \leq n$, $d, n \in \mathbb{N}$. The thin singular value decomposition of M denoted $M = U\Sigma V^T$, consists of orthonormal matrices $U \in \mathbb{R}^{n\times d}, V \in \mathbb{R}^{d\times d}$ and one diagonal matrix $\Sigma \in \mathbb{R}^{d\times d} = D_\sigma$. Where $\sigma \in \mathbb{R}^d_{\geq 0}$ is a vector comprising the singular values $\sigma_1 \geq \ldots \geq \sigma_d \geq 0$ of M.*

We denote by $\sigma_{\max} = \sigma_1$ the largest and by $\sigma_{\min} = \sigma_d$ the smallest singular value of M and write $\sigma_i(M)$ to stress that we mean the singular values that belong to M. We denote by $\operatorname{rank}(M)$ the rank of M which denotes the dimension of the columnspace of M and equals the number of non-zero elements in σ. If M has full rank, i.e., $\operatorname{rank}(M) = d$, the two orthonormal matrices span the columnspace and rowspace of M. We say M^TM is non-singular, if $\operatorname{rank}(M) = d$, in which case there exists an inverse matrix $(M^TM)^{-1}$ such that $(M^TM)^{-1}M^TM = M^TM(M^TM)^{-1} = I_d$. The trace of M^TM equals the sum of squared singular values of M, denoted by $\operatorname{tr}\left(M^TM\right) = \sum_{i=1}^d \sigma_i^2(M)$.

In the remainder of this manuscript, we assume w.l.o.g. that all matrices have full rank which is a common assumption, e.g. in linear regression analysis, cf. [67]. We stress that our proofs carry out similarly to our presentation if the matrices are of lower rank. One only needs to replace a matrix of low rank by a basis for its row- resp. columnspace, whose existence is guaranteed by the singular value decomposition. However, doing so in our calculations may require recursive application of the singular value decomposition and makes notation unnecessarily tedious. Note that one might even use knowledge about lower rank to reduce the space and time complexities to bounds that only depend on the rank rather than on the number of dimensions and in some cases it is even possible to go below [36].

Norms and metrics We define the matrix and vector norms as well as notions of metric spaces used in this manuscript. The latter will be needed later for quantifying the distance between distributions and between (probabilistic) point sets.

Definition 2.1.2 (vector norm, cf. [65])**.** *A vector norm on $\mathbb{R}^d$ is a function $f\colon \mathbb{R}^d \to \mathbb{R}$ that satisfies the following properties for all $x, y \in \mathbb{R}^d, \alpha \in \mathbb{R}$.*

1. $f(x) \geq 0$ *(non-negativity)*

2. $f(x + y) \leq f(x) + f(y)$ *(sub-additivity)*

3. $f(\alpha x) = |\alpha| f(x)$ *(absolute homogeneity)*

Norms are denoted $f(x) = \|x\|$ where subscripts indicate special instances.

We continue with defining special instances of norms for vectors as well as for matrices.

Definition 2.1.3 (ℓ_p-norm, [65])**.** *The ℓ_p vector norm for $p \in [1, \infty), x \in \mathbb{R}^d$ is defined as*

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}.$$

The limiting case for $p = \infty$ is defined as $\|x\|_\infty = \max_{i \in [d]} |x_i|$.

We denote $(\mathbb{R}^d, \|\cdot\|_p)$ the p-normed space or simply call it ℓ_p-space. For $p = 2$ this coincides with the Euclidean space. The Euclidean vector norm $\|x\|_2 = (x^T x)^{\frac{1}{2}}$ is arguably the most important norm. It is preserved under orthonormal transformations. Let $U \in \mathbb{R}^{n \times d}$ be orthonormal, $x \in \mathbb{R}^d$, then

$$\|Ux\|_2^2 = x^T U^T U x = x^T (U^T U) x = x^T I_d \, x = x^T x = \|x\|_2^2 .$$

For a p-norm, let q such that $\frac{1}{p} + \frac{1}{q} = 1$. Then q denotes the *dual* norm of p. For dual norms Hölder's inequality states that

$$x^T y \leq \|x\|_p \|y\|_q ,$$

where the special case for $p = q = 2$ is known as Cauchy-Schwarz inequality [65].

Since $\mathbb{R}^{n \times d}$ is isomorphic to $\mathbb{R}^{nd}$, the definition of norms carries over to matrices in $\mathbb{R}^{n \times d}$ [65]. Indeed, one can also think of a vector $x \in \mathbb{R}^d$ as a matrix $x \in \mathbb{R}^{d \times 1}$. However, for historical reasons, the entry-wise 2-norm from Definition 2.1.3 is called the Frobenius norm denoted by $\|M\|_F$ for $M \in \mathbb{R}^{n \times d}$, while $\|M\|_2$ denotes the so-called spectral norm.

Definition 2.1.4 (spectral norm, cf. [65])**.** *The spectral or operator norm of a matrix $M \in \mathbb{R}^{n \times d}$ is defined as*

$$\|M\|_2 = \sup_{x \in \mathbb{R}^d \setminus \{0\}} \frac{\|Mx\|_2}{\|x\|_2},$$

where, on the right hand side $\|\cdot\|_2$ denotes the Euclidean vector norm.

A useful fact that is straightforward from Definition 2.1.4 is that the spectral norm of the matrix M equals its largest singular value, cf. [65, 81]. I.e. we have $\|M\|_2 = \sigma_{\max}(M)$. This can be used to define ℓ_p-analogues of the largest and smallest singular values of matrices, which intuitively quantify the maximum dilation resp. minimum contraction, that the matrix introduces to any non-zero vector in terms of its ℓ_p norm.

Definition 2.1.5 (ℓ_p singular values, cf. [35, 65])**.** *Let $p \in [1, \infty)$. We define for a matrix $M \in \mathbb{R}^{n \times d}$*

$$\sigma_{\max}^{(p)}(M) = \sup_{x \in \mathbb{R}^d \setminus \{0\}} \frac{\|Mx\|_p}{\|x\|_p} \quad \textit{and} \quad \sigma_{\min}^{(p)}(M) = \inf_{x \in \mathbb{R}^d \setminus \{0\}} \frac{\|Mx\|_p}{\|x\|_p}.$$

While norms introduce intuitively the notion of *length* to a space, metrics introduce the concept informally referred to as *distance.* We now define the notions of a *near-metric* and a *metric* space whose distance functions are required to satisfy some properties.

Definition 2.1.6 (near-metric, metric, cf. [56, 65])**.** *A set $\mathcal{X}$ equipped with a function $d : \mathcal{X}^2 \to \mathbb{R}$ is a* near-metric *space $(\mathcal{X}, m)$ if for every triple $a, b, c \in \mathcal{X}$ the following properties hold:*

1. $d(a, b) \geq 0$ *(non-negativity)*
2. $d(a, b) = d(b, a)$ *(symmetry)*
3. $d(a, c) \leq d(a, b) + d(b, c)$ *(triangle inequality).*

We call $(\mathcal{X}, d)$ a metric *space if additionally a fourth property holds:*

4. $d(a, b) = 0 \Leftrightarrow a = b$ *(identity of indiscernible elements).*

It is well-known that $\mathcal{X} = \mathbb{R}^d$ equipped with the distance function $d(a, b) = \|a - b\|_p$ induced by the ℓ_p-norm $\|\cdot\|_p$ is metric [65] and is thus also near-metric. We overload the notation to denote this p-normed and metric space by $(\mathbb{R}^d, \|\cdot\|_p)$.

2.2 Probability distributions

The normal distribution The normal distribution $\mathrm{N}\left(\mu, \sigma^2\right)$ for parameters $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}$ is a continuous probability distribution defined over the domain $\mathbb{R}$ and has the probability density function, cf. [88]

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|x - \mu|^2}{2\sigma^2}\right).$$

The expectation and variance of $X \sim \mathrm{N}\left(\mu, \sigma^2\right)$ are given as

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \mathbb{V}[X] = \sigma^2 .$$

We are going to discuss a straightforward extension to multivariate normal distributions over $\mathbb{R}^d$ in Section 2.3.

The p-generalized normal distribution Let $p \in [1, \infty)$. The p-generalized normal distribution, cf. [96, 132], $\mathrm{N_p}(\mu, \varsigma)$ for parameters $\mu \in \mathbb{R}, \varsigma \in \mathbb{R}_{>0}$ is a continuous probability distribution defined over the domain $\mathbb{R}$ and has the probability density function

$$f(x) = \frac{p}{2\varsigma\Gamma(1/p)} \exp\left(-\frac{|x-\mu|^p}{\varsigma^p}\right),$$

where Γ denotes the gamma function. The expectation and variance of $X \sim \mathrm{N_p}(\mu, \varsigma)$ are given as

$$\mathbb{E}[X] = \mu \quad \text{and} \quad \mathbb{V}[X] = \varsigma^2 \cdot \frac{\Gamma(3/p)}{\Gamma(1/p)} \,.$$

The Poisson distribution The Poisson distribution, cf. [90, 143], is a discrete probability distribution taking values in $\mathbb{N}_0$. Given a parameter $\lambda \in \mathbb{R}_{\geq 0}$, the probability of a random variable $X \sim \mathrm{Poi}(\lambda)$ taking the value $x \in \mathbb{N}_0$ is

$$\mathbf{Pr}\left[X = x\right] = \frac{\lambda^x}{x!} e^{-\lambda} \,.$$

The Poisson distribution has the property of equidispersion, i.e., $X \sim \mathrm{Poi}(\lambda)$ satisfies

$$\mathbb{E}[X] = \mathbb{V}[X] = \lambda \,.$$

The Bernoulli distribution The Bernoulli distribution $\mathrm{Bern}(\pi)$ takes binary values in $\{0, 1\}$ which is interpreted as indicator, whether an event happened or not, cf. [90]. It takes a single parameter $\pi \in [0, 1]$ which defines the probability of the event to happen. Let $X \sim \mathrm{Bern}(\pi)$, then

$$\mathbf{Pr}\left[X = x\right] = \pi^x (1-\pi)^{1-x} \quad , \text{for } x \in \{0, 1\}.$$

The expectation and variance of X are given as

$$\mathbb{E}[X] = 1 \cdot \mathbf{Pr}\left[X = 1\right] + 0 \cdot \mathbf{Pr}\left[X = 0\right] = \pi \quad \text{and} \quad \mathbb{V}[X] = \pi(1-\pi) \,.$$

Dirac's δ function The δ function was originally defined in [43] as

$$\delta \colon \mathbb{R} \to \{0, \infty\}$$

$$\delta(x) \mapsto \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases}$$
$$\text{such that } \int_{-\infty}^{\infty} \delta(x)\,\mathrm{d}x = 1. \tag{2.1}$$

The condition (2.1) quantifies the infinite value in the range of δ and also justifies an interpretation as a degenerate probability density function that concentrates all probability mass at zero and has zero density everywhere else. A random variable $X \sim \delta$ consequently attains a constant value $X = 0$ and thus satisfies $\mathbb{E}[X] = \mathbb{V}[X] = 0$.

Distance measures on probability distributions In order to quantify the distance between probability measures we will need some further definitions. Given two probability distributions γ, ν over $\mathbb{R}^d$, let $\Lambda(\gamma, \nu)$ denote the set of all joint probability distributions on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals γ and ν, respectively. More formally, $\lambda \in \Lambda(\gamma, \nu)$ needs to satisfy both of the following conditions, cf. [139]

$$\begin{aligned} &\int_{\mathbb{R}^d} \lambda(x, y)\,\mathrm{d}y = \gamma(x) \\ \text{and } &\int_{\mathbb{R}^d} \lambda(x, y)\,\mathrm{d}x = \nu(y). \end{aligned}$$

Since the domains of γ and ν are both equal, namely $\mathbb{R}^d$ in the present manuscript, we can define such a joint probability distribution via a bijection $g\colon \mathbb{R}^d \to \mathbb{R}^d$ using Dirac's δ function. To this end we let $\lambda(y|x) = \delta(y - g(x))$ for each fixed $x \in \mathbb{R}^d$. Since g is a bijection, it is invertible and we have at the same time $\lambda(x|y) = \delta(x - g^{-1}(y))$ for each fixed $y \in \mathbb{R}^d$. Thus

$$\begin{aligned} &\int_{\mathbb{R}^d} \lambda(x, y)\,\mathrm{d}y = \int_{\mathbb{R}^d} \lambda(y|x)\gamma(x)\,\mathrm{d}y = \gamma(x) \\ \text{and } &\int_{\mathbb{R}^d} \lambda(x, y)\,\mathrm{d}x = \int_{\mathbb{R}^d} \lambda(x|y)\nu(y)\,\mathrm{d}x = \nu(y) \end{aligned}$$

follows as desired. A careful choice of the bijection g, such that the points mapped to each other are close in ℓ_p distance, will be crucial in bounding the ℓ_p Wasserstein distance of the distributions under study.

Definition 2.2.1 (ℓ_p Wasserstein distance, [139])**.** *Given two probability measures γ, ν on $\mathbb{R}^d$ the ℓ_p Wasserstein distance between γ and ν is defined as*

$$\mathcal{W}_p(\gamma,\nu) = \left(\inf_{\lambda\in\Lambda(\gamma,\nu)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|_p^p \;\mathrm{d}\lambda(x,y)\right)^{\frac{1}{p}} = \inf_{\lambda\in\Lambda(\gamma,\nu)} \mathbb{E}_\lambda\left[\|x-y\|_p^p\right]^{\frac{1}{p}}$$

From the definition of the Wasserstein distance we can derive a measure of how much points drawn from a given distribution will spread from the origin. The *Wasserstein weight* can be thought of as a norm of a probability measure.

Definition 2.2.2 (Wasserstein weight)**.** *We define the ℓ_p Wasserstein weight of a probability measure γ as*

$$\mathcal{W}_p(\gamma) = \mathcal{W}_p(\gamma,\delta) = \left(\int_{\mathbb{R}^d} \|x\|_p^p \;\mathrm{d}\gamma\right)^{\frac{1}{2}} = \mathbb{E}_\gamma\left[\|x\|_p^p\right]^{\frac{1}{p}}$$

where δ denotes the Dirac delta function.

2.3 Linear regression

In linear regression we are given data $X \in \mathbb{R}^{n\times d}$ which carries n observations of the d independent variables and we want to find parameters $\beta \in \mathbb{R}^d$ that enable us to predict the dependent random variable $Y \in \mathbb{R}^n$ via a linear model. A linear regression model is given by

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

where ξ is a noise variable whose entries usually follow a normal distribution $\xi_i \sim \mathrm{N}\left(0,\varsigma^2\right)$ for $i \in [n]$ and model the unobservable error term, cf. [67]. Under the assumption of independent and identically distributed (i.i.d.) observations, we can multiply their individual distributions for any fixed β to obtain the distribution for the entire data. The dependent variable Y then follows a multivariate normal distribution, $Y \sim \mathrm{N}\left(X\beta, \varsigma^2 I_n\right)$. The corresponding probability density function is

$$f(Y|X\beta,\Sigma) = (2\pi)^{-\frac{n}{2}}\,|\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2\varsigma^2}\,\|X\beta - Y\|_2^2\right),$$

where $\Sigma = \varsigma^2 I_n$, cf. [21, 89]. A generalization to $\xi_i \sim \mathrm{N_p}(0, \varsigma)$ distributed according to p-generalized normal distribution is straightforward and yields to the following joint probability density function [66],

$$\begin{aligned} f(Y|X\beta, \Sigma) &= \prod_{i=1}^{n} \frac{p}{2\varsigma\Gamma(1/p)} \exp\left(-\frac{|x_i\beta - y_i|^p}{\varsigma^p}\right) \\ &= \left(\frac{p}{2\varsigma\Gamma(1/p)}\right)^n \exp\left(-\frac{1}{\varsigma^p} \|X\beta - Y\|_p^p\right). \end{aligned}$$

However, in regression we aim at learning from given data X, Y knowledge about the unknown parameter β. The above joint distributions of the data are thus interpreted as so called *likelihood* functions depending on β as a variable, given the fixed observed data,

$$\mathcal{L}(\beta|X, Y) = f(Y|X\beta, \Sigma)$$

which in general is not a probability density function for β since it might be unnormalized, i.e. $\int_{\mathbb{R}^d} \mathcal{L}(\beta|X, Y)\ \mathrm{d}\beta \neq 1$. However, by normalizing via this integral, we can assume that there exists a probability density function $q(\beta) \propto \mathcal{L}(\beta|X, Y)$ that is proportional to $\mathcal{L}(\beta|X, Y)$.

2.3.1 Maximum likelihood estimation

Given the model above, a maximum likelihood estimate of the parameter β can be obtained by minimizing the exponent of the likelihood function with respect to β. In particular, for linear ℓ_2 regression this involves solving the following optimization problem, which we call the lest squares regression problem.

Definition 2.3.1. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, the least squares regression problem is to find a parameter vector $\hat\beta \in \mathbb{R}^d$ that minimizes the least squares cost function, i.e.,*

$$\hat\beta \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|X\beta - Y\|_2^2 .$$

When we differentiate the objective with respect to β and set its gradient to zero we arrive at the *normal equation* for the least squares regression problem [21], cf. [67],

$$X^T(X\beta - Y) = 0. \tag{2.2}$$

Intuitively, the normal equation states that the residual vector of the optimal solution $X\hat\beta - Y$ is orthogonal to the columnspace spanned by X. Indeed, if X^TX is non-singular, the normal equation can be rearranged to obtain the unique maximum-likelihood estimator $\hat{\beta}$, via an orthogonal projection [21, 67]

$$\hat{\beta} = (X^T X)^{-1} X^T Y,$$

which can be solved in time $O(nd^2)$ via singular value decomposition [65].

We can similarly formulate the ℓ_p-regression problem which has no closed form solution for $p \neq 2$. However it can be interpreted as finding a maximum likelihood solution $\hat{\beta}$ when the noise variable ξ is distributed component-wise following a p-generalized normal distribution.

Definition 2.3.2. *Given $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^n$, the ℓ_p regression problem is to find a parameter vector $\hat{\beta} \in \mathbb{R}^d$ that minimizes the ℓ_p cost function, i.e.,*

$$\hat{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \left\| X\beta - Y \right\|_p .$$

2.3.2 Bayesian linear regression

Bayesian regression does not assume a fixed *optimal* solution for a data set as in the problems defined above, but introduces a distribution over the parameter space. The parameter vector $\beta \in \mathbb{R}^d$ is assumed to follow an unknown distribution $p_{\text{post}}(\beta|X, Y)$ called the posterior distribution. It is composed of the *likelihood* function $\mathcal{L}(\beta|X, \beta)$ which, as before, models the information that comes from the data. Additionally we assume a prior distribution $p_{\text{pre}}(\beta)$ which models problem-specific knowledge [21, 61].

Our goal is to explore the *posterior* distribution or its defining statistics, like moments, quantiles etc. As a consequence of Bayes Theorem, the posterior distribution is a compromise between the observed data and the prior knowledge imposed for the parameters, cf. [61].

$$p_{\text{post}}(\beta|X, Y) \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta).$$

Prior knowledge about β can be modeled in many cases as an uninformative distribution [60]. It should not constrain the solution space artificially or even unintentionally. Especially when no actual prior knowledge is available, it is convenient to consider the degenerate choice of a uniform distribution over $\mathbb{R}^d$ [61].

In general, the posterior distribution cannot be calculated analytically [61]. However, numerical algorithms may be applied to approximate the posterior distribution. Examples of modern algorithms include Markov Chain Monte Carlo sampling algorithms [59], integrated nested Laplace approximation [127], and approximate Bayesian computation [39]. We do not go into details about these methods, since the results of this manuscript do not depend on the algorithm employed for the inference. However, we refer to [61] for an extensive overview of available approaches.

2.4 Generalized linear regression models

Generalized linear models extend classical linear regression to more expressive classes of generating distributions [108]. Usually one assumes that the realizations of the dependent variable are generated from a parametrized family of distributions, based on the independent observations. Well-known examples of such distributions include the multivariate normal, Bernoulli, Poisson, and gamma distributions. The expected value of the dependent variable Y is connected to the linear term $X\beta$ via a so-called link function h,

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

For most generalized linear models, no closed solution is known. We present two examples of such generalized linear models.

2.4.1 Poisson regression

The main purpose of the Poisson regression model is to learn to predict count data based on independent real-valued observations. A comprehensible example is the task of predicting the number of rented bicycles based on calendrical or meteorological data on a daily basis. [51, 62]

The natural link function for Poisson regression is the natural logarithm function [108, 143], i.e.,

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

The model assumption is that each random variable Y_i for $i \in [n]$ with outcome y_i is distributed according to a Poisson distribution with mean λ_i that is the exponential of the linear term of the observation x_i and the unknown parameter β.

$$Y_i \sim \text{Poi}(\lambda_i),\ \lambda_i = \exp(x_i\beta) \tag{2.3}$$

The task of finding a maximum likelihood estimator $\hat{\beta}$ for the parameters amounts to solving the following optimization problem for β, cf. [108, 143].

Definition 2.4.1. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{N}_0^n$, the Poisson regression problem is to find a parameter vector $\hat{\beta} \in \mathbb{R}^d$ that minimizes the negative log-likelihood function, that we call the Poisson regression cost function. I.e. to find*

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \textstyle\sum_{i=1}^{n} \exp(x_i\beta) - y_i \cdot x_i\beta + \ln(y_i!)$$

2.4.2 Logistic regression

The main purpose of the logistic regression model is to learn to predict the probability of an event to happen based on independent observations. As an example consider the task of predicting the probability of a patient of suffering from a disease based on the patients physiological and diagnostic data.

The natural link function for logistic regression is the so called logit function [80, 108]. For a binary variable Y, it maps the probabilities $\pi = \mathbb{E}[Y]$ bijectively from the $[0, 1]$ interval to the reals, connecting them to the linear predictor, i.e.,

$$\operatorname{logit}(\mathbb{E}(Y)) = \ln\left(\frac{\pi}{1-\pi}\right) = X\beta, \tag{2.4}$$

where the term in the logarithm, $\frac{\pi}{1-\pi}$, is called the *odds* of the binary variable Y. The model assumption is that each random variable Y_i for $i \in [n]$ with outcome y_i is distributed according to a Bernoulli distribution with parameter π_i that is linked to the observation x_i and the unknown parameter β via the logit link function.

$$Y_i \sim \operatorname{Bern}(\pi_i),\ \pi_i = \frac{\exp(x_i\beta)}{1+\exp(x_i\beta)} \tag{2.5}$$

A maximum likelihood estimate $\hat{\beta}$ can be obtained by minimizing the negative log-likelihood with respect to β, which amounts to solving the following optimization problem, cf. [80, 108]. For technical and notational reasons it is more convenient to map the target variables $y_i \in \{0, 1\}$ to $y_i \in \{-1, +1\}$

Definition 2.4.2. *Given $X \in \mathbb{R}^{n\times d}, Y \in \{-1, +1\}^n$, the logistic regression problem is to find a parameter vector $\hat{\beta} \in \mathbb{R}^d$ that minimizes the negative log-likelihood function, that we call the logistic regression cost function. I.e. to find*

$$\hat{\beta} \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \textstyle\sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i\beta)).$$

2.4.3 Dependency Networks

Dependency Networks were introduced by Heckerman et al. (2000) [77]. They are probabilistic graph models comprising a collection of generalized linear models, where each element of a set of d variables is regressed on all other variables. Other graph models, like Markov random fields and Bayesian networks, are limited to undirected or acyclic structures. We refer to [97] for a general introduction and broad overview on probabilistic graph models. Dependency networks, however, allow for directed as well as cyclic structures, combining the benefits of previous models. We give a formal definition.

Definition 2.4.3 (cf. [77])**.** *Let $X = (X^{(1)}, \dots, X^{(d)})$ denote a random vector and $x = (x^{(1)}, \dots, x^{(d)})$ its instantiation. A* Dependency Network *on X is a pair $(\mathcal{G}, \Psi)$ where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a directed, possibly cyclic graph, where each vertex $i \in \mathcal{V} = [d]$ corresponds to the random variable $X^{(i)}$. These are connected via a set of directed edges $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V} \setminus \{(i,i) \,|\, i \in [d]\}$. We refer to the vertices that have an edge pointing to $X^{(i)}$ as its parents, denoted by $\mathbf{pa}_i = \{X^{(j)} \,|\, (j,i) \in \mathcal{E}\}$. $\Psi = \{p_i \,|\, i \in [d]\}$ is a set of conditional probability distributions associated with each variable such that $X^{(i)} \sim p_i$ and where*

$$p_i = p(x^{(i)} |\, \mathbf{pa}_i) = p(x^{(i)} |\, x^{\setminus i}) \ .$$

Finally, the full joint distribution is simply defined as the product of conditional distributions

$$p(x) = \prod\nolimits_{i \in [d]} p(x^{(i)} |\, x^{\setminus i}) \ ,$$

that we call pseudo likelihood, cf. [20].

Each edge in a dependency network models a dependency between variables. Note in particular that if there is no edge between i and j then the variables $X^{(i)}$ and $X^{(j)}$ are conditionally independent given the other variables $X^{\setminus i,j}$ indexed by $[d] \setminus \{i,j\}$ in the dependency network.

Note, that defining the pseudo likelihood as the product of the conditional distributions neglects the fact that in general the distributions are not independent and indeed, the existence of a consistent joint distribution of which they are the conditionals is not guaranteed [18, 19, 20].

Learning dependency networks amounts to determining the conditional probability distributions from a given set of n training instances over d variables. Assuming that $p(x^{(i)} |\, \mathbf{pa}_i)$ is parameterized as a generalized linear model, this amounts to estimating the parameters $\beta^{(i)}$ of the generalized linear model associated with each variable $X^{(i)}$, since then the local distributions are completely determined by the parameters $\beta^{(i)} \in \mathbb{R}^{(d-1)}$. However, $p(x^{(i)} | \mathbf{pa}_i)$ will possibly depend on all other variables in the network, and these dependencies define the edges of its graph structure.

Dependency networks have several interesting applications, like collaborative filtering [77], phylogenetic analysis [27], genetic analysis [121], network inference from sequencing data [7], and traffic [70] as well as topic modeling [71]. However, we do not go into details about possible applications.

2.5 Data reduction methods

Here we finally introduce the data reduction methods, that we are going to use to make classic algorithms for our problems scalable to massive data sets following the general scheme of the *sketch and solve* paradigm [145] that stems from the theory of streaming algorithms: first reduce the massive data set using one of the methods presented here and then solve the problem on the reduced data using a classical or only slightly modified algorithm. The general approach is given in the following scheme, where Π is a problem specific map, i.e., an algorithm that maps the large data set $X \in \mathbb{R}^{n\times d}$ to a significantly smaller data set $\Pi(X) \in \mathbb{R}^{k\times d}$ where $k \ll n$. By $f(\beta \mid X)$ we denote some function $f : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ that depends on the data and that we want to approximate over all $\beta \in \mathbb{R}^d$ or special choices of $\beta \in \mathbb{R}^d$ depending on an approximation parameter ε which is indicated by the symbol $\approx_\varepsilon$.

$$\begin{tikzcd} X \arrow[r, "\Pi"] \arrow[d] & \Pi(X) \arrow[d] \\ f(\beta \mid X) \arrow[r, phantom, "\approx_\varepsilon"] & f(\beta \mid \Pi(X)). \end{tikzcd} \tag{2.6}$$

The function f might be thought of as a probability density function or an objective function of an optimization problem. Reducing the data first is faster or at least more space efficient than performing the statistical analysis or optimization directly on the original data. If this data is really massive in volume, the direct approach might not be possible at all due to resource restrictions of the computing device. We now present different methods to realize the data reduction map Π.

2.5.1 Coresets

Coresets are arguably one of the first formalized approach for such data reduction methods. The concept of coresets is related to the early works [48, 49, 107] on *data-squashing*. To our knowledge the term *coreset* was introduced in [15] in the context of shape fitting and clustering problems. Some of their coresets were so-called *weak* coresets, that only guarantee a $(1+\varepsilon)$-approximation in the optimum of the problem on the reduced data set. But since its size is very little, sometimes even independent of the number of input points, an exhaustive search can often efficiently solve the problem. However, in this manuscript we focus on *strong* coresets and omit the distinction in the remainder unless stated differently. A strong coreset is a (possibly) weighted and considerably smaller data set that serves as a proxy for the original data and approximates the given objective function for all candidate solutions, cf [122].

Definition 2.5.1 (strong $(1\pm\varepsilon)$-coreset)**.** *Let X be a set of points of size n from a universe A and let B be a set of candidate solutions. Let $f : 2^A \times B \to \mathbb{R}_{\geq 0}$ be a non-negative objective function. Then a set C of size $k \ll n$ is a strong $(1\pm\varepsilon)$-coreset of X for f, if*

$$\forall \beta \in B : |f(X,\beta) - f(C,\beta)| \leq \varepsilon \cdot f(X,\beta).$$

Not distinguishing between weak or strong, we can give a brief and incomplete list of results. Small coresets have been developed, e.g., for shape fitting problems [3, 4, 13, 14], clustering [15, 54, 55, 105], classification [73, 75, 124], ℓ_2-regression [44, 45], ℓ_1-regression [30, 35, 131], ℓ_p-regression [40, 146], M-estimators [33, 34] and generalized linear models [84, 124, 135]. See [122] for a recent and extensive survey on coreset results and [114] for a technical introduction to coreset construction methods.

Before we introduce the so-called sensitivity framework of [53, 103] as a very general method for obtaining coresets in Section 2.5.3, we first focus on a related data reduction tool that is closer to the needs of problems in the area of randomized linear algebra.

2.5.2 Subspace embeddings

The following definition of so called ε-subspace embeddings is central to parts of this manuscript. Such an embedding can be used to reduce the size of a given data matrix while preserving the entire structure of its spanned subspace up to $(1\pm\varepsilon)$ distortion. Subspace embeddings were introduced in [129]. Before we summarize several methods to construct a subspace embedding for a given input matrix, we give a formal definition and present some results that follow from this definition. An ε-subspace embedding can be considered as a coreset construction for the squared Euclidean norm of vectors in the columnspace of an input matrix $X = U\Sigma V^T$ which is completely determined by U derived from its singular value decomposition.

Definition 2.5.2 (ε-subspace embedding, cf. [129, 144])**.** *Given a matrix $U \in \mathbb{R}^{n\times d}$ with orthonormal columns, an integer $k \leq n$ and an approximation parameter $0 < \varepsilon \leq 1/2$, an ε-subspace embedding for U is a map $\Pi : \mathbb{R}^n \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)\, \|Ux\|_2^2 \leq \|\Pi U x\|_2^2 \leq (1+\varepsilon)\, \|Ux\|_2^2 \tag{2.7}$$

holds for all $x \in \mathbb{R}^d$, or, equivalently (cf. [36, 120])

$$\left\|U^T \Pi^T \Pi U - I_d\right\|_2 \leq \varepsilon. \tag{2.8}$$

Inequality (2.7) makes clear that the norms of all vectors in the embedded subspace are preserved up to little distortion, while the equivalent inequality (2.8) makes clear that the embedded subspace remains close to an orthonormal basis, not introducing much scale or rotation. As a consequence an ε-subspace embedding Π for the columnspace of a matrix M preserves its squared singular values up to little distortion, which in particular means that it also preserves its rank.

Observation 1. *Let Π be an ε-subspace embedding for the columnspace of $M \in \mathbb{R}^{n\times d}$. Then*

$$(1-\varepsilon)\, \sigma_i^2(M) \leq \sigma_i^2(\Pi M) \leq (1+\varepsilon)\, \sigma_i^2(M)$$

and

$$(1-2\varepsilon)\, \sigma_i^{-2}(M) \leq \sigma_i^{-2}(\Pi M) \leq (1+2\varepsilon)\, \sigma_i^{-2}(M).$$

Proof. For the first claim, we make use of a min-max representation of the singular values that is known as the Courant-Fischer theorem, cf. [81]. In the following derivation we choose x^* to be the maximizer of (2.9) and S^* the minimizer of (2.10).

$$\begin{aligned}
\sigma_i^2(\Pi M) &= \min_{S\in\mathbb{R}^{(i-1)\times d}} \max_{Sx=0, \|x\|_2=1} \|\Pi M x\|_2^2 \\
&\leq \max_{S^*x=0, \|x\|_2=1} \|\Pi M x\|_2^2 \qquad (2.9)\\
&= \|\Pi M x^*\|_2^2 \\
&\leq (1+\varepsilon)\ \|Mx^*\|_2^2 \\
&\leq (1+\varepsilon) \max_{S^*x=0, \|x\|_2=1} \|Mx\|_2^2 \\
&= (1+\varepsilon) \min_{S\in\mathbb{R}^{(i-1)\times d}} \max_{Sx=0, \|x\|_2=1} \|Mx\|_2^2 \qquad (2.10)\\
&= (1+\varepsilon)\, \sigma_i^2(M).
\end{aligned}$$

The lower bound can be derived analogously using the lower bound of (2.7). The second claim follows from the first.

$$\begin{aligned}\left|\frac{1}{\sigma_i^2(M)}-\frac{1}{\sigma_i^2(\Pi M)}\right| &= \frac{|\sigma_i^2(M)-\sigma_i^2(\Pi M)|}{\sigma_i^2(M)\sigma_i^2(\Pi M)} \leq \frac{\varepsilon\sigma_i^2(M)}{(1-\varepsilon)\,\sigma_i^4(M)} \\ &= \frac{\varepsilon}{(1-\varepsilon)}\,\sigma_i^{-2}(M) \leq 2\varepsilon\,\sigma_i^{-2}(M). \qquad \square\end{aligned}$$

Another useful fact that we are going to exploit is that an ε-subspace embedding implies that it also approximates matrix multiplication with respect to the spectral norm. We formalize this fact in the following lemma which is taken from [36].

Lemma 2.5.3 ([36]). *Let $C = [A, B]$, and let Π be an ε-subspace embedding for the columnspace of C. Then it holds that*

$$\left\|A^T\Pi^T\Pi B - A^T B\right\|_2 \leq \varepsilon \left\|A\right\|_2 \left\|B\right\|_2. \tag{2.11}$$

Constructions based on random projections There are several ways to construct an ε-subspace embedding. One of the more recent methods is using a so called *graph-sparsifier*, which was initially introduced for the efficient construction of sparse sub-graphs with good expansion properties [17]. A follow-up work [24] adapted the technique to ℓ_2-problems like ordinary least squares regression. While the initial construction was deterministic, they also gave alternative constructions combining the deterministic decision rules with non-uniform random sampling techniques.

In principle our approximation results are independent of the actual method used to calculate the embedding as long as the property given in Definition 2.5.2 is satisfied. However, we want to tackle really massive data or deal with a data stream, in which case it can only be read once due to given time and space constraints. In order to construct ε-subspace embeddings in a single pass over the data, we first consider the approach of so called *oblivious* subspace embeddings. These can be viewed as distributions over appropriately structured $k \times n$ matrices from which we can draw a realization Π independent of the input data. It is then guaranteed that for any fixed matrix U as in Definition 2.5.2 and failure probability $0 < \eta \leq 1/2$, the realization Π is an ε-subspace embedding with probability at least $1 - \eta$. We survey the following approaches for obtaining oblivious ε-subspace embeddings. The first two approaches are modern adaptions of the seminal Johnson-Lindenstrauss embeddings [91], see also [1, 6]. The third method that we present is based on the so called *count-sketch* which was introduced in the context of approximation algorithms for finding frequent items in data streams [29]. It was shown only significantly later in [32] to be useful for generating ε-subspace embeddings that work in *input sparsity time*, i.e., the running time of their application to X is $O(\mathrm{nnz}(X))$ where $\mathrm{nnz}(X)$ denotes the number of non-zero elements of X and can be considerably smaller than nd.

The Rademacher Matrix Π is obtained by choosing each entry independently from $\{-1,1\}$ with equal probability. The matrix is then rescaled by $\frac{1}{\sqrt{k}}$. This method has been shown in [129] to form an ε-subspace embedding with probability at least $1-\eta$ when choosing essentially $k \in O\left(\frac{d\log(d/\eta)}{\varepsilon^2}\right)$. This was later improved to $k \in O\left(\frac{d+\log(1/\eta)}{\varepsilon^2}\right)$ in [31], which was recently shown to be optimal in [119]. While this method yields the best reduction among the different constructions that we consider in this manuscript, the Rademacher matrix has the disadvantage that we need $\Theta(ndk)$ time to apply it to an $n \times d$ matrix using standard multiplication when reading the input in general. If the input is given row by row or at least block by block, a fast matrix multiplication algorithm can be applied block wise. We remark that it is provably sufficient that the $\{-1,1\}$-entries in each row of the Rademacher matrix are basically four wise independent, i.e., when considering up to four entries of the same row, these behave as if they were fully independent. Such random numbers can be generated using a hashing scheme that generates Bose-Chaudhuri-Hocquenghem (BCH) codes using a seed of size $O(\log n)$. This has first been noticed in a seminal work by Alon et al. [9] based on a technique from [8]. An overview over different generating methodsis given in Rusu and Dobra [128].

The Subsampled Randomized Hadamard Transform This embedding method is originally from [6]. It is chosen to be $\Pi = RH_mD$ where D is an $m \times m$ diagonal matrix where each entry is independently chosen from $\{-1,1\}$ with equal probability. The value of m is assumed to be a power of two. Moreover, it is convenient to choose the smallest integer that satisfies $m \geq n$. H_m is the *Walsh-Hadamard-matrix* of order m, which is recursively defined, cf. [6], as

$$H_1 = [1], \quad H_m = \begin{bmatrix} H_{m/2} & H_{m/2} \\ H_{m/2} & -H_{m/2} \end{bmatrix}.$$

Finally, R is a $k \times m$ row sampling matrix. That is, each row of R contains exactly one 1-entry and is 0 everywhere else. The index of the 1-entry is chosen uniformly from $[m]$ i.i.d. for every row. The matrix is then rescaled by $\frac{1}{\sqrt{k}}$. Since m is often larger than n, the input data must be padded with 0-entries to compute the product ΠX. Of course, it is not necessary to do this explicitly since all multiplications by zero can be omitted. The target dimension needed to form an ε-subspace embedding with probability at least $1-\eta$ using this family of matrices was shown in [36] to be $k \in O\left(\frac{(d+\log(1/(\varepsilon\eta)))\log(d/\eta)}{\varepsilon^2}\right)$, which improved upon previous results from [23, 46, 136]. Note that this is close to the known lower bound of $\Omega(d \log d)$ given in [72, 136]. Compared to the Rademacher matrix, the dependency on the dimension d is worse by essentially a factor of $O(\log d)$. The benefit that we get is that due to the inductive structure of the Walsh-Hadamard matrix, the embedding can be applied in roughly $O(nd \log k) \subseteq O(nd \log d)$ time omitting constant factors, which is considerably faster. It has been noticed in the original paper [6] that the construction is closely related to four wise independent BCH codes. To our knowledge, there is no explicit proof that it is sufficient to use random bits of little independence.

The Clarkson Woodruff embedding This is the most recent construction that we consider here. As noted before, it is actually well-known as the count-sketch [29]. But it was shown by Clarkson and Woodruff in [32] that it can be used to construct subspace embeddings. The embedding is obtained as $\Pi = \Phi D$. Each entry of the diagonal matrix D is chosen as $D_{ii} \in \{-1, 1\}$ with equal probability. Given a random hash map $h : [n] \to [k]$ such that the image of every $i \in [n]$ is $h(i) = t \in [k]$ with probability $\frac{1}{k}$, again Φ is a binary matrix whose 1-entries can be defined by $\Phi_{h(i),i} = 1$. All other entries are 0. This is obviously the fastest embedding, due to its sparse construction. It can be applied to any matrix $X \in \mathbb{R}^{n \times d}$ in $O(\text{nnz}(X)) \subseteq O(nd)$ time, where $\text{nnz}(X)$ denotes the number of non-zero entries in X. This is referred to as *input sparsity time* and is clearly optimal up to small constants, since this is the time needed to actually read the input from a data stream or external memory. However, its disadvantage is that the target dimension is $k \in \Omega(d^2)$ [118]. Roughly spoken, this is necessary due to the need to obliviously and perfectly hash d of the standard basis vectors spanning $\mathbb{R}^n$. Upper bounds given in [117] improved over the original ones [32] and showed that $k \in O\left(\frac{d^2}{\varepsilon^2\eta}\right)$ is sufficient to draw an ε-subspace embedding from this distribution of matrices with probability at least $1-\eta$. This reference [117] also shows that it is sufficient to use only four wise independent random bits to generate the diagonal matrix D. Again, the four wise independent BCH scheme from [8, 9, 128] can be used. Moreover, Φ can be constructed using only pairwise independent entries. This can be achieved very efficiently using the fast universal hashing scheme introduced in [42]. The space requirement is only $O(\log n)$ for a hash function from this class. For a really fast implementation using bit-wise operations, the actual size parameters of the embedding can be chosen to be the smallest powers of two that are larger than the required size parameters n and k.

Note, in particular the trade-off behavior between time and space complexity of the presented embedding methods. While usually one is interested in the fastest possible application time, memory constraints might make it impossible to apply the Clarkson and Woodruff embedding due to its quadratic dependency on d. Taking it the other way, for a fixed embedding size k, this method will give the weakest approximation guarantee, cf. [148]. For really large d, even the $O(d \log d)$ factor of the subsampled randomized Hadamard transform might be too large so that we have to rely on the slowest Rademacher embedding method.

Extensions to streaming and distributed environments The presented reduction techniques are already useful whenever we deal with medium to large sized data for reducing time and space requirements. However, when the data grows massive, we need to put more care on computational requirements. We therefore want to briefly discuss and give references to some of these technical details. For example, the dimensions of the resulting embeddings do not depend on n. However, this is not true for the embedding matrices $\Pi \in \mathbb{R}^{k \times n}$. However, the embedding matrices presented above can be stored implicitly by using the different hash functions of limited independence. The hash functions that are suitable to implement the embedding methods are the four wise independent BCH scheme used in the seminal works [8, 9] and the universal hashing scheme from [42]. These can be evaluated very efficiently using bit-wise operations and can be stored using a seed whose size is only $O(\log n)$. Note that this small dependency on n is only needed in the embedding phase. After the embedding has been applied to the data, the space requirements for further computations will be independent of n. A survey and evaluation of alternative hashing schemes including BCH can be found in [128].

The linearity of the embeddings allows for efficient application in sequential streaming and in distributed environments, see e.g. [31, 93, 146]. The sketches can be updated in the most flexible dynamic setting, which is commonly referred to as the *turnstile* model, see [116]. In this model, think of an initial matrix of all zero values. The stream consists of updates of the form (i, j, u) meaning that the entry X_{ij} will be updated to $X_{ij} + u$. A single entry can be defined by one single update or by a sequence of not necessarily consecutive updates. For example a stream $S = \{\dots, (i, j, +5), \dots, (i, j, -3), \dots\}$ will result in $X_{ij} = 2$. Even deletions are possible in this setting by using negative updates. At first sight this model might seem very technical and unnatural. But the usual form of storing data in a table is not appropriate or performant for massive data sets. The data is rather stored as a sequence of *(key, value)* pairs in arbitrary order [63, 130]. For dealing with such unstructured data, the design of algorithms working in the turnstile model is of high importance.

For distributed computations, note that the embedding matrices can be communicated efficiently to every machine in a computing cluster of l machines. This is due to the small implicit representation by hash functions. Now, suppose the data is given as $X = \sum_{i=1}^{l} X_{(i)}$ where $X_{(i)}$ is the part of X stored on the machine with index $i \in [l]$. Note that by the above data representation in form of updates, $X_{(i)}$ can consist of rows, columns or single entries of X. Again, multiple updates to the same entry are possible and may be distributed to different machines. Every machine $i \in [l]$ can compute a small embedding on its own part of the data $X_{(i)}$ and efficiently communicate it to one dedicated central server. An embedding of the entire data set can be obtained by summing up the single embeddings since $\Pi X = \sum_{i=1}^{l} \Pi X_{(i)}$. For more details, the reader is referred to [93, 104, 146].

The above discussions make clear that these methods suit the criteria – identified in [142] – that need to be satisfied when dealing with massive data. Specifically, the number of data items that need to be accessed at a time is only a small subset of the whole data set, particularly independent of the total number of observations n. The algorithms should work on data streams. Moreover, the algorithms should be amenable to distributed computing environments like MapReduce [41].

ℓ_2-Sampling Another approach is subspace preserving subsampling of rows from the data matrix. The general technique is due to [58] in the context of low-rank approximation. It was further developed and transferred to approximating ℓ_2 regression in [44] and improved in [45]. Finally, it was generalized to more general subspace sampling for the ℓ_p-spaces in [40]. We will deal with this generalization later and now focus on ℓ_2.

Technically, an importance sampling scheme is performed proportional to the so called *statistical leverage scores*. The roots of these importance measures for linear regression can be traced back to [37]. The sampling based methods are in principle applicable whenever it is possible to read the input multiple times. For instance, one needs two passes over the data to perform the subspace preserving sampling procedure; one for preprocessing the input matrix and another for computing the probabilities and for the actual sampling [47]. The latter reference showed how to obtain good constant approximations to the leverage scores faster than the time needed to solve the linear regression problem exactly. Previously this was an important open problem for a long time, posed in [44].

We can construct a sampling and reweighting matrix Π which forms an ε-subspace embedding with constant probability in the following way [45]. Let U be any orthonormal basis for the columnspace of X. This basis can for example be obtained from the singular value decomposition $X = U\Sigma V^T$ of the data matrix. Now we define the *leverage scores* for the columnspace of X.

$$l_i = \frac{\|U_i\|_2^2}{\|U\|_F^2} = \frac{\|U_i\|_2^2}{d}, \text{ for } i \in [n].$$

Now we fix a sampling size parameter $k \in O(d \log(d/\varepsilon)/\varepsilon^2)$ and sample the input points one-by-one with probability $q_i = \min\{1, k \cdot l_i\}$. We reweight their contribution to the cost function by $w_i = \frac{1}{q_i}$. For the least squares cost function, this corresponds to defining a diagonal (sampling and reweighting) matrix Π by $\Pi_{ii} = \frac{1}{\sqrt{q_i}}$ with probability q_i and $\Pi_{ii} = 0$ otherwise. Also note, that the expected number of samples is $k \in O(d \log(d/\varepsilon)/\varepsilon^2)$, which also holds with constant probability by Markov's inequality. Moreover, to give an intuition why this works, note that for any fixed $\beta \in \mathbb{R}^d$, we have

$$\mathbb{E}\left[\|SX\beta\|_2^2\right] = \sum\nolimits_{i=1}^n \left(\frac{x_i\beta}{\sqrt{q_i}}\right)^2 q_i = \sum\nolimits_{i=1}^n (x_i\beta)^2 = \|X\beta\|_2^2 .$$

Note, that for some non-linear cost functions, the weights have to be stored separately for a similar result. The significantly stronger property of forming an ε-subspace embedding, according to Definition 2.5.2, follows from a matrix approximation bound given in [45, 126]. The result that we need is only implicitly given in the latter reference. We thus give a self-contained proof.

Lemma 2.5.4. *Let $X \in \mathbb{R}^{n\times d}$ be an input matrix. Let Π be a sampling matrix constructed as stated above with sampling size parameter $k \in O(d \log(d/\varepsilon)/\varepsilon^2)$. Then Π forms an ε-subspace embedding for the columnspace of X with constant probability.*

Proof. Let $X = U\Sigma V^T$ be the singular value decomposition of X. By Theorem 7 in [45] there exists an absolute constant $C > 1$ such that

$$\mathbb{E}\left[\left\|U^TS^TSU - U^TU\right\|_2\right] \leq C\sqrt{\frac{\log k}{k}}\,\|U\|_F\,\|U\|_2 \leq C\sqrt{\frac{\log k}{k}}\sqrt{d} \;\leq\; \varepsilon,$$

where we used the fact that $\|U\|_F = \sqrt{d}$ and $\|U\|_2 = 1$ by orthonormality of U. The last inequality holds by choice of $k = Dd \log(d/\varepsilon)/\varepsilon^2$ for a sufficiently large absolute constant $D > 1$ such that $\frac{1+\log D}{D} < \frac{1}{4C^2}$, since

$$\frac{\log k}{k} = \frac{\log(Dd \log(d/\varepsilon)/\varepsilon^2)}{Dd \log(d/\varepsilon)/\varepsilon^2} \leq \frac{2\varepsilon^2 \log(Dd \log(d/\varepsilon)/\varepsilon)}{Dd \log(d/\varepsilon)}$$

$$\leq \frac{4\varepsilon^2(\log(d/\varepsilon) + \log D)}{Dd\log(d/\varepsilon)} \leq \frac{4\varepsilon^2}{d}\left(\frac{1+\log D}{D}\right) < \frac{\varepsilon^2}{C^2 d}\,.$$

By an application of Markov's inequality and rescaling ε, we can assume with constant probability

$$\left\|U^T\Pi^T\Pi U - U^T U\right\|_2 \leq \varepsilon. \qquad \square$$

The failure probability can be reduced to arbitrary failure probability $0 < \eta \leq \frac{1}{2}$ via standard probability amplification, as noted in [45]. Another probability amplification scheme that applies to arbitrary ε-subspace embeddings is given in [104, 144]. It thus applies also to the sparse embedding by Clarkson and Woodruff [32] to reduce the linear dependency on η. The overhead incurred by both methods is a factor of $O(\log\frac{1}{\eta})$ independent repetitions.

The question arises whether we can do better than $O(d\log(d/\varepsilon)/\varepsilon^2)$ in the case of sampling via leverage scores. One can show by reduction from the coupon collectors theorem that there is a lower bound of $\Omega(d\log d)$ [136] matching the upper bound up to its dependency on ε. The hard instance is a $d^m \times d, m \in \mathbb{N}$ orthonormal matrix in which the scaled canonical basis $I_d/\sqrt{d^{m-1}}$ is stacked d^{m-1} times. The leverage scores are all equal to $1/d^m$, implying a uniform sampling distribution with probability $1/d$ for each basis vector. Any rank d preserving sample must comprise at least one of them. This is exactly the coupon collectors theorem with d coupons which has a lower bound of $\Omega(d\log d)$ [113]. The fact that the sampling is without replacement does not change this, since the reduction holds for arbitrary large m creating sufficient multiple copies of each element to simulate the sampling with replacement, see [136] for details.

Extensions to ℓ_p-spaces The definition of ε-subspace embeddings can be naturally extended to ℓ_p spaces for all $p \in [1,\infty)$, which was done implicitly in [40].

Definition 2.5.5 ((ε, p)-subspace embedding, cf. [40])**.** *Given a matrix $M \in \mathbb{R}^{n\times d}$, an integer $k \leq n$ and an approximation parameter $0 < \varepsilon \leq 1/2$, an (ε, p)-subspace embedding for the columnspace of M is a map $\Pi : \mathbb{R}^n \to \mathbb{R}^k$ such that*

$$(1-\varepsilon)\left\|Mx\right\|_p \leq \left\|\Pi Mx\right\|_p \leq (1+\varepsilon)\left\|Mx\right\|_p \tag{2.12}$$

holds for all $x \in \mathbb{R}^d$.

Consider the special case of an $(\varepsilon, 2)$-subspace embedding, which coincides with our previous definition of an ε-subspace embedding up to a small constant factor.

Observation 2. *Let $M \in \mathbb{R}^{n \times d}$. If Π is an ε-subspace embedding for M then Π is an $(\varepsilon, 2)$-subspace embedding for M. If Π is an $(\varepsilon, 2)$-subspace embedding for M then Π is a 3ε-subspace embedding for M.*

Proof. The first direction follows from the fact that

$$\|\Pi Mx\|_2^2 \leq (1+\varepsilon)\, \|\Pi Mx\|_2^2 \leq (1+\varepsilon)^2\, \|\Pi Mx\|_2^2$$

and

$$\|\Pi Mx\|_2^2 \geq (1-\varepsilon)\, \|\Pi Mx\|_2^2 \geq (1-\varepsilon)^2\, \|\Pi Mx\|_2^2$$

by taking the square root of every single term in both inequalities.

The reverse relationship follows similarly since $(1+\varepsilon)^2 = 1 + 2\varepsilon + \varepsilon^2 \leq 1 + 3\varepsilon$ and $(1-\varepsilon)^2 = 1 - 2\varepsilon + \varepsilon^2 \geq 1 - 2\varepsilon \geq 1 - 3\varepsilon$. □

Since the construction methods have only a polynomial dependency on $\frac{1}{\varepsilon}$, folding the constant into ε does not change the asymptotic complexities. We can thus treat both definitions as equivalent.

We will need at a later point another lemma which states that padding an embedded matrix with another matrix yields an embedding for the padded original matrix for the same parameters (ε, p).

Lemma 2.5.6. *Let $M = [M_1^T, M_2^T]^T \in \mathbb{R}^{(n_1+n_2) \times d}$ be an arbitrary matrix, and $p \in [1, \infty)$. Suppose Π is an (ε, p)-subspace embedding for the columnspace of M_1. Let $I_{n_2} \in \mathbb{R}^{(n_2 \times n_2)}$ be the identity matrix. Then*

$$P = \begin{bmatrix} \Pi & 0 \\ 0 & I_{n_2} \end{bmatrix} \in \mathbb{R}^{(k+n_2) \times (n_1+n_2)}$$

is an (ε, p)-subspace embedding for the columnspace of M.

Proof. Fix an arbitrary $x \in \mathbb{R}^d$. We have

$$\begin{aligned}
(1-\varepsilon)^p\, \|Mx\|_p^p &= (1-\varepsilon)^p(\|M_1x\|_p^p + \|M_2x\|_p^p) \\
&\leq (1-\varepsilon)^p\, \|M_1x\|_p^p + \|M_2x\|_p^p \\
&\leq \|\Pi M_1x\|_p^p + \|M_2x\|_p^p \\
&= \|PMx\|_p^p \\
&= \|\Pi M_1x\|_p^p + \|M_2x\|_p^p \\
&\leq (1+\varepsilon)^p\, \|M_1x\|_p^p + \|M_2x\|_p^p
\end{aligned}$$

$$\leq (1+\varepsilon)^p(\|M_1x\|_p^p + \|M_2x\|_p^p) = (1+\varepsilon)^p \, \|Mx\|_p^p$$

which concludes the proof by taking the pth root. □

Random projections for ℓ_p We can construct an (ε, p)-subspace embedding similar to the ℓ_2 case. However, there are several complications. In principle, we want to use oblivious ℓ_p embeddings via random projections but we cannot rely only on this technique, as we will discuss now.

There exist oblivious subspace embeddings via random projections for ℓ_1 [35, 131]. This can be generalized by exploiting the concept of *p-stable distributions* [110, 131] which exist only for $p \in [1, 2]$ [85]. This settled the problem of designing subspace embeddings limited to the cases $p \in [1, 2]$. To overcome this limitation, Andoni introduced the notion of *max-stability* of reciprocal exponential random variables [10] which led to a general construction working for all $p \in [1, \infty)$ in [141, 146]. The main issue is that the subspace approximation guarantee on the distortion of these embeddings is not bounded by $(1 \pm \varepsilon)$ any more. Instead, we only have weak bounds in the order of $O((d \log d)^{\frac{1}{p}})$ for dilation and contraction. These have recently been proven to be tight up to small $(\log d)^{O(1/p)}$ factors [141] for $p \in [1, 2)$. Therefore, the direct embedding is used only for a coarse preprocessing step followed by weighted sampling of rows from the original input matrix similar to the case of ℓ_2 sampling described above. We will deal with the sampling approach later.

Another problem we face is that while the embedding dimension is small $k \in O(\mathrm{poly}(\frac{d}{\varepsilon}))$ and independent of n for all $p \in [1, 2]$, the tight lower bounds [11] of $\Omega(n^{1-\frac{2}{p}} \log n)$ on approximating the pth-frequency moments for $p > 2$ imply that the embedding dimension must be polynomial in n and becomes quickly linear as $p \to \infty$. We summarize the oblivious ℓ_p-subspace embedding results of [146] in a slightly simplified theorem.

Theorem 2.5.7 (cf. [146])**.** *For every $p \in [1, \infty)$ there exists a family of random matrices $\Pi \in \mathbb{R}^{k \times n}$ such that for any basis U of a d-dimensional subspace of $(\mathbb{R}^n, \|\cdot\|_p)$ we have with constant probability for some $1 \leq \mu \in O((d \log d)^{\frac{1}{p}})$*

$$\forall x \in \mathbb{R}^d : \frac{1}{\mu} \, \|Ux\|_p \leq \|\Pi U x\|_q \leq \mu \, \|Ux\|_p,$$

where

$$(q,k) = \begin{cases} (2, O(d^2)) & \text{if } p \in [1,2) \\ (\infty, O(n^{1-\frac{2}{p}} \log n (d \log d)^{1+\frac{2}{p}} + d^{5+4p})) & \text{if } p \in (2,\infty). \end{cases}$$

The random projection of [146] can be composed as $\Pi = PD$ where P is any of the ε-subspace embeddings for ℓ_2 described above, but the embedding dimension k is changed according to Theorem 2.5.7. $D \in \mathbb{R}^{n\times n}$ denotes a diagonal matrix whose entries are $D_{ii} = 1/\lambda_i^{1/p}$ where each $\lambda_i \sim \exp(1)$ is drawn from a standard exponential distribution, whose density function is $p(x) = \exp(-x)$, for $x \in [0, \infty)$.

Now towards a sampling based algorithm, recall that the ℓ_2 leverage scores required an orthonormal basis for the columnspace spanned by the data matrix. Since ℓ_p norms are not rotationally invariant, we need as an analogue the notion of an (α, β, p)-well-conditioned basis which is a basis that does not distort the ℓ_p norms of vectors in the subspace too much. Its idea is due to [40].

Definition 2.5.8 ([40, 146])**.** *Given a matrix $M \in \mathbb{R}^{n\times d}$ and $p \in [1, \infty)$ let q denote the dual norm of p, i.e., $1/p + 1/q = 1$. We say $U \in \mathbb{R}^{n\times d}$ is an (α, β, p)-well-conditioned basis for the columnspace of M if*

1. *$\|M\|_p \leq \alpha$, and*
2. *$\forall x \in \mathbb{R}^d$: $\|x\|_q \leq \beta \|Mx\|_p$.*

We show that the random matrix Π from Theorem 2.5.7 yields an (α, β, p)-well-conditioned basis for the columnspace of a given matrix M.

Lemma 2.5.9. *Let Π be an embedding matrix that satisfies Theorem 2.5.7 for some $1 \leq \mu \in O((d \log d)^{\frac{1}{p}})$. Let $M \in \mathbb{R}^{n\times d}$. Consider the embedded matrix ΠM and let $\Pi M = U\Sigma V^T = UR$ be its singular value decomposition. Then $Q = MR^{-1}$ is an (α, β, p)-well-conditioned basis for the columnspace of M, where*

$$(\alpha, \beta) = \begin{cases} (\mu d, \mu) & \text{if } p \in [1, 2) \\ (\mu d, \mu d) & \text{if } p \in (2, \infty). \end{cases}$$

Proof. We are going to use the fact that $U = \Pi M R^{-1}$ is an orthonormal basis. Let e_i for $i \in [d]$ denote the ith standard basis vector. We have

$$\begin{aligned} \|Q\|_p = \left\|MR^{-1}\right\|_p = \left\| M \sum_{i=1}^{d} (R^{-1})^{(i)} e_i^T \right\|_p &= \left\| \sum_{i=1}^{d} M (R^{-1})^{(i)} e_i^T \right\|_p \\ \leq \sum_{i=1}^{d} \left\| M (R^{-1})^{(i)} e_i^T \right\|_p &= \sum_{i=1}^{d} \left\| M (R^{-1})^{(i)} \right\|_p \end{aligned} \tag{2.13}$$

Now suppose $p \in (2, \infty)$.

$$
\begin{aligned}
(2.13) &\leq \mu \sum_{i=1}^{d} \left\| \Pi M (R^{-1})^{(i)} \right\|_\infty \leq \mu\sqrt{d} \left(\sum_{i=1}^{d} \left\| \Pi M (R^{-1})^{(i)} \right\|_\infty^2 \right)^{\frac{1}{2}} \\
&\leq \mu\sqrt{d} \left(\sum_{i=1}^{d} \left\| \Pi M (R^{-1})^{(i)} \right\|_2^2 \right)^{\frac{1}{2}} \leq \mu\sqrt{d} \left(\sum_{i=1}^{d} \underbrace{\left\| U^{(i)} \right\|_2^2}_{=1} \right)^{\frac{1}{2}} = \mu d
\end{aligned}
$$

For arbitrary $x \in \mathbb{R}^d$ it holds that

$$
\|x\|_q \leq \sqrt{d} \, \|x\|_2 = \sqrt{d} \, \|Ux\|_2 = \sqrt{d} \left\| \Pi M R^{-1} x \right\|_2 \leq d \left\| \Pi M R^{-1} x \right\|_\infty \leq d\mu \, \|Qx\|_p .
$$

Consequently Q is $(\mu d, \mu d, p)$-well-conditioned. Next suppose $p \in [1, 2)$. Again we bound

$$
\begin{aligned}
(2.13) &\leq \mu \sum_{i=1}^{d} \left\| \Pi M (R^{-1})^{(i)} \right\|_2 \leq \mu\sqrt{d} \left(\sum_{i=1}^{d} \left\| \Pi M (R^{-1})^{(i)} \right\|_2^2 \right)^{\frac{1}{2}} \\
&\leq \mu\sqrt{d} \left(\sum_{i=1}^{d} \underbrace{\left\| U^{(i)} \right\|_2^2}_{=1} \right)^{\frac{1}{2}} = \mu d
\end{aligned}
$$

Also, since $p \leq 2$, its dual norm satisfies $q \geq 2$. Fix an arbitrary $x \in \mathbb{R}^d$. It follows that

$$
\|x\|_q \leq \|x\|_2 = \|Ux\|_2 = \left\| \Pi M R^{-1} x \right\|_2 \leq \mu \, \|Qx\|_p
$$

It follows that Q is even $(\mu d, \mu, p)$-well-conditioned in this case. □

ℓ_p-sampling The following theorem is due to Dasgupta et al. [40] and informally states that if we have a well conditioned basis, this implies a sampling scheme to construct a subspace embedding.

Theorem 2.5.10 ([40]). *Given a matrix $M \in \mathbb{R}^{n \times d}$, $p \in [1, \infty)$ and an (α, β, p)-well-conditioned basis for the columnspace of M, we can construct a sampling and reweighting matrix $\Pi \in \mathbb{R}^{k \times n}$, where $k \in O((\alpha\beta)^p (d \log(\frac{1}{\varepsilon}) + \log(\frac{1}{\eta}))/\varepsilon^2)$, such that with probability $1 - \eta$ we have for all $x \in \mathbb{R}^d$*

$$
(1 - \varepsilon) \, \|Mx\|_p \leq \|\Pi M x\|_p \leq (1 + \varepsilon) \, \|Mx\|_p .
$$

The algorithm works as follows, cf. [40, 146]. First, we embed the data matrix $M = [X, Y] \in \mathbb{R}^{n\times d}$ via the weak subspace embeddings P of Theorem 2.5.7 to obtain PM. Next we compute its singular value decomposition $PM = U\Sigma V^T = UR$ and an (α, β, p)-well conditioned basis $Q = MR^{-1}$ via Lemma 2.5.9. We define the ℓ_p-*leverage scores* with respect to Q.

$$l_i = \frac{\|Q_i\|_p^p}{\|Q\|_p^p}, \text{ for } i \in [n].$$

Now we fix a sampling size parameter $k \in O((\alpha\beta)^p(d\log(\frac{1}{\varepsilon}) + \log(\frac{1}{\eta}))/\varepsilon^2)$ depending on the conditioning properties of Q. Specifically, we have by Lemma 2.5.9 the following quantities

$$k \in O\left(d^{p+3}\log^2 d\left(\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\eta}\right)\right)\Big/\varepsilon^2\right) \qquad \text{for } p \in [1, 2)$$

$$k \in O\left(d^{2p+3}\log^2 d\left(\log\left(\frac{1}{\varepsilon}\right) + \log\left(\frac{1}{\eta}\right)\right)\Big/\varepsilon^2\right) \qquad \text{for } p \in (2, \infty).$$

Next we sample the input points one-by-one with probability $q_i = \min\{1, k \cdot l_i\}$. We reweight their contribution by $w_i = \frac{1}{q_i}$. This corresponds to defining the diagonal (sampling and reweighting) matrix Π by $\Pi_{ii} = 1/{q_i}^{1/p}$ with probability q_i and $\Pi_{ii} = 0$ otherwise. Also note, that the expected number of samples is k, which also holds with high probability via standard concentration inequalities, see [40].

2.5.3 Sampling via Sensitivity Scores

Now that we have reviewed several results on coresets for ℓ_p related norm functions, we turn our attention to a more general class of functions. Suppose we are given a data set $X \in \mathbb{R}^{n\times d}$ together with weights $w \in \mathbb{R}^n_{>0}$. The function under study is

$$f_w(X\beta) = \sum\nolimits_{i=1}^n w_i \cdot g(x_i\beta),$$

where $g : \mathbb{R} \to \mathbb{R}_{\geq 0}$. Examples include pth powers of the ℓ_p-norms where $g(x_i\beta) = |x_i\beta|^p$, and logistic regression where $g(x_i\beta) = \ln(1 + \exp(x_i\beta))$. When we associate with each point x_i the function $g_i(\beta) = g(x_i\beta)$, we can obtain a sampling based coreset construction with the following approach, called sensitivity sampling [25, 53, 103]. Then we have the following definition.

Definition 2.5.11 ([57, 103])**.** *Consider a family of functions $\mathcal{F} = \{g_1, \dots, g_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and weighted by $w \in \mathbb{R}^n_{>0}$. The sensitivity of g_i for $f_w(\beta) = \sum w_i g_i(\beta)$ is*

$$\varsigma_i = \sup \frac{w_i g_i(\beta)}{f_w(\beta)} \tag{2.14}$$

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

The sensitivity of a point measures its worst-case importance for approximating the objective function on the entire input data set. Performing importance sampling proportional to the sensitivities of the input points thus yields a good approximation. Computing the sensitivities is often intractable and involves solving the original optimization problem to near-optimality, which is the problem we want to solve in the first place, as pointed out in [25]. To get around this, it was shown that any upper bound on the sensitivities $s_i \geq \varsigma_i$ also has provable guarantees. However, the number of samples depends on the sum of their estimates $S = \sum s_i \geq \sum \varsigma_i = \mathfrak{S}$, so we need to carefully control this quantity. Another complexity measure that plays a crucial role in the sampling complexity is the Vapnik–Chervonenkis (VC) dimension of the range space induced by the set of functions under study.

Definition 2.5.12 ([57])**.** *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 $|\{G \cap R \mid R \in \text{ranges}\}| = 2^{|G|}$.*

Definition 2.5.13 ([57])**.** *Let $\mathcal{F}$ be a finite set of functions mapping from $\mathbb{R}^d$ to $\mathbb{R}_{\geq 0}$. For every $\beta \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$, let $\text{range}_{\mathcal{F}}(\beta, r) = \{f \in \mathcal{F} \mid f(\beta) \geq r\}$, and $\text{ranges}(\mathcal{F}) = \{\text{range}_{\mathcal{F}}(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}$, and $\mathfrak{R}_{\mathcal{F}} = (\mathcal{F}, \text{ranges}(\mathcal{F}))$ be the range space induced by $\mathcal{F}$.*

Recently a framework combining the sensitivity scores with the theory on the VC dimension of range spaces was developed in [25]. For technical reasons we use a slightly modified but unpublished version due to [57].

Theorem 2.5.14 ([57])**.** *Consider a family of functions $\mathcal{F} = \{f_1, \dots, f_n\}$ mapping from $\mathbb{R}^d$ to $[0, \infty)$ and a vector of weights $w \in \mathbb{R}^n_{>0}$. Let $\varepsilon, \eta \in (0, 1/2)$. Given $s_i \geq \varsigma_i$ for $i \in [n]$, let $S = \sum_{i=1}^n s_i \geq \mathfrak{S}$. We can compute in time $O(|\mathcal{F}|)$ a set $R \subset \mathcal{F}$ of*

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

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

$$\left| \sum_{f_j \in \mathcal{F}} w_j f_j(\beta) - \sum_{f_i \in R} u_i f_i(\beta) \right| \leq \varepsilon \sum_{f_j \in \mathcal{F}} w_j f_j(\beta).$$

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* $\mathcal{F}^*$ *that can be obtained by defining* $\mathcal{F}^*$ *to be the set of functions* $f_j \in \mathcal{F}$ *where each function is scaled by* $\frac{Sw_j}{s_j|R|}$.

We will need a bound on the VC dimension of a range space induced by an ℓ_1 related family of functions that yields an $(\varepsilon, 1)$-subspace embedding via the sensitivity framework.

Lemma 2.5.15. *Let* $X \in \mathbb{R}^{n\times d}, w \in \mathbb{R}^n_{>0}$. *The range space induced by* $\mathcal{F}_{\ell_1} = \{h_i(\beta) = w_i|x_i\beta| \mid i \in [n]\}$ *satisfies* $\Delta(\mathfrak{R}_{\mathcal{F}_{\ell_1}}) \leq 10(d+1)$.

Proof. Fix an arbitrary $G \subseteq \mathcal{F}_{\ell_1}$. Let $\Omega = \mathbb{R}^d \times \mathbb{R}_{\geq 0}$. We attempt to bound the quantity

$$\begin{aligned}
\left|\{G \cap R \mid R \in \text{ranges}(\mathcal{F}_{\ell_1})\}\right| &= \left|\{\text{range}_G(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}\right| \\
&= \left| \bigcup_{(\beta,r)\in\Omega} \Big\{ \{h_i \in G \mid h_i(\beta) \geq r\} \Big\} \right| \\
&= \left| \bigcup_{(\beta,r)\in\Omega} \Big\{ \{h_i \in G \mid w_i x_i \beta \geq r \vee -w_i x_i \beta \geq r\} \Big\} \right| \\
&\leq \left| \bigcup_{(\beta,r)\in\Omega} \Big\{ \{h_i \in G \mid w_i x_i \beta \geq r\} \Big\} \right| \\
&\qquad \cdot \left| \bigcup_{(\beta,r)\in\Omega} \Big\{ \{h_i \in G \mid -w_i x_i \beta \geq r\} \Big\} \right| \\
&= \left| \bigcup_{(\beta,r)\in\Omega} \Big\{ \{h_i \in G \mid w_i x_i \beta \geq r\} \Big\} \right|^2 . \qquad (2.15)
\end{aligned}$$

The inequality holds, since each non-empty set in the collection on the left hand side satisfies either of the conditions of the sets in the collections on the right hand side, or both, and is thus the union of two of those sets, one from each collection. It can thus comprise at most all unions obtained from combining any two of these sets. The last equality holds since for each fixed β we also union over $-\beta$ as we reach over all $\beta \in \mathbb{R}^d$. The two sets are thus equal.

Now note that each set $\{h_i \in G \mid w_i x_i \beta \geq r\}$ equals the set of weighted points that is shattered by the affine hyperplane classifier $w_i x_i \mapsto \mathbb{1}_{\{w_i x_i \beta - r \geq 0\}}$. Note that the VC dimension of the set of hyperplane classifiers is $d+1$ [94, 138]. To conclude the claimed bound on $\Delta(\mathfrak{R}_{\mathcal{F}_{\ell_1}})$ it is sufficient to show that the above term (2.15) is bounded strictly below $2^{|G|}$ for $|G| = 10(d+1)$. By a bound given in [22, 94] we have for this particular choice

$$\begin{aligned}(2.15) &\leq \left|\left\{\{h_i \in G \mid w_i x_i \beta - r \geq 0\} \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}\right\}\right|^2 \\ &\leq \left(\frac{e|G|}{d+1}\right)^{2(d+1)} < 2^{2(d+1)\log(30)} \leq 2^{2(d+1)5} = 2^{|G|}\end{aligned}$$

which implies that $\Delta(\mathfrak{R}_{\mathcal{F}_{\ell_1}}) < 10(d+1)$. □

It is noteworthy that similar results can be obtained for the corresponding ℓ_p related families, but we do not go into details about this.

Now we show that the VC dimension of the range space induced by the set of functions studied later in logistic regression can be related to the VC dimension of the set of linear classifiers. We first start with a fixed common weight and subsequently generalize to more general weights.

Lemma 2.5.16 (cf. [84])**.** *Let $X \in \mathbb{R}^{n \times d}, c \in \mathbb{R}_{>0}$. The range space induced by*

$$\mathcal{F}_{log}^c = \{g_i(\beta) = c\ln(1 + \exp(x_i\beta)) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \leq d+1$.

Proof. For all $G \subseteq \mathcal{F}_{log}^c$, we have

$$|\{G \cap R \mid R \in \mathrm{ranges}(\mathcal{F}_{log}^c)\}| = |\{\mathrm{range}_G(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}|$$

Note that $g(z) = c\ln(1+\exp(z))$ is invertible and monotone. Also note that g^{-1} maps $\mathbb{R}_{\geq 0}$ surjectively into $\mathbb{R}$. For all $\beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}$ we thus have

$$\begin{aligned}\mathrm{range}_G(\beta, r) &= \{g_i \in G \mid g_i(\beta) \geq r\} \\ &= \{g_i \in G \mid g(x_i\beta) \geq r\} \\ &= \{g_i \in G \mid x_i\beta \geq g^{-1}(r)\}.\end{aligned}$$

Now note that $\{g_i \in G \mid x_i\beta \geq g^{-1}(r)\}$ corresponds to the set of points that is shattered by the affine hyperplane classifier $x_i \mapsto \mathbb{1}_{\{x_i\beta - g^{-1}(r) \geq 0\}}$. We can conclude that

$$\left|\{\text{range}_G(\beta, r) \mid \beta \in \mathbb{R}^d, r \in \mathbb{R}_{\geq 0}\}\right| = \left|\Big\{\{g_i \in G \mid x_i\beta - s \geq 0\} \mid \beta \in \mathbb{R}^d, s \in \mathbb{R}\Big\}\right|$$

which means that the VC dimension of $\mathfrak{R}_{\mathcal{F}_{log}}$ is $d+1$ since the VC dimension of the set of hyperplane classifiers is $d+1$. [94, 138]. □

We finally generalize the result to a more general finite set of distinct weights.

Lemma 2.5.17. *Let $X \in \mathbb{R}^{n \times d}$, weighted by $w \in \mathbb{R}^n$ where for all $i \in [n], w_i \in \{v_1, \ldots, v_k\}$. The range space induced by*

$$\mathcal{F}_{log} = \{g_i(\beta) = w_i \ln(1 + \exp(x_i\beta)) \mid i \in [n]\}$$

satisfies $\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \in O(dk \log k)$.

Proof. We partition the functions into classes according to their weights. Let $F_i = \{g_j \in \mathcal{F}_{log} \mid w_j = v_i\}$, for $i \in [k]$. In each class, the weights are equal. Lemma 2.5.16 thus yields $\Delta(\mathfrak{R}_{F_i}) \leq d+1$ for each $i \in [k]$. Now we consider the rangespace induced by $\mathcal{F}_{log}$ and note that each $\text{range}_{\mathcal{F}_{log}} \in \text{ranges}(\mathcal{F}_{log})$ is the union of ranges from the sets ranges(F_i). More precisely, we have for each $\beta \in \mathbb{R}^d$ and $r \in \mathbb{R}_{\geq 0}$

$$\text{range}_{\mathcal{F}_{log}}(\beta, r) = \bigcup\nolimits_{i=1}^{k} \text{range}_{F_i}(\beta, r)\,.$$

Similarly to the proof of Lemma 2.5.15 where we had a 2-fold union, we can argue that all elements contained in the union must be contained in at least one range from the k individual rangespaces. This argument has been used to prove a general upper bound for k-fold unions (and intersections) of ranges [22] which now implies $\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \in O(dk \log k)$. □

2.5.4 A lower bounding technique based on communication complexity

We briefly introduce results from two-party one-way communication complexity. For a broader introduction on that topic we refer to [101]. In a communication game, Alice and Bob are given input strings $x \in X, y \in Y$. Their goal is to compute some boolean function $f\colon X \times Y \to \{0, 1\}$ by exchanging as little information as possible. The difficulty lies in the fact that Alice only knows x and Bob only knows y. One-way communication protocols are even more restrictive. Alice can derive a message $m(x)$ from her input, whose size is $|m(x)|$ bits. Then she sends the message to Bob whose task is to compute the output only based on this message and his own input string. No further communication is allowed. We denote by $R_\eta(f)$ the minimum number of bits communicated by any randomized two-party one-way protocol $P(x, y)$ that computes $f(x, y)$ with error probability at most η. More formally, if

$$\mathbf{Pr}\left[P(x, y) \neq f(x, y)\right] \leq \eta,$$

then the number of communicated bits satisfies $|m(x)| \geq R_\eta(f)$.

We will use the following problem known in the literature as the indexing problem. Alice is given a string $x \in \{0, 1\}^n$ and Bob has an index $i \in [n]$. Bobs task is to compute $IND\colon \{0, 1\}^n \times [n] \rightarrow \{0, 1\}$ mapping $(x, i) \mapsto x_i$, based on Alice's message and his index. The following result is known regarding the one-way communication complexity of IND. The difficulty of the problem is inherently one-way; otherwise Bob could simply send his index of size $O(\log n)$ to Alice, cf. [98]. If the entire communication consists of only a single message sent by Alice to Bob, the message must contain $\Omega(n)$ bits.

Theorem 2.5.18. *([87, 98]) For every constant* $0 < \eta \leq \frac{1}{3}$ *we have* $R_\eta(IND) \in \Omega(n)$.

We note that the result holds for any constant failure probability $< \frac{1}{2}$ via standard probability amplification [113]. We are going to use this result to derive space lower bounds on the size of coresets for specific problems. The high-level idea is that we assume there exists a randomized coreset construction for the problem under study that succeeds with constant probability. Then we can create the following protocol for the indexing problem. Alice produces a large and carefully designed point set X that depends on her input $x \in \{0, 1\}^n$. Now, she computes a coreset $C(X)$ of X and sends it as a message to Bob. Bob takes the coreset and computes a reasonably good approximation to the problem based on the coreset. From this approximation he concludes the value of x_i. If this protocol fails only with constant probability, then by Theorem 2.5.18 the size of the coreset $C(X)$ must be at least $\Omega(n)$ bits. Such arguments have been applied for various problems in the streaming and coreset literature, e.g. in [3, 31].

2.6 Learning from probabilistic points

The input to a problem on probabilistic data is a probabilistic set of points. This can be defined, cf. [38] as a set $\mathcal{D} = \{D_1, \ldots, D_n\}$ of n discrete and independent probability distributions. Each distribution D_i is defined over a set of z possible locations $q_{i1}, \ldots, q_{iz} \in \mathbb{R}^d \cup \{\bot\}$, where $\bot$ indicates that the ith point is not present in a sampled set, i.e. $\{q_{ij}\} = \emptyset$ if $q_{ij} = \bot$. A probability p_{ij} is associated with each location such that $\sum_{j=1}^{z} p_{ij} = 1$

for every $i \in [n] = \{1, \ldots, n\}$. Thus, the probabilistic points can be considered to be independent random variables $X_i \sim D_i$. The locations together with the probabilities specify distributions $\mathbf{Pr}\left[X_i = q_{ij}\right] = p_{ij}$ for every $i \in [n]$ and $j \in [z]$. A probabilistic set $X \sim \mathcal{D}$ consisting of probabilistic points is therefore also a random variable. The underlying sample space can be described by $\Omega = [z]^n$. Therefore every random choice of elements $(j_1, \ldots, j_n) \in \Omega$ determines a realization $P_{(j_1,\ldots,j_n)} = X(j_1, \ldots, j_n) = (q_{1j_1}, \ldots, q_{nj_n})$ with $\mathbf{Pr}\left[X = P_{(j_1,\ldots,j_n)}\right] = \prod_{i\in[n]} p_{ij_i}$ by independence. We will slightly abuse notation and identify with each realization $P_{(j_1,\ldots,j_n)}$ the multiset $P = \{q_{1j_1}, \ldots, q_{nj_n}\}$.

Before we get to the problems that we are going to tackle in the probabilistic data setting, we define the notion of a ball for ℓ_p spaces.

Definition 2.6.1. *We define the ball with respect to $(\mathbb{R}^d, \|\cdot\|_p)$ for $p \in [1, \infty)$ centered at $c \in \mathbb{R}^d$ with radius $r \in \mathbb{R}_{\geq 0}$ as*

$$\mathrm{B_p}(c, r) = \{x \in \mathbb{R}^d \mid \|x - c\|_p \leq r\}.$$

We drop the subscript in the special case of a Euclidean ball, with $p = 2$, when this is clear from the context.

We will mainly investigate the (probabilistic) smallest enclosing ball problem. We will also deal with 1-median problems in this context. We thus need definitions for both. Here, we will focus on the Euclidean distance of arbitrary points $a, b \in \mathbb{R}^d \cup \{\perp\}$, i.e., $\|a - b\|_2$, where we define $\|a - \perp\|_2 = 0$. We also assume that the maximum and sum taken over an empty set equals zero.

Definition 2.6.2. *Given a finite set of points $P \subset \mathbb{R}^d$, the* 1-median cost *of a given center $c \in \mathbb{R}^d$ is defined as*

$$\mathrm{cost}_{\mathrm{MED}}(P, c) = \sum_{p\in P} \|p - c\|_2 \,.$$

The 1-center cost*, or cost of the smallest enclosing ball of P that is centered at c, is its maximum distance over the input points,*

$$\mathrm{cost}_{\mathrm{SEB}}(P, c) = \max_{p\in P} \|p - c\|_2 \,.$$

Now we define the probabilistic versions of the above problems. Note that the expectations are taken over the randomness of drawing realizations P of $X \sim \mathcal{D}$ according to the n input distributions.

Definition 2.6.3. *Let $\mathcal{D}$ be a set of n discrete distributions, where each distribution is defined over z locations in $\mathbb{R}^d \cup \{\perp\}$. The probabilistic 1-median problem is to find a center $c \in \mathbb{R}^d$ that minimizes the expected 1-median cost, i.e.,*

$$c \in \operatorname{argmin}_{c' \in \mathbb{R}^d} \mathbb{E}\left[\mathrm{cost}_{\mathrm{MED}}(X, c')\right] .$$

The probabilistic smallest enclosing ball problem is to find a center $c \in \mathbb{R}^d$ that minimizes the expected smallest enclosing ball cost, i.e.,

$$c \in \operatorname{argmin}_{c' \in \mathbb{R}^d} \mathbb{E}\left[\mathrm{cost}_{\mathrm{SEB}}(X, c')\right] .$$

In both cases the expectation is taken over the randomness of $X \sim \mathcal{D}$.

We now formally define the notion of a $(1 + \varepsilon)$-approximation for both of the problems defined above.

Definition 2.6.4. *Let $\mathcal{D}$ be a set of n discrete distributions, where each distribution is defined over z locations in $\mathbb{R}^d \cup \{\perp\}$. Let $\varepsilon > 0$. A $(1+\varepsilon)$-approximation to the probabilistic 1-median problem is a center $c \in \mathbb{R}^d$ that satisfies*

$$\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c)] \leq (1 + \varepsilon) \min_{c' \in \mathbb{R}^d} \mathbb{E}\left[\mathrm{cost}_{\mathrm{MED}}(X, c')\right] .$$

A $(1 + \varepsilon)$-approximation to the probabilistic smallest enclosing ball problem is a center $c \in \mathbb{R}^d$ that satisfies

$$\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, c)] \leq (1 + \varepsilon) \min_{c' \in \mathbb{R}^d} \mathbb{E}\left[\mathrm{cost}_{\mathrm{SEB}}(X, c')\right] .$$

In both cases the expectations are taken over the randomness of $X \sim \mathcal{D}$.

Cormode and McGregor [38] introduced the study of *probabilistic clustering problems.* They developed approximation algorithms for the probabilistic settings of k-means, k-median as well as k-center clustering. For the k-center clustering problem their results are $O(1)$-approximation algorithms with a blow-up on the number of centers k and apply to arbitrary metrics. If the probability distributions are restricted to the cases that a point exists or not, then they achieve a $(1 + \varepsilon)$-approximation but again with a substantially inflated number of centers by a factor of $O(\varepsilon^{-1} \log^2 n)$. Guha and Munagala [68] improved upon the previous work. They achieved $O(1)$-approximations for very large constants while preserving the number of centers.

Lammersen et al. [102] developed the first k-median clustering algorithms for uncertain datasets in the streaming setting via the first coreset constructions for probabilistic points. Huang et al. [83] developed ε-kernel for stochastic data.

Huang and Li [82] generalized the $(1+\varepsilon)$-approximation to the probabilistic smallest enclosing ball problem presented in this manuscript to Euclidean k-center in $\mathbb{R}^d$ for fixed constant k and d. It was noted that assuming k is a constant is necessary for obtaining a $(1+\varepsilon)$-approximation in polynomial time, since even the deterministic Euclidean k-center problem in $\mathbb{R}^2$ is hard for the set of optimization problems that allow polynomial-time constant-factor approximation algorithms (APX) when k is unbounded [52].

We note that the algorithm of [82] is only a *polynomial time approximation scheme* (PTAS) rather than a *fully polynomial time approximation scheme* (FPTAS) since its running time is polynomial in the input size but exponential in $\frac{1}{\varepsilon}$, specifically $O(n^{O(\varepsilon^{-(2d+2)}dk^5)})$, even for $k=1$ and $d \in O(1)$. We note that our result presented later in this manuscript is still the only FPTAS for the probabilistic Euclidean 1-center problem in constant dimension.

3 Subspace embeddings for Bayesian regression

3.1 Approximate Bayesian ℓ_2-regression

In this section we show that a data reduction via ε-subspace embeddings preserves the posterior distribution in Bayesian regression models based on normal distributions, cf. Scheme (2.6). Specifically, we bound the Wasserstein distance between the original posterior and its counterpart that is defined only on the considerably smaller embedding. To this end, we begin with a known result on ordinary least squares regression. We combine this result later with a more involved argument on the covariance structure to show that the likelihoods are close to each other. Finally, we show how to extend the result to posterior approximation for normal priors.

3.1.1 Ordinary least squares regression

Note that for multivariate normal distributions their means equal their modes [89]. They can be obtained by finding a maximum likelihood estimator via solving the least squares regression problem, see Section 2.3.1. We can thus approximate them via an approximation to the latter optimization problem.

In our first lemma we show that using an $\frac{\varepsilon}{2}$-subspace embedding Π for the columnspace of $[X, Y]$, we can approximate the least squares regression problem up to a factor of $1+\varepsilon^2$. That is, we can find a solution ν by projecting ΠY into the columnspace of ΠX such that $\|X\nu - Y\|_2 \leq (1+\varepsilon^2)\min_{\beta\in\mathbb{R}^d}\|X\beta - Y\|_2$. The proof follows the outline of Theorem 3.1 from [31].

Lemma 3.1.1. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $\gamma = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\|X\beta - Y\|_2^2$ and similarly $\nu = \operatorname{argmin}_{\beta\in\mathbb{R}^d}\|\Pi(X\beta - Y)\|_2^2$. Then*

$$\|X\nu - Y\|_2^2 \leq (1+\varepsilon^2)\,\|X\gamma - Y\|_2^2\,.$$

Proof. Let $X = U\Sigma V^T$ be the singular value decomposition of X. Recall that by the normal equation (2.2) for the original problem we have

$$U^T(X\gamma - Y) = X^T(X\gamma - Y) = 0,$$

and similarly we have for the embedded problem

$$U^T\Pi^T\Pi(X\nu - Y) = X^T\Pi^T\Pi(X\nu - Y) = 0.$$

We will need a bound on the squared distance of the two solutions in the columnspace of X, i.e., on $\|X(\gamma - \nu)\|_2^2$. To this end, consider the vector $\vartheta = \Sigma V^T(\gamma - \nu)$ and note that its norm is the same, since

$$\begin{aligned}\|X(\gamma - \nu)\|_2^2 &= (\gamma - \nu)^T V\Sigma U^T U\Sigma V^T(\gamma - \nu)\\ &= (\gamma - \nu)^T V\Sigma I_d\Sigma V^T(\gamma - \nu) = \left\|\Sigma V^T(\gamma - \nu)\right\|_2^2.\end{aligned}$$

It is thus sufficient to bound ϑ instead. By the triangle inequality and definition of the spectral norm, we have

$$\begin{aligned}\|\vartheta\|_2 &\le \left\|U^T\Pi^T\Pi U\vartheta\right\|_2 + \left\|U^T\Pi^T\Pi U\vartheta - \vartheta\right\|_2\\ &\le \left\|U^T\Pi^T\Pi U\vartheta\right\|_2 + \left\|U^T\Pi^T\Pi U - I\right\|_2 \|\vartheta\|_2\\ &\le \left\|U^T\Pi^T\Pi U\vartheta\right\|_2 + \varepsilon\,\|\vartheta\|_2\end{aligned}$$

where the last inequality is a direct application of the subspace embedding property (2.8). It thus follows that

$$\|\vartheta\|_2 \le \left\|U^T\Pi^T\Pi U\vartheta\right\|_2 /(1 - \varepsilon) \le 2\left\|U^T\Pi^T\Pi U\vartheta\right\|_2.$$

We still need a bound on $\left\|U^T\Pi^T\Pi U\vartheta\right\|_2$. Using the normal equation for the embedded problem, and since X and U are only different bases for the same linear subspace, we have

$$\begin{aligned}U^T\Pi^T\Pi U\vartheta &= U^T\Pi^T\Pi U\Sigma V^T(\gamma - \nu)\\ &= U^T\Pi^T\Pi X(\gamma - \nu)\\ &= U^T\Pi^T\Pi X(\gamma - \nu) + \underbrace{U^T\Pi^T\Pi(X\nu - Y)}_{=0}\\ &= U^T\Pi^T\Pi(X\gamma - Y).\end{aligned}$$

Next, we can apply Lemma 2.5.3 on approximate matrix multiplication, since Π is an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace spanned by U. This yields

$$\begin{aligned}\left\|U^T\Pi^T\Pi U\vartheta\right\|_2 &= \left\|U^T\Pi^T\Pi(X\gamma - Y)\right\|_2\\ &\leq \frac{\varepsilon}{2}\left\|U\right\|_2\left\|X\gamma - Y\right\|_2\\ &= \frac{\varepsilon}{2}\left\|X\gamma - Y\right\|_2\end{aligned}$$

We thus have $\|X(\gamma-\nu)\|_2^2 \leq \varepsilon^2\|X\gamma - Y\|_2^2$, which enables us to conclude our proof. Again, the normal equations for the original problem imply that $X\gamma - Y$ and $X(\gamma-\nu)$ are orthogonal. So we can apply the Pythagorean Theorem to get

$$\begin{aligned}\|X\nu - Y\|_2^2 &= \|X\gamma - Y\|_2^2 + \|X(\gamma-\nu)\|_2^2\\ &\leq (1+\varepsilon^2)\|X\gamma - Y\|_2^2. \qquad \square\end{aligned}$$

3.1.2 Embedding the likelihood

Now, we study the distributions proportional to the likelihood functions $p \propto \mathcal{L}(\beta|X,Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$ and bound their Wasserstein distance.

The following observation is similar to known results (cf. [64, 92]) and will help us to derive a bound on the ℓ_2 Wasserstein distance of two normal distributions. It allows us to investigate their means and their covariance structure separately.

Observation 3. *Let $Z_1, Z_2 \in \mathbb{R}^d$ be random variables with finite first moments $m_1, m_2 < \infty$ and let $Z_1^m = Z_1 - m_1$, respectively, $Z_2^m = Z_2 - m_2$ be their mean-centered counterparts. Then it holds that*

$$\mathbb{E}\left[\|Z_1 - Z_2\|_2^2\right] = \|m_1 - m_2\|_2^2 + \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right].$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\|Z_1 - Z_2\|_2^2\right] &= \mathbb{E}\left[\|Z_1 - m_1 + m_1 - Z_2 + m_2 - m_2\|_2^2\right]\\ &= \mathbb{E}\left[\|Z_1^m - Z_2^m + m_1 - m_2\|_2^2\right]\\ &= \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2 + \|m_1 - m_2\|_2^2\right] + 2(m_1 - m_2)^T\underbrace{\mathbb{E}[Z_1^m - Z_2^m]}_{=0}\\ &= \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right] + \|m_1 - m_2\|_2^2 \qquad \square\end{aligned}$$

We begin our investigation of p, p' with a bound on the distance of the means γ and ν. Previous results [44, 129] considered specific embedding methods, while we generalize to arbitrary ε-subspace embeddings.

Lemma 3.1.2. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $\gamma = \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|X\beta - Y\|_2^2$. Similarly let $\nu = \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|\Pi(X\beta - Y)\|_2^2$. Then*

$$\|\gamma - \nu\|_2^2 \leq \frac{\varepsilon^2}{\sigma_{\min}^2(X)} \|X\gamma - Y\|_2^2 .$$

Proof. Let $X = U\Sigma V^T$ denote the singular value decomposition of X. Let $\vartheta = V^T(\gamma - \nu)$. First note that γ and ν are both contained in the columnspace of V, cf. [129], which means that V^T is a proper rotation with respect to $\gamma - \nu$. Thus,

$$\begin{aligned}\|X(\gamma-\nu)\|_2^2 &= \left\|U\Sigma V^T(\gamma-\nu)\right\|_2^2 = \left\|\Sigma V^T(\gamma-\nu)\right\|_2^2 \\ &= \sum\nolimits_{i=1}^d \sigma_i^2\vartheta_i^2 \geq \sum\nolimits_{i=1}^d \sigma_{\min}^2\vartheta_i^2 \\ &= \sigma_{\min}^2 \left\|V^T(\gamma-\nu)\right\|_2^2 = \sigma_{\min}^2 \|\gamma-\nu\|_2^2 .\end{aligned}$$

Consequently, it remains to bound $\|X(\gamma - \nu)\|_2^2$. This can be done by using the fact that the minimizer γ is obtained by projecting Y orthogonally onto the columnspace of X by the normal equation $X^T(X\gamma - Y) = 0$. Furthermore, by Lemma 3.1.1 it holds that $\|X\nu - Y\|_2^2 \leq (1 + \varepsilon^2) \|X\gamma - Y\|_2^2$. Again putting this into the Pythagorean theorem and rearranging we get that

$$\|X(\gamma-\nu)\|_2^2 = \|X\nu - Y\|_2^2 - \|X\gamma - Y\|_2^2 \leq \varepsilon^2 \|X\gamma - Y\|_2^2 .$$

Combining these equations yields the claim

$$\|\gamma-\nu\|_2^2 \leq \frac{1}{\sigma_{\min}^2(X)} \|X(\gamma-\nu)\|_2^2 \leq \frac{\varepsilon^2}{\sigma_{\min}^2(X)} \|X\gamma - Y\|_2^2 . \qquad \square$$

Now it remains to consider the covariances. By Observation 3, we can center both distributions to their means and assume w.l.o.g. $\gamma = \nu = 0$. More specifically, we derive a bound on $\inf_{\lambda\in\Lambda(p,p')} \mathbb{E}\left[\|Z_1^m - Z_2^m\|_2^2\right]$, i.e., the least expected squared Euclidean distance of two points drawn from a joint distribution whose marginals are the *mean-centered* original distribution and its embedded counterpart. The idea behind our next lemma is that we can define a properly chosen joint distribution and bound the expected squared distance for this particular choice.

Lemma 3.1.3. *Given $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $p \propto \mathcal{L}(\beta|X,Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Let Z_1^m, Z_2^m be the mean-centered versions of the random variables $Z_1 \sim p$ and $Z_2 \sim p'$ that are distributed according to p and p' respectively. Then we have*

$$\inf_{\lambda \in \Lambda(p,p')} \mathbb{E}_\lambda \left[\|Z_1^m - Z_2^m\|_2^2 \right] \leq \varepsilon^2 \operatorname{tr}\left((X^T X)^{-1} \right).$$

Proof. Our plan is to design a joint distribution that deterministically maps points from one distribution to another in such a way that we can bound the distance of every pair of points. This can be done by using a bijective map $g\colon \mathbb{R}^d \to \mathbb{R}^d$, that implicitly defines the joint distribution $\lambda \in \Lambda(p,p')$ via Dirac's delta function, as described in Section 2.2. It thus remains to find such a map and bound the distance between every pair of points x and $g(x)$.

According to Observation 1, the columnspace of a matrix is expanded or contracted by a factor of at most $(1 \pm \varepsilon)$, when we apply the embedding Π. We will use this fact as follows. Let $X = U\Sigma V^T$ and $\Pi X = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ denote the singular value decompositions of X and ΠX. Now, to define the x-y-pairs that will be mapped to each other via g, we consider vectors $x, x', y, y' \in \mathbb{R}^d$ where x' and y' are contained in the columnspaces of V and $\tilde{V}$, respectively. To obtain the bijection g, let the vectors have the following properties for any fixed radius $\rho \geq 0$:

1. $\|x'\|_2 = \|y'\|_2 = \rho$
2. $x = \Sigma V^T x'$
3. $y = \tilde{\Sigma}\tilde{V}^T y'$
4. $\exists \tau \geq 0\colon x = \tau y$.

By the first property, x' and y' lie on a d-dimensional sphere with radius ρ centered at 0. Thus, there exists a rotation matrix $R \in \mathbb{R}^{d \times d}$ such that $y' = Rx'$. Such a map is bijective by definition. The second item defines a bijection of such spheres to ellipsoids, which remain centered at the origin by linearity. The third property is defined analogously. The fourth property ensures that x and y have the same orientation but possibly different norm, where τ quantifies the scaling factor. Consequently they lie on a ray starting from the origin. Note that every such ray intersects each ellipsoid exactly once.

Our bijection can be defined as

$$g : \mathbb{R}^d \to \mathbb{R}^d$$

$$x \ \mapsto \tilde{\Sigma}\tilde{V}^T R V \Sigma^{-1} x$$

by composing the map ΣV^T, defined in the second item, with the rotation R and finally with $\tilde{\Sigma}\tilde{V}^T$ from the third property. The composition is bijective since each of its components are bijections.

Now, in order to bound the distance $\|Z_1^m - Z_2^m\|_2^2$ for any realization of (Z_1^m, Z_2^m) according to their joint distribution implicitly defined by g. We have by the fourth property that $Z_2^m = g(Z_1^m) = \tau Z_1^m$ and we need a bound on the parameter τ. We combine the second and third properties with the fourth, which yields

$$\Sigma V^T x' = \tau \tilde{\Sigma}\tilde{V}^T y'.$$

We thus have by rearranging

$$\begin{aligned} y'^T y' \tau &= (y'^T \tilde{V})\tilde{\Sigma}^{-1}\Sigma(V^T x') \\ &= \sum\nolimits_{i=1}^{d} (y'^T \tilde{V})_i (V^T x')_i \frac{\sigma_i}{\tilde{\sigma}_i} \\ &\leq \sum\nolimits_{i=1}^{d} (y'^T \tilde{V})_i (V^T x')_i \frac{\sigma_i}{\sigma_i\sqrt{1-\varepsilon}} \\ &\leq (1+\varepsilon) \sum\nolimits_{i=1}^{d} (y'^T \tilde{V})_i (V^T x')_i \\ &\leq (1+\varepsilon)\, \rho^2. \end{aligned}$$

The first inequality follows from $\tilde{\sigma}_i \geq \sqrt{1-\varepsilon}\ \sigma_i$, see Observation 1. This eventually means that $\tau \leq (1+\varepsilon)$ since $y'^T y' = \rho^2$ by the first property. A lower bound of $\tau \geq (1-\varepsilon)$ can be derived analogously by using $\tilde{\sigma}_i \leq \sqrt{1+\varepsilon}\ \sigma_i$.

Now we can conclude our proof. It follows that

$$\begin{aligned} \inf_{\lambda' \in \Lambda(p,p')} \mathbb{E}_{\lambda'}\Big[\|Z_1^m - Z_2^m\|_2^2\Big] \leq \mathbb{E}_{\lambda}\Big[\|Z_1^m - Z_2^m\|_2^2\Big] &\leq \mathbb{E}_{\lambda}\Big[\|\varepsilon Z_1^m\|_2^2\Big] \\ &= \varepsilon^2\ \mathbb{E}_{\lambda}\Big[\|Z_1^m\|_2^2\Big] = \varepsilon^2 \ \mathrm{tr}\left((X^T X)^{-1}\right). \end{aligned}$$

The last equality holds since the expected squared norm of the mean-centered random variable, i.e. its second moment, is just the trace of its covariance matrix. □

Combining the above results we get the following lemma.

Lemma 3.1.4. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $p \propto \mathcal{L}(\beta|X,Y)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Then*

$$\mathcal{W}_2^2(p,p') \leq \frac{\varepsilon^2}{\sigma_{\min}^2(X)} \, \|X\gamma - Y\|_2^2 + \varepsilon^2 \operatorname{tr}\left((X^TX)^{-1}\right).$$

Proof. The lemma follows from Definition 2.2.1, Observation 3, Lemma 3.1.2 and Lemma 3.1.3. □

In the following, we assume that there exists some constant $\vartheta \in (0,1]$ such that $\|X\gamma\|_2 \geq \vartheta \|Y\|_2$, cf. [44]. This is very natural in the setting of linear regression since it means that at least a constant fraction of the dependent variable Y can be explained within the columnspace of the data X. If this is not true, it indicates that a linear model is not appropriate at all for the given data [67]. This mild assumption yields a $(1+O(\varepsilon))$-approximation of the likelihood with respect to the Wasserstein weight.

Corollary 3.1.5. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $p \propto \mathcal{L}(\beta|X,Y)$ and similarly let $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Let $\kappa(X) = \sigma_{\max}(X)/\sigma_{\min}(X)$ be the condition number of X. Assume that for some $\vartheta \in (0,1]$ we have $\|X\gamma\|_2 \geq \vartheta \|Y\|_2$. Then*

$$\mathcal{W}_2(p') \leq \left(1 + \frac{\kappa(X)}{\vartheta}\,\varepsilon\right) \mathcal{W}_2(p).$$

Proof. By definition, the squared ℓ_2 Wasserstein weight of p equals its second moment. Since p is a normal distribution with mean γ and covariance matrix $(X^TX)^{-1}$, we thus have

$$\mathcal{W}_2^2(p) = \|\gamma\|_2^2 + \operatorname{tr}\left((X^TX)^{-1}\right)$$

and

$$\mathcal{W}_2^2(p') = \|\nu\|_2^2 + \operatorname{tr}\left((X^T\Pi^T\Pi X)^{-1}\right).$$

Since Π is an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X we know from Observation 1, that all the squared singular values of X are approximated up to less than $(1 \pm \varepsilon)$ error and so are their inverses. Therefore, we have

$$\operatorname{tr}\left((X^T\Pi^T\Pi X)^{-1}\right) \leq (1+\varepsilon) \operatorname{tr}\left((X^TX)^{-1}\right). \tag{3.1}$$

It remains to bound $\|\nu\|_2^2$. To this end we use the assumption that for some $\vartheta \in (0,1]$ we have $\|X\gamma\|_2 \geq \vartheta \|Y\|_2$. By the normal equation $X^T(X\gamma - Y) = 0$ we can apply the Pythagorean Theorem. This yields

$$\|X\gamma - Y\|_2^2 = \|Y\|_2^2 - \|X\gamma\|_2^2 \leq \|X\gamma\|_2^2 \left(\frac{1}{\vartheta^2} - 1\right) \leq \frac{\|X\gamma\|_2^2}{\vartheta^2}. \tag{3.2}$$

Now we can apply the triangle inequality, Lemma 3.1.2, Inequality (3.2) and Definition 2.1.4 to get

$$\begin{aligned}
\|\nu\|_2 &\leq \|\gamma\|_2 + \|\nu - \gamma\|_2 \\
&\leq \|\gamma\|_2 + \frac{\varepsilon}{\sigma_{\min}(X)} \|X\gamma - Y\|_2 \\
&\leq \|\gamma\|_2 + \frac{\varepsilon}{\vartheta\sigma_{\min}(X)} \|X\gamma\|_2 \\
&\leq \|\gamma\|_2 + \frac{\varepsilon}{\vartheta\sigma_{\min}(X)} \|X\|_2 \|\gamma\|_2 \\
&= \|\gamma\|_2 + \frac{\varepsilon}{\vartheta} \kappa(X) \|\gamma\|_2 \\
&= \left(1 + \frac{\kappa(X)}{\vartheta}\varepsilon\right) \|\gamma\|_2 .
\end{aligned}$$

Combining this with Inequality (3.1), the claim follows since $\frac{\kappa(X)}{\vartheta} \geq 1$ and therefore $(1+\varepsilon) \leq (1 + \frac{\kappa(X)}{\vartheta}\varepsilon)^2$ and finally taking square roots on both sides. □

3.1.3 Bayesian posterior approximation

So far we have shown that using subspace embeddings to compress a given data set for regression yields a good approximation to the likelihood. Note that in a Bayesian regression setting, Lemma 3.1.4 already implies a similar approximation error for the posterior distribution if the prior for β is a uniform distribution over $\mathbb{R}^d$. This is an improper, non-informative choice, $p_{\text{pre}}(\beta) = \mathbb{1}_{\mathbb{R}^d}(\beta)$. From this, it follows that

$$p_{\text{post}}(\beta|X,Y) \propto \mathcal{L}(\beta|X,Y) \cdot \mathbb{1}_{\mathbb{R}^d}(\beta) = \mathcal{L}(\beta|X,Y).$$

The remaining term is simply the likelihood which is a proper normal distribution up to normalization. For regression models, especially on data sets with large n, this covers a considerable amount of the cases of interest, especially if there is no actual prior knowledge, cf. [61]. We will extend this to arbitrary normal priors $p_{\text{pre}}(\beta)$ leading to our main result: an approximation guarantee for normal Bayesian linear regression in its most general form.

To this end, let m be the mean of the prior distribution and let S be derived from its covariance matrix by $\Sigma = \varsigma^2(S^TS)^{-1}$. Now, the posterior distribution is given by

$$\begin{aligned} &p_{\text{post}}(\beta|X,Y) \propto \mathcal{L}(\beta|X,Y) \cdot p_{\text{pre}}(\beta) \\ &= \frac{1}{(2\pi\varsigma^2)^{n/2}} \cdot \exp\left(-\frac{1}{2\varsigma^2}\left\|X\beta - Y\right\|_2^2\right) \cdot \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2\varsigma^2}\left\|S(\beta-m)\right\|_2^2\right). \end{aligned}$$

Thus, we know that up to some constants that are independent of β, the exponent of the posterior can be described by

$$\|X\beta - Y\|_2^2 + \|S(\beta - m)\|_2^2 \tag{3.3}$$

which contains all the information to define the mean and covariance structure of the posterior distribution. Now let

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

With these definitions we can rewrite Equation (3.3) above as $\|Z\beta - z\|_2^2$. This, in turn, can be treated as a (frequentist) regression problem to which we can apply Lemma 3.1.4. To this end, however, we have to use a subspace embedding for the columnspace of Z instead of only for X. By Lemma 2.5.6 it is not necessary to do this explicitly. Embedding only the data matrix $[X, Y]$ with an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X is sufficient to have an embedding for the entire columnspace defined by the data and the prior information in Z, and therefore, to have a proper approximation of the posterior distribution defined on $[Z, z]$. This finally enables us to prove our main theoretical result.

Theorem 3.1.6. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $p_{\text{pre}}(\beta)$ be an arbitrary normal distribution with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma = \varsigma^2(S^TS)^{-1} \in \mathbb{R}^{d\times d}$. Let*

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \textit{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

Let $\mu = \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|Z\beta - z\|_2^2$ be the posterior mean. Let $p \propto \mathcal{L}(\beta|X,Y) \cdot p_{\text{pre}}(\beta)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y) \cdot p_{\text{pre}}(\beta)$. Then

$$\mathcal{W}_2^2(p,p') \leq \frac{\varepsilon^2}{\sigma_{\min}^2(Z)} \|Z\mu - z\|_2^2 + \varepsilon^2 \operatorname{tr}\left((Z^TZ)^{-1}\right).$$

Proof. From our previous reasoning we know that approximating the posterior distribution can be reduced to approximating a likelihood function that is defined in terms of the data as well as the parameters of the prior distribution. This has been shown by rewriting Equation (3.3) above as $\|Z\beta - z\|_2^2$. For that reason, we can apply Lemma 3.1.4 to get the desired result if we are given an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of Z. Using Lemma 2.5.6 we know that to achieve this, it is sufficient to use an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X independent of the covariance and mean that define the prior distribution. □

Similar to Corollary 3.1.5 we have the following result concerning the posterior distribution.

Corollary 3.1.7. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $\frac{\varepsilon}{2}$-subspace embedding for the columnspace of X. Let $p_{\mathrm{pre}}(\beta)$ be an arbitrary normal distribution with mean $m \in \mathbb{R}^d$ and covariance matrix $\Sigma = \varsigma^2(S^TS)^{-1} \in \mathbb{R}^{d\times d}$. Let*

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

Let $\mu = \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|Z\beta - z\|_2^2$ be the posterior mean. Let $p \propto \mathcal{L}(\beta|X,Y)\cdot p_{\mathrm{pre}}(\beta)$ and $p' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)\cdot p_{\mathrm{pre}}(\beta)$. Let $\kappa(Z)$ be the condition number of Z. Assume that for some $\vartheta \in (0,1]$ we have $\|Z\mu\|_2 \geq \vartheta\,\|z\|_2$. Then we have

$$\mathcal{W}_2(p') \leq \left(1 + \frac{\kappa(Z)}{\vartheta}\,\varepsilon\right)\,\mathcal{W}_2(p).$$

Both Theorem 3.1.6 and Corollary 3.1.7 show that the sketch preserves the mean and the covariance structure of the posterior distribution very well. Note that for normal distributions, these parameters fully characterize the distribution; they are thus called *sufficient statistics*. Therefore, one can see the corresponding parameters based on the sketched data set as very accurate approximations for the sufficient statistics of the posterior distribution.

3.2 Extension to Bayesian ℓ_p-regression

In this section we show how to extend the results from Section 3.1 to ℓ_p-regression for general but fixed $p \in [1,\infty)$. We show that a data reduction via (ε,p)-subspace embeddings preserves the posterior distribution in Bayesian regression models based on p-generalized normal distributions. Specifically, we bound the ℓ_p Wasserstein distance between the original posterior and its counterpart that is defined only on the considerably smaller embedding. The outline is similar to the ℓ_2 case.

3.2.1 Maximum likelihood approximation

As in the previous section we begin with an approximation to the maximum likelihood estimate. Note however that this might not be unique. An example is a linear subspace spanned by the columns of X that fully contains a facet of the ℓ_1 ball of optimal radius centered at Y. All points on this facet have equal ℓ_1 distance to the dependent variable and thus maximize the likelihood function, i.e., are modes of the distribution. Our further calculations will focus on any mode of the p-generalized normal distribution induced by the data. For the multivariate distributions of the regression parameters the modes do not necessarily coincide with their means.

Now consider the maximum likelihood problem for frequentist ℓ_p-regression. Similar proof techniques previously appeared in [35] and [24] in the context of ℓ_1 and ℓ_2 regression, respectively.

Lemma 3.2.1. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the columnspace of $[X,Y]$. Let $\gamma \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|X\beta - Y\|_p$ and similarly we let $\nu \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|\Pi(X\beta - Y)\|_p$. Then*

$$\|X\nu - Y\|_p \leq (1+\varepsilon)\ \|X\gamma - Y\|_p .$$

Proof. Let $[X,Y] = U\Sigma V^T$ be the singular value decomposition of $[X,Y]$. Now define $\vartheta_1 = \Sigma V^T[\gamma^T, -1]^T$ and $\vartheta_2 = \Sigma V^T[\nu^T, -1]^T$. Note that $U\vartheta_1 = X\gamma - Y$ and $U\vartheta_2 = X\nu - Y$. We have

$$(1-\frac{\varepsilon}{3})\ \|U\vartheta_2\|_p \leq \|\Pi U\vartheta_2\|_p \leq \|\Pi U\vartheta_1\|_p \leq (1+\frac{\varepsilon}{3})\ \|U\vartheta_1\|_p .$$

The middle inequality follows from the optimality of ν in the embedded subspace. The other two inequalities are direct applications of the subspace embedding property, cf. (2.12). Now, after rearranging and resubstituting we conclude

$$\|X\nu - Y\|_p \leq \left(\frac{1+\frac{\varepsilon}{3}}{1-\frac{\varepsilon}{3}}\right) \|X\gamma - Y\|_p \leq (1+\varepsilon)\ \|X\gamma - Y\|_p . \qquad \square$$

3.2.2 Embedding the ℓ_p-likelihood

We analyze the distributions proportional to the likelihood functions $q \propto \mathcal{L}(\beta|X,Y)$ and $q' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$ which are now based on p-generalized normal distributions. We would like to conduct the analysis in terms of their modes and pth moments separately. This will be helpful for bounding their ℓ_p-Wasserstein distance. To this end, we make the following observation, which can be derived via a generalized triangle inequality, cf. [40, 111].

Observation 4. *Let $Z_1, Z_2 \in \mathbb{R}^d$ be random variables and let m_1, m_2 be values that they can attain. Let $Z_1^m = Z_1 - m_1$, respectively, $Z_2^m = Z_2 - m_2$ be their centered counterparts. Then for any fixed $p \in [1, \infty)$ it holds that*

$$\mathbb{E}\left[\|Z_1 - Z_2\|_p^p\right] \leq 2^{p-1}\left(\|m_1 - m_2\|_p^p + \mathbb{E}\left[\|Z_1^m - Z_2^m\|_p^p\right]\right).$$

Proof. By convexity of the p-norm we can apply Jensen's inequality, which yields for arbitrary $x, y \in \mathbb{R}^d$ a generalized triangle inequality,

$$\|x+y\|_p^p = 2^p \left\|\frac{x+y}{2}\right\|_p^p \leq 2^p \frac{\|x\|_p^p + \|y\|_p^p}{2} = 2^{p-1}(\|x\|_p^p + \|y\|_p^p).$$

We thus have

$$\begin{aligned}
\mathbb{E}\left[\|Z_1 - Z_2\|_p^p\right] &= \mathbb{E}\left[\|Z_1 - m_1 + m_1 - Z_2 + m_2 - m_2\|_p^p\right] \\
&= \mathbb{E}\left[\|Z_1^m - Z_2^m + m_1 - m_2\|_p^p\right] \\
&\leq 2^{p-1}\,\mathbb{E}\left[\|Z_1^m - Z_2^m\|_p^p + \|m_1 - m_2\|_p^p\right] \\
&= 2^{p-1}\left(\mathbb{E}\left[\|Z_1^m - Z_2^m\|_p^p\right] + \|m_1 - m_2\|_p^p\right).
\end{aligned}$$

□

Observation 4 allows us to analyze the distance of q and q' in terms of their modes and pth moments separately, as desired. Recall that our analysis holds for arbitrary modes γ and ν of q and q', respectively. Hence, we start by bounding the distance of any of their modes.

Lemma 3.2.2. *Given $X \in \mathbb{R}^{n \times d}, Y \in \mathbb{R}^n$, let Π be an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the columnspace of $[X, Y]$. Now let $\gamma \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \|X\beta - Y\|_p$ and similarly define $\nu \in \operatorname{argmin}_{\beta \in \mathbb{R}^d} \|\Pi(X\beta - Y)\|_p$. Then*

$$\|\gamma - \nu\|_p \leq \frac{2+\varepsilon}{\sigma_{\min}^{(p)}(X)} \|X\gamma - Y\|_p .$$

Proof. Using the ℓ_p singular values, see Definition 2.1.5, the triangle inequality and Lemma 3.2.1, we have

$$\begin{aligned}\|\gamma-\nu\|_p &\le \frac{1}{\sigma^{(p)}_{\min}(X)}\,\|X(\gamma-\nu)\|_p\\ &\le \frac{1}{\sigma^{(p)}_{\min}(X)}\left(\|X\gamma-Y\|_p+\|X\nu-Y\|_p\right)\\ &\le \frac{2+\varepsilon}{\sigma^{(p)}_{\min}(X)}\,\|X\gamma-Y\|_p\,. \qquad\square\end{aligned}$$

Note that the application of the triangle inequality is tight for the ℓ_1 norm (similarly for ℓ_∞) since the columns of X can be aligned in such a way, that its columnspace touches $\mathrm{B}_1(Y,\|X\gamma-Y\|_p)$ at one vertex v and at the same time passes through another vertex w of $\mathrm{B}_1(Y,(1+\varepsilon)\,\|X\gamma-Y\|_p)$ that is adjacent to the vertex v' corresponding to v in the expanded ball. Without imposing further assumptions, the ℓ_1 distance between $X\gamma$ and $X\nu$ can thus be exactly $(2+\varepsilon)\,\|X\gamma-Y\|_1$. It remains open for now if one can parametrize the loss via the norm parameter p to obtain a stronger inequality.

Observation 4 allows us to assume w.l.o.g. $\gamma=\nu=0$ when we consider the distance of q and q' in terms of their pth moments. It remains to derive a bound on $\inf\mathbb{E}\left[\|Z_1^m-Z_2^m\|_p^p\right]$, where the two points are drawn from a joint distribution whose marginals are the original likelihood distribution and its counterpart derived from the subspace embedding, each centered at one of their modes. The infimum minimizes over all possible choices. We bound this quantity by choosing a particular joint distribution and bounding the expected distance with respect to the pth moment for this particular instance.

Lemma 3.2.3. *Given $X\in\mathbb{R}^{n\times d}, Y\in\mathbb{R}^n$, let Π be an $(\frac{\varepsilon}{3},p)$-subspace embedding for the columnspace of $[X,Y]$. Let $q\propto\mathcal{L}(\beta|X,Y)$ and $q'\propto\mathcal{L}(\beta|\Pi X,\Pi Y)$. Let Z_1^m,Z_2^m be the mode-centered versions of the random variables $Z_1\sim q$ and $Z_2\sim q'$ that are distributed according to q and q' respectively. Then we have*

$$\inf_{\lambda\in\Lambda(q,q')}\mathbb{E}_\lambda\left[\|Z_1^m-Z_2^m\|_p^p\right]\le\left(\frac{2+\varepsilon}{\sigma^{(p)}_{\min}(X)}\right)^p\mathbb{E}_q\left[\|XZ_1^m-Y\|_p^p\right].$$

Proof. We construct a bijective map $g\colon\mathbb{R}^d\to\mathbb{R}^d$, that implicitly defines a joint distribution $\lambda\in\Lambda(q,q')$ of q and q' via Dirac's delta function, as described in Section 2.2. The joint distribution deterministically maps points from one distribution to another in such a way that we can bound the distance of every pair of points.

The high level idea is similar as in the proof of Lemma 3.1.3. We want g to map points to each other that lie in the same direction from the common mode and have the same ℓ_p-distance to their corresponding dependent variables Y and ΠY, and thus have the same density up to normalization. But unlike the $p = 2$ case, we cannot work completely inside the d dimensional spaces.

So for each $\rho \geq 0$, g maps the points $\alpha, \beta \in \mathbb{R}^d$ to each other such that

1. $\|X\alpha - Y\|_p = \|\Pi(X\beta - Y)\|_p = \rho$

2. $\exists \tau \geq 0\colon \alpha = \tau\beta$.

By the first property, α and β have the same ℓ_p distance from Y. They thus, share the same p-generalized normal density up to normalization.

It is important to note that the ℓ_p balls $\mathrm{B_p}(Y, \rho)$ and $\mathrm{B_p}(\Pi Y, \rho)$ are convex sets. Thus, their intersections with the linear columnspaces spanned by X and ΠX are also convex. Since the variables are centered at any fixed modes of their distributions we can assume that both of these modes equal 0 and more importantly, lie within these convex sets.

The second property ensures that α and β share the same orientation but have a possibly different scale, where τ quantifies the scaling factor. Consequently they lie on a ray starting from the origin and by convexity any such ray intersects the surface of each convex set exactly once. We have thus defined the bijection that we need here.

It remains to bound the distance of each pair of points α, β mapped to each other. This is done by taking a detour into the ambient spaces which are connected via the subspace embedding.

$$\begin{aligned}
\sigma_{\min}^{(p)}(X)\,\|\alpha - \beta\|_p &\leq \|X(\alpha - \beta)\|_p \\
&\leq \|X\alpha - Y\|_p + \|X\beta - Y\|_p \\
&\leq \|X\alpha - Y\|_p + (1+\varepsilon)\,\|\Pi(X\beta - Y)\|_p \\
&= (2+\varepsilon)\,\|X\alpha - Y\|_p
\end{aligned}$$

We can thus conclude

$$\begin{aligned}
\inf_{\lambda' \in \Lambda(q,q')} \mathbb{E}_{\lambda'}\left[\|Z_1^m - Z_2^m\|_p^p\right] &\leq \mathbb{E}_{\lambda}\left[\|Z_1^m - Z_2^m\|_p^p\right] \\
&\leq \left(\frac{2+\varepsilon}{\sigma_{\min}^{(p)}(X)}\right)^p \mathbb{E}_{\lambda}\left[\|XZ_1^m - Y\|_p^p\right] = \left(\frac{2+\varepsilon}{\sigma_{\min}^{(p)}(X)}\right)^p \mathbb{E}_{q}\left[\|XZ_1^m - Y\|_p^p\right]. \qquad \square
\end{aligned}$$

Combining our results we get the following lemma.

Lemma 3.2.4. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the columnspace of $[X, Y]$. Let $q \propto \mathcal{L}(\beta|X,Y)$ and $q' \propto \mathcal{L}(\beta|\Pi X, \Pi Y)$. Let Z_1^m, Z_2^m be the mode-centered versions of the random variables $Z_1 \sim q$ and $Z_2 \sim q'$ that are distributed according to q and q' respectively. Then*

$$\mathcal{W}_p(q,q') \leq \frac{4+2\varepsilon}{\sigma_{\min}^{(p)}(X)} \left(\|X\gamma - Y\|_p + \mathbb{E}_q\Big[\|XZ_1^m - Y\|_p^p\Big]^{\frac{1}{p}} \right).$$

Proof. The lemma follows from Definition 2.2.1, Observation 4, Lemma 3.2.2 and Lemma 3.2.3. Using the notation from these results, we have

$$\begin{aligned}
\mathcal{W}_p(q,q') &= \inf_{\lambda' \in \Lambda(q,q')} \mathbb{E}_{\lambda'}\left[\|Z_1 - Z_2\|_p^p\right]^{\frac{1}{p}} \\
&\leq 2^{1-\frac{1}{p}} \inf_{\lambda' \in \Lambda(q,q')} \mathbb{E}_{\lambda'}\left[\|m_1 - m_2\|_p^p + \|Z_1^m - Z_2^m\|_p^p\right]^{\frac{1}{p}} \\
&\leq 2^{1-\frac{1}{p}} \left(\|\gamma - \nu\|_p + \inf_{\lambda' \in \Lambda(q,q')} \mathbb{E}_{\lambda'}\left[\|Z_1^m - Z_2^m\|_p^p\right]^{\frac{1}{p}} \right) \\
&\leq 2 \left(\|\gamma - \nu\|_p + \inf_{\lambda' \in \Lambda(q,q')} \mathbb{E}_{\lambda'}\left[\|Z_1^m - Z_2^m\|_p^p\right]^{\frac{1}{p}} \right) \\
&\leq \frac{4+2\varepsilon}{\sigma_{\min}^{(p)}(X)} \left(\|X\gamma - Y\|_p + \mathbb{E}_q\Big[\|XZ_1^m - Y\|_p^p\Big]^{\frac{1}{p}} \right).
\end{aligned}$$

□

3.2.3 Bayesian posterior approximation

So far we have shown that using subspace embeddings to compress a given data set for regression yields a good approximation to the likelihood. Again, in a Bayesian regression setting, Lemma 3.2.4 already implies a similar approximation error for the posterior distribution if the prior for β is a uniform distribution over $\mathbb{R}^d$. As in the ℓ_2 case, though the prior is degenerate, the posterior will be a proper distribution after normalizing. The extension to arbitrary p-generalized normal priors $p_{\text{pre}}(\beta)$ is similar to the case of normal distributions.

Let m be a mode parameter and consider an arbitrary $S \in \mathbb{R}^{d\times d}$ of full rank to define the p-generalized normal prior distribution. Now, the posterior distribution is given by

$$\begin{aligned}
&p_{\text{post}}(\beta|X,Y) \propto \mathcal{L}(\beta|X,Y) \cdot p_{\text{pre}}(\beta) \\
&= \left(\frac{p}{2\varsigma\Gamma(1/p)}\right)^n \cdot \exp\left(-\frac{1}{\varsigma}\|X\beta - Y\|_p^p\right) \cdot \left(\frac{p}{2\varsigma\Gamma(1/p)}\right)^d \cdot \exp\left(-\frac{1}{\varsigma}\|S(\beta - m)\|_p^p\right).
\end{aligned}$$

Thus, we know that up to some constants that are independent of β, the exponent of the posterior can be described by

$$\|X\beta - Y\|_p^p + \|S(\beta - m)\|_p^p \tag{3.4}$$

which contains all defining parameters of the posterior distribution. Now let

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \text{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

With these definitions we can rewrite Equation (3.4) above as $\|Z\beta - z\|_p^p$. This, again, can be treated as a (frequentist) regression problem to which we can apply Lemma 3.2.4 using a subspace embedding for the columnspace of $[Z, z]$ instead of only embedding $[X, Y]$. Again, it is not necessary to do this explicitly. An embedding of the data is sufficient by Lemma 2.5.6. This yields a proper approximation of the posterior distribution.

Theorem 3.2.5. *Given $X \in \mathbb{R}^{n\times d}, Y \in \mathbb{R}^n$, let Π be an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the columnspace of $[X, Y]$. Let $p_{\text{pre}}(\beta)$ be an arbitrary p-generalized normal distribution with mode $m \in \mathbb{R}^d$ and variance parameters $\varsigma \in \mathbb{R}_{>0}$ and $S \in \mathbb{R}^{d\times d}$. Let*

$$Z = \begin{bmatrix} X \\ S \end{bmatrix} \quad \textit{and} \quad z = \begin{bmatrix} Y \\ Sm \end{bmatrix}.$$

Let $\mu \in \operatorname{argmin}_{\beta\in\mathbb{R}^d} \|Z\beta - z\|_p$ be a posterior mode. Let $q \propto \mathcal{L}(\beta|X, Y) \cdot p_{\text{pre}}(\beta)$ and $q' \propto \mathcal{L}(\beta|\Pi X, \Pi Y) \cdot p_{\text{pre}}(\beta)$. Let Z_1^m be the mode-centered version of the random variable $Z_1 \sim q$ distributed according to q. Then

$$\mathcal{W}_p(q, q') \leq \frac{4 + 2\varepsilon}{\sigma_{\min}^{(p)}(Z)} \left(\|Z\mu - z\|_p + \mathbb{E}_q\left[\|XZ_1^m - Y\|_p^p\right]^{\frac{1}{p}} \right).$$

Proof. From our previous reasoning we know that approximating the posterior distribution can be reduced to approximating a likelihood function that is defined in terms of the data as well as the parameters of the prior distribution. This has been shown by rewriting Equation (3.4) above as $\|Z\beta - z\|_p^p$. For that reason, we can apply Lemma 3.2.4 to get the desired result if we are given an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the space spanned by the columns of Z and z. Using Lemma 2.5.6 we know that to this end, it is sufficient to use an $(\frac{\varepsilon}{3}, p)$-subspace embedding for the columnspace of $[X, Y]$ independent of the parameters that define the prior distribution. □

We note that the bounds for ℓ_p spaces are much worse than in the ℓ_2 case. This affects the size of the embeddings (see Section 2.5.2) as well as the approximation guarantees. We have pointed out at several steps, that the loss incurred in our inequalities is tight if we do not impose additional assumptions or parameterize the inequalities. This means in particular, that we can construct worst-case instances attaining these bounds.

4 Coresets for dependency networks and generalized linear models

4.1 Core dependency networks

As we have learned in Section 2.4.3, learning dependency networks consists of determining the conditional probability distributions from a given set of n training instances $x_i \in \mathbb{R}^d$ representing the rows of the data matrix $X \in \mathbb{R}^{n\times d}$ over d variables. If we assume that each of these distributions $p(x^{(i)} | \mathbf{pa}_i)$ is parametrized as a generalized linear model [108], this means we have to estimate the parameters $\beta^{(i)}$ of the generalized linear model associated with each variable $X^{(i)}$. This completely determines the local distributions $p(x^{(i)} | \mathbf{pa}_i)$. These will possibly depend on all other variables in the network, and these dependencies define the structure of the network revealed in the regression analysis. This view of training dependency networks as fitting d generalized linear models to the data allows us to develop *core dependency networks*. On a high level, the idea is to construct a coreset and train a dependency network over certain members of the family of generalized linear models on the coreset. For Gaussian dependency networks, we can do this via ε-subspace embeddings, which can be obtained via random projection techniques or by sampling and reweighting a small number of input points. For other generalized linear models, we will concentrate on the sampling approach, since it turns out to be more versatile when extending to more general models and it preserves special features of the data, like integrality when dealing with count data.

4.1.1 Coresets for Gaussian dependency networks

Consider $(\mathcal{G}, \Psi)$, a Gaussian dependency network, i.e., a collection of linear regression models according to normal distributions,

$$\Psi = \left\{ p_i(X^{(i)} | X^{\setminus i}, \beta^{(i)}) = \mathrm{N}\left(X^{\setminus i}\beta^{(i)}, \sigma^2 \right) \mid i \in [d] \right\}$$

on an arbitrary digraph structure $\mathcal{G}$ [77].

The logarithm of the *(pseudo-)likelihood* [20] of the above model is given by

$$\ln \mathcal{L}(\Psi) = \ln \prod\nolimits_{i=1}^{d} p_i = \sum\nolimits_{i=1}^{d} \ln p_i.$$

A maximum likelihood estimate can be obtained by maximizing this function with respect to $\beta = (\beta^{(1)}, \ldots, \beta^{(d)})$ which is equivalent to minimizing the cost function of Gaussian dependency networks

$$f_G(X, \beta) = \sum\nolimits_{i=1}^{d} \left\| X^{\setminus i}\beta^{(i)} - X^{(i)} \right\|_2^2.$$

In general, we can compute a coreset for each individual term in the sum and thus construct a coreset for every single generalized linear model, but the total size of the collection of coresets will be multiplied by a factor of d. Since most coreset constructions are randomized, another $\log d$ factor might be necessary to amplify the failure probability from η to $\frac{\eta}{d}$ and union bound over the d coresets, to finally obtain a total failure probability of η again.

Indeed, when the local distributions are all normal distributions, we can use ε-subspace embeddings Π_i for the columnspaces of the matrices $X^{\setminus i}$. This is the structural key property to show that $\Pi_i[X^{\setminus i}, X^{(i)}]$ are coresets for the normal linear regression models and their collection is a coreset for the Gaussian dependency network.

However, it is noteworthy that computing one single coreset for the columnspace of X is sufficient in the case of normal distributions, rather than computing d coresets for the d different subspaces spanned by $X^{\setminus i}$.

Theorem 4.1.1. *Given $X \in \mathbb{R}^{n\times d}$, let Π be an ε-subspace embedding for the columnspace of X. Then ΠX is a $(1 \pm \varepsilon)$-coreset for the Gaussian dependency network cost function.*

Proof. Fix an arbitrary $\beta = (\beta^{(1)}, \ldots, \beta^{(d)}) \in \mathbb{R}^{d(d-1)}$. Consider the affine map $\Phi : \mathbb{R}^{d-1} \times [d] \to \mathbb{R}^d$, defined by $\Phi(\beta^{(i)}) = I_d^{\setminus i}\beta^{(i)} - e_i$. The map Φ has the effect of extending its argument from $d-1$ to d dimensions by inserting a -1 entry at position i and leaving the other entries in their original order. Let $\gamma^{(i)} = \Phi(\beta^{(i)}) \in \mathbb{R}^d$. Note that for each $i \in [d]$ we have

$$X\gamma^{(i)} = X\Phi(\beta^{(i)}) = X^{\setminus i}\beta^{(i)} - X^{(i)}, \tag{4.1}$$

and each $\gamma^{(i)}$ is a vector in $\mathbb{R}^d$. Thus, the triangle inequality and the universal quantifier in Definition 2.5.2 guarantee that

$$\left| \sum\nolimits_{i=1}^{d} \left\| \Pi X\gamma^{(i)} \right\|_2^2 - \sum\nolimits_{i=1}^{d} \left\| X\gamma^{(i)} \right\|_2^2 \right| = \left| \sum\nolimits_{i=1}^{d} \left(\left\| \Pi X\gamma^{(i)} \right\|_2^2 - \left\| X\gamma^{(i)} \right\|_2^2 \right) \right|$$

$$
\begin{aligned}
&\leq \sum\nolimits_{i=1}^{d} \left| \left\| \Pi X\gamma^{(i)} \right\|_2^2 - \left\| X\gamma^{(i)} \right\|_2^2 \right| \\
&\leq \sum\nolimits_{i=1}^{d} \varepsilon \left\| X\gamma^{(i)} \right\|_2^2 \; = \; \varepsilon \sum\nolimits_{i=1}^{d} \left\| X\gamma^{(i)} \right\|_2^2 .
\end{aligned}
$$

Resubstituting Identity (4.1) yields the proposition. □

From Theorem 4.1.1 it is straightforward to show that the minimizer found for the coreset is a good approximation of the minimizer for the original data, see Scheme (2.6).

Corollary 4.1.2. *Let C be a $(1 \pm \varepsilon)$-coreset of X for the Gaussian dependency network cost function, let $\tilde{\beta} \in \operatorname{argmin}_{\beta \in \mathbb{R}^{d(d-1)}} f_G(C, \beta)$. Then it holds that*

$$f_G(X, \tilde{\beta}) \leq (1 + 4\varepsilon) \min_{\beta \in \mathbb{R}^{d(d-1)}} f_G(X, \beta).$$

Proof. Let $\beta^* \in \operatorname{argmin}_{\beta \in \mathbb{R}^{d(d-1)}} f_G(X, \beta)$. Then

$$
\begin{aligned}
f_G(X, \tilde{\beta}) \leq \frac{1}{1-\varepsilon} f_G(C, \tilde{\beta}) \quad &\leq \frac{1}{1-\varepsilon} f_G(C, \beta^*) \\
\leq \frac{1+\varepsilon}{1-\varepsilon} f_G(X, \beta^*) &\leq (1 + 4\varepsilon) f_G(X, \beta^*).
\end{aligned}
$$

The first and third inequalities are direct applications of the coreset property, the second holds by optimality of $\tilde{\beta}$ for the coreset, and the last follows from $\varepsilon < \frac{1}{2}$. □

In Section 3.1, we have shown for (Bayesian) linear regression models that the entire multivariate normal likelihood distribution over the parameter space is approximately preserved by ε-subspace embeddings, which in particular applies to the local distributions in the dependency networks studied here. This suggests that the coreset yields a useful approximation to Markov Chain Monte Carlo sampling from Gaussian dependency networks via sampling from the local distributions, as in the pseudo-Gibbs sampler in [77].

Naturally, the question arises whether coresets exist for dependency networks over all generalized linear models? In the following we would like to extend our coreset results to models other than the normally distributed model. Specifically we study extensions to Poisson regression to model count data as well as to logistic regression to model binary Bernoulli distributed data.

4.2 On coresets for Poisson regression

We first deal with the Poisson regression model. Recall that we are given a data matrix $X \in \mathbb{R}^{n\times d}$, and labels $Y \in \mathbb{N}_0^n$ and the cost function of the Poisson regression problem [108, 143] is

$$\ell(\beta) = \ell(\beta|X, Y) = \sum\nolimits_{i=1}^{n} \exp(x_i\beta) - y_i \cdot x_i\beta + \ln(y_i!).$$

The problem consists in minimizing this function over $\beta \in \mathbb{R}^d$ to obtain a maximum likelihood estimate for the parameters. Again we would like to construct a coreset as a means for reducing the size of the input and thus save time and space in the optimization task, cf. Scheme (2.6).

4.2.1 Lower bound

Unfortunately, it turns out that there exists no coreset construction for the problem of Poisson regression, which implies the result for Poisson dependency networks [71]. We show this formally by reduction from the communication complexity problem indexing, see Section 2.5.4.

Theorem 4.2.1. *Let Σ_D be a data structure for $D = [X, Y] \in \mathbb{R}^{n\times(d+1)}, d \geq 3$ that approximates likelihood queries $\Sigma_D(\beta)$ for Poisson regression, such that for some $\varphi \geq 1$*

$$\forall\beta \in \mathbb{R}^d : \varphi^{-1} \cdot \ell(\beta|D) \leq \Sigma_D(\beta) \leq \varphi \cdot \ell(\beta|D).$$

If $\varphi < \frac{\exp(\frac{n}{4})}{2n^2}$ then Σ_D requires $\Omega(n)$ bits of memory.

Proof. We reduce from the indexing problem for which we know by Theorem 2.5.18 it has one-way randomized communication complexity $R_{1/3}(IND) \in \Omega(n)$. We construct a protocol as follows. Alice is given a vector $b \in \{0, 1\}^n$. She produces for every i with $b_i = 1$ the points $x_i = (r \cdot \omega^i, -1) \in \mathbb{R}^3$, where $\omega^i, i \in \{0, \ldots, n-1\}$ denote the nth unit roots in the plane, i.e., the vertices of a regular n-polygon in canonical order. We set the radius to $r = n/(1 - \cos(\frac{2\pi}{n})) \leq n^3$. The corresponding counts are set to $y_i = 1$. She builds and sends Σ_D of size $s(n)$ to Bob, whose task is to guess the bit b_j. He chooses to query $\beta = (\omega^j, r \cdot \cos(\frac{2\pi}{n})) \in \mathbb{R}^3$. Note that this affine hyperplane separates $r \cdot \omega^j$ from the other scaled unit roots since it passes exactly through $r \cdot \omega^{(j-1) \bmod n}$ and $r \cdot \omega^{(j+1) \bmod n}$. Also, all points are within distance $2r$ from each other by construction and consequently from the hyperplane. Thus, $-2r \leq x_i\beta \leq 0$ for all $i \neq j$.

If $b_j = 0$, then x_j does not exist and the cost is at most

$$\ell(\beta) = \sum\nolimits_{i=1}^{n} \exp(x_i\beta) - y_i \cdot x_i\beta + \ln(y_i!) \leq \sum\nolimits_{i=1}^{n} 1 + 2r + 1 \leq 2n + 2nr \leq 4n^4 \ .$$

If $b_j = 1$ then x_j is in the expensive halfspace and at distance exactly

$$\begin{aligned} x_j\beta &= (r\omega^j)^T\omega^j - r \cdot \cos\left(\frac{2\pi}{n}\right) \\ &= r \cdot \left(1 - \cos\left(\frac{2\pi}{n}\right)\right) \ = \ n \end{aligned}$$

So the cost is bounded below by $\ell(\beta) \geq \exp(n) - n + 1 \geq \exp(\frac{n}{2})$.

Given $\varphi < \frac{\exp(\frac{n}{4})}{2n^2}$, Bob can distinguish these two cases based on the data structure only, by deciding whether $\Sigma_D(\beta)$ is strictly smaller or larger than $\exp(\frac{n}{4}) \cdot 2n^2$. Consequently $s(n) \in \Omega(n)$, since this solves the indexing problem. □

Note that the bound is given in bit complexity, and that it holds already in $d = 3$ dimensions. Restricting the data structure to a coreset consisting of points in $\mathbb{R}^d$ and assuming a point can be expressed in $O(d \log n)$ bits, this means we can still give a lower bound of $k \in \Omega(\frac{n}{\log n})$ points.

Corollary 4.2.2. *Every coreset of $D = [X, Y] \in \mathbb{R}^{n\times(d+1)}$ consisting of points in $\mathbb{R}^{(d+1)}$, for $d \geq 3$ for Poisson regression with approximation factor $\varphi < \frac{\exp(\frac{n}{4})}{2n^2}$ as in Theorem 4.2.1 must comprise at least $k \in \Omega(\frac{n}{\log n})$ points.*

Proof. The details are the same as in the proof of Theorem 4.2.1. If we had a coreset construction with $o(\frac{n}{\log n})$ points, we could create a protocol for the indexing problem as follows. Alice computes a coreset for her point set and sends it to Bob. Bob evaluates the point β depending on his index. He can thus solve indexing using $o(n)$ communication, which contradicts Theorem 2.5.18. So Alice's coreset cannot exist. □

4.2.2 Approximation for count data

So far, we have a quite pessimistic view on extending core dependency networks beyond normal distributions. In the linear regression setting, where the cost is measured in squared Euclidean distance, the number of important points, i.e., having significantly large leverage scores, is bounded essentially by $O(d)$. This is implicit in the original early works [44, 45] and has been explicitly formalized later [32, 103] partly in the more general context of sensitivity sampling. It is crucial to understand that this is an inherent property of the

Euclidean norm function, and thus holds for arbitrary data. For the Poisson generalized linear model, in contrast, we have shown that its cost function does not come with such properties from scratch. We constructed a worst case scenario, where basically every single input point may be important for the model and needs to appear or be represented implicitly in the coreset. This implies that it must be large.

Usually, this is not the case with statistical models, where the data is assumed to be generated i.i.d. from some generating distribution that fits the model assumptions. Consider for instance a data reduction for linear regression via leverage score sampling vs. uniform sampling. It was shown that given the data follows the model assumption of a normal distribution, the two approaches behave very similarly. Or, to put it another way, the leverage scores are quite uniform. In the presence of more and more outliers generated by the heavier tails of t-distributions, sampling via leverage scores increasingly outperforms uniform sampling [106].

The Poisson model (2.3) though being the standard model for count data, suffers from its inherent limitation on equidispersed data since $\mathbb{E}[Y_i|x_i] = \mathbb{V}[Y_i|x_i] = \exp(x_i\beta)$. Count data, however, is often overdispersed especially for large counts. This may be due to unobserved variables or problem specific heterogeneity and contagion-effects. The Poisson log-normal model is known to be inferior for data which specifically follows the Poisson model, but turns out to be more powerful in modeling the effects that can not be captured by the simple Poisson model. It has wide applications for instance in econometric elasticity problems. We review the Poisson log-normal model for count data [143]

$$\begin{aligned} Y_i &\sim \mathrm{Poi}(\lambda_i), \\ \lambda_i &= \exp(x_i\beta)u_i = \exp(x_i\beta + v_i), \\ v_i \;=\; \ln u_i &\sim \mathrm{N}\left(\mu, \sigma^2\right). \end{aligned}$$

A natural choice for the parameters of the log-normal distribution is $\mu = -\frac{\sigma^2}{2}$ in which case we have

$$\begin{aligned} \mathbb{E}[Y_i|x_i] &= \exp(x_i\beta + \mu + \sigma^2/2) \\ &= \exp(x_i\beta)\,, \\ \mathbb{V}[Y_i|x_i] &= \mathbb{E}[y_i|x_i] + (\exp(\sigma^2) - 1)\mathbb{E}[Y_i|x_i]^2\,. \end{aligned}$$

It follows that $\mathbb{V}[Y_i|x_i] = \exp(x_i\beta) + \Omega(\exp(x_i\beta)^2) > \exp(x_i\beta)$, where a constant σ^2 that is independent of x_i, controls the amount of overdispersion. Taking the limit for $\sigma \to 0$

we arrive at the simple model (2.3), since the distribution of $v_i = \ln u_i$ tends to δ, the deterministic Dirac delta distribution which puts all mass on 0. The inference might aim for the Poisson log-normal model directly as in [149], or it can be performed by maximum likelihood estimation of the simple Poisson model. The latter provides a consistent estimator as long as the log-linear mean function is correctly specified, even if higher moments do not possess the limitations inherent in the simple Poisson model [143].

After this brief review on the count modeling perspective, we can sum up that it is crucial for the consistency of the estimator in any Poisson model to preserve the log-linear mean function. Moreover, modeling count data in a Poisson log-normal model gives us intuition why an ε-subspace embedding for the columnspace of the data matrix may be used to capture the underlying linear model accurately although the universal approximation guarantee of strong coresets is out of reach. To this end, slightly abusing notation, we define $\ln w$ for a vector w to be the vector that results from a entry-wise application of the natural logarithm function to w.

In the Poisson log-normal model, u follows a log-normal distribution. It thus holds for

$$\ln \lambda = X\beta + \ln u \; = \; X\beta + v,$$

that

$$v \sim \mathrm{N}\left(-\frac{\sigma^2}{2} \cdot \mathbf{1}, \sigma^2 I_n\right)$$

by independence of the observations, which implies

$$\ln \lambda \sim \mathrm{N}\left(X\beta - \frac{\sigma^2}{2} \cdot \mathbf{1}, \sigma^2 I_n\right).$$

Now, let $\zeta = \ln \lambda + \frac{\sigma^2}{2} \cdot \mathbf{1}$. We notice that this yields again an ordinary least squares problem

$$\min_{\beta \in \mathbb{R}^d} \|X\beta - \zeta\|_2^2$$

defined in the columspace of X which we know how to deal with using an ε-subspace embedding for X.

The complication here is that λ and consequently also ζ are only implicitly given in the data, but are not explicitly available. We thus cannot simply compute a maximum likelihood estimator via least squares regression. However, this indicates that using a sampling based ε-subspace embedding will yield a good approximation to the maximum likelihood estimator for the Poisson log-normal regression model, cf. Scheme (2.6). The sampling based approach is required to preserve the relationship of $Y_i \sim \mathrm{Poi}(\lambda_i)$ as well as integrality of the count values y_i. Both are necessary in the Poisson log-normal model.

4.3 On coresets for logistic regression

Another important generalized linear model is logistic regression. Recall that in the logistic regression problem we are given a data matrix $Z \in \mathbb{R}^{n \times d}$, and labels $Y \in \{-1, 1\}^n$ and the cost function of the logistic regression problem is [108]

$$\ell(\beta) = \ell(\beta|Z, Y) = \sum\nolimits_{i=1}^{n} \ln(1 + \exp(-Y_i Z_i \beta)).$$

From a learning and optimization perspective, this is the objective function that we aim to minimize over $\beta \in \mathbb{R}^d$ to obtain a maximum likelihood estimator for the parameter. For notational brevity we fold the labels Y_i as well as the factor -1 in the exponent into row vectors $x_i = -Y_i Z_i$ for all $i \in [n]$. In order to speed up the computation and to lower memory and storage requirements we would like to significantly reduce the number of observations without losing much information in the original data, see Scheme (2.6). To achieve this, our plan is to design a coreset construction for the objective function. For technical reasons, to obtain coresets via the sensitivity framework, we deal with a weighted version for weights $w \in \mathbb{R}^n_{>0}$, where each weight satisfies $w_i > 0$ and any positive scaling of the all ones vector $\mathbf{1}$ corresponds to the unweighted case. For brevity, let $g(z) = \ln(1 + \exp(z))$. The objective function becomes

$$f_w(X\beta) = \sum\nolimits_{i=1}^{n} w_i g(x_i\beta) = \sum\nolimits_{i=1}^{n} w_i \ln(1 + \exp(x_i\beta)).$$

4.3.1 Lower bound

We will again reduce from the indexing communication problem leveraging one-way randomized communication complexity of $R_{1/3}(IND) \in \Omega(n)$, cf. Theorem 2.5.18. In the following we will show why this implies that no strongly sublinear summaries or coresets for logistic regression can exist in general, even if we assume the points to lie in 3-dimensional Euclidean space.

Theorem 4.3.1. *Let Σ_D be a data structure for $D = [Z, Y] \in \mathbb{R}^{n\times(d+1)}, d \geq 3$ that allows insertions and approximates likelihood queries $\Sigma_D(\beta)$ for logistic regression, such that the optimum value of $\Sigma_D(\beta)$ can be found in finite time and for some $\varphi \geq 1$*

$$\forall \beta \in \mathbb{R}^d : \varphi^{-1} \cdot \ell(\beta|D) \leq \Sigma_D(\beta) \leq \varphi \cdot \ell(\beta|D).$$

If $\varphi < \infty$ then Σ_D requires $\Omega(n)$ bits of memory.

Proof. Assume we are given an instance of the indexing problem, i.e., Alice has a string $b \in \{0,1\}^n$ and Bob has an index $j \in [n]$. For each i with $b_i = 1$, she produces the points $z_i = (\omega^i, -1) \in \mathbb{R}^3$, where $\omega^i, i \in \{0, \dots, n-1\}$ denote the nth unit roots in the plane in canonical order, i.e., the vertices of a regular n-polygon of unit radius. Note that all of these points have equal Euclidean norm and hence any single point may be linearly separated from the others. All of Alice's points have label $y_i = 1$. She builds and communicates Σ_D of size $s(n)$ to Bob, whose task is to guess the bit b_j. Bob adds the point $z_n = ((1-\vartheta)\cdot\omega^j, -1) \in \mathbb{R}^3$ for a sufficiently small $\vartheta > 0$ with label $y_n = -1$. Bob now finds the optimum on Σ_D augmented by his point z_n labeled by y_n.

If Alice added z_j and hence $b_j = 1$ then the optimal solution for the original instance will have cost at least $\ln(2)$ since for any choice of β, there will be at least one misclassification with $\operatorname{sgn}(z_m\beta) \neq y_m$ which contributes $\ell(\beta) \geq \ln(1+\exp(-y_m z_m\beta)) \geq \ln(1+\exp(0)) = \ln(2)$. If, on the other hand, Alice did not add z_j and hence $b_j = 0$, then the two differently labeled point sets are linearly separable and the cost tends to 0. Distinguishing between these two cases, i.e. approximating the cost of logistic regression via Σ_D below a factor $\lim\limits_{x\to 0} \frac{\ln(2)}{x} = \infty$ solves the indexing problem. Consequently $s(n) \in \Omega(n)$. □

The same reduction also holds if Alice's message is restricted to consist of points forming a coreset. Hence, the following corollary holds, where we assume as before that a point can be expressed in $O(d\log n) \subseteq O(\log n)$ bits.

Corollary 4.3.2. *Every coreset of $D = [X, Y] \in \mathbb{R}^{n\times(d+1)}$ consisting of points in $\mathbb{R}^{(d+1)}$, for $d \geq 3$ for logistic regression with approximation factor $\varphi < \infty$ as in Theorem 4.3.1 must comprise at least $k \in \Omega(\frac{n}{\log n})$ points.*

Proof. If we had a coreset construction with $o(\frac{n}{\log n})$ points, we could give a protocol for the indexing problem as follows. Alice computes a coreset for her point set defined in the proof of Theorem 4.3.1 and sends it to Bob. Bob computes an optimal solution on the union of the coreset and his point. This solves indexing using $o(n)$ communication, which contradicts the lower bound given in Theorem 2.5.18. So Alice's coreset cannot exist. □

We note that independently, a linear lower bound appeared in [135] for sums of monotonic functions which includes logistic regression. The bound is based on a worst case instance to the sensitivity approach from [84]. Our lower bounds and theirs are incomparable. They show that if a coreset can only consist of input points it comprises the entire data set in the worst-case. We show that no coreset with $o(\frac{n}{\log n})$ can exist, irrespective of whether input points are used. While the distinction may seem minor, a number of coreset constructions in literature rely on non-input points, see e.g. [4, 55].

4.3.2 μ-complex data sets

We have shown that in general, there exists no sublinear summary or coreset construction that works for all data sets. For the sake of developing coreset constructions that work *reasonably well*, as well as conducting a formal analysis beyond worst-case instances, we introduce a measure μ that quantifies the complexity of compressing a given data set. Our analysis is parametrized with the value of μ, setting it into the light of beyond worst-case analysis [16, 125]. Recall that for a weight vector w, D_w denotes a diagonal matrix carrying the entries of w.

Definition 4.3.3. *Given a data set $X \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ and a vector $\beta \in \mathbb{R}^d$ let $(D_w X\beta)^-$ denote the vector comprising only the negative entries of $D_w X\beta$. Similarly let $(D_w X\beta)^+$ denote the vector of positive entries. We define for X weighted by w*

$$\mu_w(X) = \sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{\|(D_w X\beta)^+\|_1}{\|(D_w X\beta)^-\|_1}.$$

X weighted by w is called μ-complex if $\mu_w(X) \leq \mu$.

The size of our coreset constructions for logistic regression for a given μ-complex data set X will have low polynomial dependency on μ, d, ε but only sublinear dependency on its original size parameter n. So for μ-complex data sets having sufficiently small $\mu(X) \leq \mu$ we have the first $(1 \pm \varepsilon)$-coreset of provably sublinear size. The above definition implies, for $\mu(X) \leq \mu$, the following inequalities. The reader should keep in mind that for all $\beta \in \mathbb{R}^d$

$$\mu^{-1}\|(D_w X\beta)^-\|_1 \leq \|(D_w X\beta)^+\|_1 \leq \mu\|(D_w X\beta)^-\|_1 \,.$$

The parameter $\mu(X)$ has an intuitive interpretation and might be of independent interest. The odds of a binary random variable V are defined as $\frac{\mathbf{Pr}[V=1]}{\mathbf{Pr}[V=0]}$. The model assumption of logistic regression, see Equation 2.4, is that for every sample X_i, the logit, i.e. the logarithm of the odds, is a linear function of $X_i\beta$. For a candidate β, multiplying all odds and taking the logarithm is then exactly $\|X\beta\|_1$. Our definition thus relates the logit of the probabilities due to incorrectly classified points and the logit of the probabilities due to the correctly classified points. We say that the ratio between these two is upper bounded by μ. For logistic regression, assuming they are within some order of magnitude is not uncommon. One extreme is the (degenerate) case where the data set is exactly separable. Choosing β to parameterize a separating hyperplane for which $X\beta$ is all positive, implies that $\mu(X) = \infty$. Another case is when we have a large ratio between the number of positively and negatively labeled points which is a lower bound to μ. Note that under either of these conditions, logistic regression exhibits methodological weaknesses due to the separation or imbalance between the given classes, cf. [76, 79, 99, 109].

4.3.3 Coresets for μ-complex data

Our sampling based coreset constructions are obtained via the sensitivity sampling framework [25, 103], see Section 2.5.3. The main parameters that determine the size of a coreset are the VC dimension of the rangespace induced by functions under study, and their total sensitivity. First we derive sufficiently tight and efficiently computable upper bounds on the sensitivities. We will use Lemma 2.5.17, which bounds the VC dimension of logistic regression $\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \in O(dt \log t)$ where t will be related to the number of distinct weighted sensitivities, which we are going to bound essentially by $t \in O(\log n)$.

Base algorithm We show that sampling proportional to the square root of the ℓ_2-leverage scores augmented by $w_i / \sum_{j=1}^n w_j$ yields a coreset whose size is roughly linear in μ and the dependency on the input size is roughly $\sqrt{n}$. In what follows, let $\mathcal{W} = \sum_{i=1}^n w_i$.

We make a case distinction covered by lemmas 4.3.4 and 4.3.5. The intuition in the first case is that for a sufficiently large positive entry z, we have that $|z| \leq g(z) \leq 2|z|$. The lower bound holds even for all non-negative entries. Moreover, for μ-complex inputs we are able to relate the ℓ_1 norm of all entries to the positive ones, which will yield the desired bound, inspired by the techniques of [34] although adapted here for logistic regression.

Lemma 4.3.4. *Let $X \in \mathbb{R}^{n \times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of $D_w X$. If for index i the supreme β in the definition of sensitivities (2.14) satisfies $0.5 \leq x_i\beta$ then*

$$w_i g(x_i\beta) \leq 2(1+\mu)\|U_i\|_2 f_w(X\beta).$$

Proof. Let $D_wX = U\Sigma V^T = UR$, be derived from the singular value decomposition, where U is an orthonormal basis for the columnspace of D_wX. It follows from $0.5 \leq x_i\beta$ and monotonicity of g that

$$\begin{aligned}
w_i g(x_i\beta) &= w_i g\left(\frac{w_i x_i\beta}{w_i}\right) = w_i g\left(\frac{U_i R\beta}{w_i}\right)\\
&\leq w_i g\left(\frac{\|U_i\|_2\|R\beta\|_2}{w_i}\right) = w_i g\left(\frac{\|U_i\|_2\|UR\beta\|_2}{w_i}\right)\\
&= w_i g\left(\frac{\|U_i\|_2\|D_wX\beta\|_2}{w_i}\right) \leq w_i \frac{2}{w_i}\|U_i\|_2\|D_wX\beta\|_2\\
&\leq 2\|U_i\|_2\|D_wX\beta\|_1 \leq 2\|U_i\|_2(1+\mu)\|(D_wX\beta)^+\|_1\\
&= 2\|U_i\|_2(1+\mu)\sum_{j:w_jx_j\beta\geq 0} w_j|x_j\beta|\\
&\leq 2\|U_i\|_2(1+\mu)\sum_{j:x_j\beta\geq 0} w_j g(x_j\beta)\\
&\leq 2\|U_i\|_2(1+\mu)f_w(X\beta).
\end{aligned}$$

□

In the second case, the element under study is bounded above by a constant and thus the previous linear upper bound on $g(z)$ fails. We consider two sub cases. If there are a lot of contributions, which are not too small, and thus cost at least a constant each, then we can lower bound the total cost by a constant times their total weight. If on the other hand there are many very small negative values, then this implies again by μ complexity that the cost is within a μ fraction of the total weight.

Lemma 4.3.5. *Let $X \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. If for index i the supreme β in the definition of sensitivities (2.14) satisfies $x_i\beta \leq 0.5$ then*

$$w_i g(x_i\beta) \leq \frac{(20+\mu)w_i}{\mathcal{W}} f_w(X\beta).$$

Proof. Let $K^- = \{j \in [n] \mid x_j\beta \leq -2\}$ and $K^+ = \{j \in [n] \mid x_j\beta > -2\}$. Note that $g(-2) > 1/10$ and $g(x_i\beta) \leq g(0.5) < 1$. Also, $\sum_{j\in K^-} w_j + \sum_{j\in K^+} w_j = \mathcal{W}$.

Thus if $\sum_{j\in K^+} w_j \geq \frac{1}{2}\mathcal{W}$ then

$$f_w(X\beta) = \sum\nolimits_{j=1}^n w_j g(x_j\beta) \geq \frac{\sum_{j=1}^n w_j}{20} \geq \frac{\mathcal{W}}{20w_i}\cdot w_i g(x_i\beta).$$

If on the other hand $\sum_{j\in K^+} w_j < \frac{1}{2}\mathcal{W}$ then $\sum_{j\in K^-} w_j \geq \frac{1}{2}\mathcal{W}$. Thus

$$\begin{aligned} f_w(X\beta) &\geq \|(D_wX\beta)^+\|_1 \geq \|(D_wX\beta)^-\|_1/\mu \\ &\geq \left(2\cdot\frac{\sum_{j=1}^n w_j}{2}\right)\Big/\mu \geq \frac{\mathcal{W}}{\mu w_i}\cdot w_i g(x_i\beta). \end{aligned} \qquad \square$$

Combining both lemmas yields general upper bounds on the sensitivities that we can use as an importance sampling distribution. We also derive an upper bound on the total sensitivity that will be used to bound the sampling complexity.

Lemma 4.3.6. *Let $X \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n_{>0}$ be μ-complex. Let U be an orthonormal basis for the columnspace of D_wX. For each $i \in [n]$, the sensitivity of $g_i(\beta) = g(x_i\beta)$ for the weighted logistic regression function is bounded by $\varsigma_i \leq s_i = (20+2\mu)\cdot(\|U_i\|_2 + w_i/\mathcal{W})$. The total sensitivity is bounded by $\mathfrak{S} \leq S \leq 44\mu\sqrt{nd}$.*

Proof. From Lemma 4.3.4 and Lemma 4.3.5 we have for each i

$$\begin{aligned} \varsigma_i = \sup_\beta \frac{w_i g(x_i\beta)}{f_w(X\beta)} &\leq 2(1+\mu)\|U_i\|_2 + (20+\mu)\frac{w_i}{\mathcal{W}} \\ &\leq (20+2\mu)\Big(\|U_i\|_2 + \frac{w_i}{\mathcal{W}}\Big) \end{aligned}$$

From this, the second claim follows via the Cauchy-Schwarz inequality and using the fact that the Frobenius norm satisfies $\|U\|_F = \sqrt{d}$ due to orthonormality of U. We have

$$\begin{aligned} \mathfrak{S} = \sum\nolimits_{i=1}^n \varsigma_i &\leq (20+2\mu)\sum\nolimits_{i=1}^n \Big(\|U_i\|_2 + \frac{w_i}{\mathcal{W}}\Big) \\ &\leq 22\mu(\sqrt{n}\|U\|_F + 1) \leq 44\mu\sqrt{nd}\,. \end{aligned} \qquad \square$$

We combine our results into the following theorem on the data reduction, cf. Scheme (2.6).

Theorem 4.3.7. *Let $X \in \mathbb{R}^{n\times d}$ weighted by $w \in \mathbb{R}^n$ be μ-complex. Let $\omega = \frac{w_{\max}}{w_{\min}}$ be the ratio between the maximum and minimum weight in w. Let $\varepsilon \in (0, 1/2)$. There exists a $(1 \pm \varepsilon)$-coreset of X, w for logistic regression of size*

$$k \in O\left(\frac{\mu\sqrt{n}}{\varepsilon^2} d^{3/2} \log(\mu nd)\log(\omega n)\log\log(\omega n)\right).$$

Such a coreset can be constructed in time $O(nd^2)$ with probability $1 - 1/n^c$ for any absolute constant $c > 1$.

Proof. The algorithm computes the singular value decomposition $D_w X = U\Sigma V^T = UR$ of $D_w X$. Note that U is an orthonormal basis for the columnspace of $D_w X$. It uses the upper bounds s_i on the sensitivities from Lemma 4.3.6 but modifies them to obtain upper bounds s_i' such that each value $\frac{s_i'}{w_i}$ corresponds to $\frac{s_i}{w_i}$ but is rounded up to the closest power of two. It thus holds $s_i \leq s_i' \leq 2s_i$ for all $i \in [n]$. The input points are sampled proportional to the sampling probabilities $p_i = \frac{s_i'}{\sum_{j=1}^n s_j'}$. From Lemma 4.3.6 we know that $S' = \sum_{j=1}^n s_j' \leq 2S \in O(\mu\sqrt{nd})$.

In the proof of Theorem 2.5.14, see [57], the VC dimension bound is applied to a set of functions which are reweighted by $\frac{S' w_i}{s_i' k}$. We denote this set of functions $\mathcal{F}_{log}$.

Now note that the sensitivities satisfy

$$\frac{2}{w_{\min}} \geq \frac{2}{w_i} \geq \frac{s_i'}{w_i} \geq \frac{s_i}{w_i} = \sup_\beta \frac{g(x_i\beta)}{\sum_{j=1}^n w_j g(x_j\beta)} \overset{\beta=0}{\geq} \frac{1}{\sum_{j=1}^n w_j} \geq \frac{1}{n w_{\max}}. \tag{4.2}$$

Now note that k and S' are fixed values. Since the values $\frac{s_i'}{w_i}$ are scaled to powers of two, by (4.2) there can be at most $O(\log \frac{n w_{\max}}{w_{\min}}) \subseteq O(\log(\omega n))$ distinct values of $\frac{S' w_i}{s_i' k}$. Putting this into Lemma 2.5.17, we have $\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \in O(d \log(\omega n) \log\log(\omega n))$.

Putting all these pieces into Theorem 2.5.14 for error parameter $\varepsilon \in (0, 1/2)$ and failure probability $\eta = n^{-c}$, we have that a reweighted random sample of size

$$\begin{aligned} k &\in O\left(\frac{S'}{\varepsilon^2}\left(\Delta(\mathfrak{R}_{\mathcal{F}_{log}}) \log S' + \log\left(\frac{1}{\eta}\right)\right)\right) \\ &\subseteq O\left(\frac{\mu\sqrt{nd}}{\varepsilon^2}\left(d \log(\mu\sqrt{nd}) \log(\omega n) \log\log(\omega n) + \log\left(n^c\right)\right)\right) \\ &\subseteq O\left(\frac{\mu\sqrt{n}}{\varepsilon^2} d^{3/2} \log(\mu n d) \log(\omega n) \log\log(\omega n)\right) \end{aligned}$$

is a $(1 \pm \varepsilon)$ coreset with probability $1 - 1/n^c$ as claimed.

It remains to bound the running time. We can compute the singular value decomposition of $D_w X$ in time $O(nd^2)$, see [65]. Once U is available, we can inspect it row-by-row computing $\|U_i\|_2 + w_i/\mathcal{W}$ and give it as input together with x_i to k independent copies of a weighted reservoir sampler [28], which takes $O(nd)$ time. This gives a total running time of $O(nd^2)$ since the computations are dominated by the singular value decomposition. □

Recursive algorithm Here we develop a recursive algorithm, inspired by the recursive sampling technique of [33] for the Huber M-estimator, adapted here for logistic regression. This yields a better dependency on the input size. More specifically, we can diminish the leading $\sqrt{n}$ factor to only $\log^c(n)$ for an absolute constant c. A complication is that the parameter μ grows in the recursion, which we need to control. We thus have to deal with the ℓ_1 norm of all solutions set into the columnspace of the subsamples.

Our plan is to apply the Algorithm of Theorem 4.3.7 recursively. To do so, we need to ensure that after one stage of subsampling and reweighting, the resulting data set remains μ'-complex for a value μ' that is not too much larger than μ. To this end, we use the bound on the VC dimension of a range space induced by the ℓ_1 related family of functions $\mathcal{F}_{\ell_1} = \{h_i(\beta) = w_i|x_i\beta| \mid i \in [n]\}$. We know from Lemma 2.5.15 that $\Delta(\mathfrak{R}_{\mathcal{F}_{\ell_1}}) \leq 10(d+1)$. Applying the sensitivity sampling via Theorem 2.5.14 to $\mathcal{F}_{\ell_1}$ implies that the subsample of Theorem 4.3.7 satisfies the $(\varepsilon, 1)$-subspace embedding property for the columnspace of the weighted data matrix with respect to ℓ_1. Note that, by linearity of the ℓ_1-norm, we can fold the weights into D_wX.

Lemma 4.3.8. *Let T be a sampling and reweighting matrix according to Theorem 4.3.7. That is TD_wX is the resulting reweighted sample when Theorem 4.3.7 is applied to μ-complex input X, w. Then we have with probability* $1 - 1/n^c$

$$(1-\varepsilon)\|D_wX\beta\|_1 \leq \|TD_wX\beta\|_1 \leq (1+\varepsilon)\|D_wX\beta\|_1$$

for all $\beta \in \mathbb{R}^d$ simultaneously.

Proof. Consider any fixed $\beta \in \mathbb{R}^d$. Let $D_wX = U\Sigma V^T = UR$ be derived from the singular value decomposition where U is an orthonormal basis for the columnspace of D_wX. As in Lemma 4.3.4 we have for each index i

$$\begin{aligned} |w_ix_i\beta| &= |U_iR\beta| \leq \|U_i\|_2\|R\beta\|_2 = \|U_i\|_2\|UR\beta\|_2 \\ &= \|U_i\|_2\|D_wX\beta\|_2 \leq \|U_i\|_2\|D_wX\beta\|_1 \end{aligned} \tag{4.3}$$

The sensitivity for the ℓ_1 norm function of $x_i\beta$ is thus

$$\sup_{\beta\in\mathbb{R}^d\setminus\{0\}} \frac{w_i|x_i\beta|}{\|D_wX\beta\|_1} \leq \|U_i\|_2.$$

Note that our upper bounds on the sensitivities satisfy $s_i \geq \|U_i\|_2$. Thus also $S = \sum_{i=1}^n s_i \geq \sum_{i=1}^n \|U_i\|_2$ holds. Also, by Lemma 2.5.15, we have a bound of $O(d)$ on the VC dimension of the class of functions $\mathcal{F}_{\ell_1}$. Now, rescaling the error probability parameter η that we put into Theorem 2.5.14 by a factor of $\frac{1}{2}$, and union bound over the two sets of functions $\mathcal{F}_{log}$, and $\mathcal{F}_{\ell_1}$, the sample in Theorem 4.3.7 satisfies at the same time the claims of Theorem 4.3.7 and this lemma. □

Using this, we can show that the μ-complexity property is not violated too much after one stage of sampling.

Lemma 4.3.9. *Let T be a sampling and reweighting matrix according to Theorem 4.3.7 with parameter $\varepsilon' \leq \varepsilon/(\mu+1)$. That is TD_wX is the resulting reweighted sample when Theorem 4.3.7 succeeds on μ-complex input X, w. Suppose that simultaneously Lemma 4.3.8 holds. Let*

$$\mu' = \mu_{Tw}(X) = \sup_{\beta \in \mathbb{R}^d \setminus \{0\}} \frac{\|(TD_wX\beta)^+\|_1}{\|(TD_wX\beta)^-\|_1}.$$

Then we have $\mu' \leq (1+\varepsilon)\mu$.

Proof. For brevity of presentation let $X' = D_wX$. First note that by Lemma 4.3.8 we have for all $\beta \in \mathbb{R}^d$

$$\left(1-\varepsilon'\right)\|X'\beta\|_1 \leq \|TX'\beta\|_1 \leq \left(1+\varepsilon'\right)\|X'\beta\|_1.$$

Note that since the weights are non-negative, sampling and reweighting does not change the sign of the entries. This implies for $\vartheta^+ = |\|(TX'\beta)^+\|_1 - \|(X'\beta)^+\|_1|$ and $\vartheta^- = |\|(TX'\beta)^-\|_1 - \|(X'\beta)^-\|_1|$ that $\max\{\vartheta^+, \vartheta^-\} \leq \vartheta^+ + \vartheta^- = |\|TX'\beta\|_1 - \|X'\beta\|_1| \leq \varepsilon'\|X'\beta\|_1$.

From this and $\|X'\beta\|_1 = \|(X'\beta)^+\|_1 + \|(X'\beta)^-\|_1 \leq (\mu+1)\min\{\|(X'\beta)^+\|_1, \|(X'\beta)^-\|_1\}$ it follows for any $\beta \in \mathbb{R}^d$

$$\begin{aligned}
\frac{\|(TX'\beta)^+\|_1}{\|(TX'\beta)^-\|_1} &\leq \frac{\|(X'\beta)^+\|_1 + \varepsilon'\|X'\beta\|_1}{\|(X'\beta)^-\|_1 - \varepsilon'\|X'\beta\|_1} \\
&\leq \frac{\|(X'\beta)^+\|_1 + \varepsilon'(\mu+1)\|(X'\beta)^+\|_1}{\|(X'\beta)^-\|_1 - \varepsilon'(\mu+1)\|(X'\beta)^-\|_1} \\
&\leq \frac{\|(X'\beta)^+\|_1(1+\varepsilon)}{\|(X'\beta)^-\|_1(1-\varepsilon)} \\
&\leq \mu\,\frac{1+\varepsilon}{1-\varepsilon} \leq (1+4\varepsilon)\mu\,.
\end{aligned}$$

The claim follows by folding the constant $\frac{1}{4}$ into ε. □

Now we are ready to prove our theorem regarding the recursive subsampling algorithm.

Theorem 4.3.10. *Let $X \in \mathbb{R}^{n\times d}$ be μ-complex. Let $\varepsilon \in (0, 1/2)$, and $\varepsilon' = \varepsilon/(\mu+1)$. There exists a $(1 \pm \varepsilon')$-coreset of X for logistic regression of size*

$$k \in O\left(\frac{\mu^6}{\varepsilon^4} d^3 \log^2(\mu n d) \log^2 n (\log\log n)^8\right).$$

Such a coreset can be constructed in time $O(d^2 n \log\log n)$ with probability $1 - 1/n^c$ for any absolute constant $c > 1$.

Proof. Recall, due to Lemma 4.3.9, the μ'-complexity at the ith recursion level is upper bounded by $\mu(1+\varepsilon)^i$. We thus apply Theorem 4.3.7 recursively $l = \log\log n$ times with parameter $\varepsilon_i = \frac{\varepsilon}{2l(\mu+1)(1+\varepsilon)^i}$ for $i \in \{0 \dots l-1\}$. First we bound the approximation ratio, which is the product of the single stages. We have

$$\begin{aligned}\prod\nolimits_{i=0}^{l-1}(1+\varepsilon_i) &\leq \prod\nolimits_{i=0}^{l-1}\left(1 + \frac{\varepsilon}{(1+\varepsilon)^i 2l\mu}\right)\\ &\leq \left(1 + \frac{\varepsilon}{2l\mu}\right)^l \leq \exp\left(\frac{\varepsilon}{2\mu}\right) \leq 1 + \frac{\varepsilon}{\mu}.\end{aligned}$$

Also

$$\begin{aligned}\prod\nolimits_{i=0}^{l-1}(1-\varepsilon_i) &\geq \prod\nolimits_{i=0}^{l-1}\left(1 - \frac{\varepsilon}{(1+\varepsilon)^i 2l\mu}\right)\\ &\geq \prod\nolimits_{i=0}^{l-1}\left(1 - \frac{\varepsilon}{2l\mu}\right) \geq 1 - \sum\nolimits_{i=0}^{l-1}\frac{\varepsilon}{2l\mu} \geq 1 - \frac{\varepsilon}{2\mu}.\end{aligned}$$

Initially all weights are equal to one. So in the first application of Theorem 4.3.7 we have $\omega = 1$. This value might grow as the weights are reassigned. However, from Inequality (4.2) and the discussion below it follows, that the value of ω can grow only by a factor of $2n$ in each recursive iteration. So it remains bounded by $\omega \leq (2n)^{\log\log n}$ in all levels of our recursion. Its contribution to the lower order terms given in Theorem 4.3.7 is thus bounded by

$$O\left(\log((2n)^{1+\log\log n}) \log\log((2n)^{1+\log\log n})\right) \subseteq O\left(\log n(\log\log n)^2\right).$$

The size of the data set at recursion level $i+1$ thus satisfies

$$\begin{aligned}n_{i+1} &\leq \sqrt{n_i} \cdot \frac{Cl^2(1+\varepsilon)^{3i}\mu^3}{\varepsilon^2} d^{3/2} \log((1+\varepsilon)^i \mu n d) \log n(\log\log n)^2\\ &\leq \sqrt{n_i} \cdot \frac{Cl^2 8^i \mu^3}{\varepsilon^2} d^{3/2} \log(2^i \mu n d) \log n(\log\log n)^2\end{aligned}$$

for some constant $C > 1$. Solving the recursion until we reach $n_0 = n$ we get the following bound on n_l. We use that for our choice $l = \log\log n$ we have $2^l = \log n$ and $n^{2^{-l}} = 2^{\frac{\log n}{2^l}} = 2$.

$$\begin{aligned}
n_l &\leq n^{2^{-l}} \prod\nolimits_{i=0}^{l} \left(C \cdot \frac{l^2 8^i \mu^3}{\varepsilon^2} d^{3/2} \log(2^i \mu n d) \log n (\log\log n)^2 \right)^{2^{-i}} \\
&\leq 2 \prod\nolimits_{i=0}^{l} 8^{i2^{-i}} \prod\nolimits_{i=0}^{l} \left(C \cdot \frac{l^2 \mu^3}{\varepsilon^2} d^{3/2} \log(2^l \mu n d) \log n (\log\log n)^2 \right)^{2^{-i}} \\
&\leq 2 \prod\nolimits_{i=0}^{l} 8^{i2^{-i}} \prod\nolimits_{i=0}^{l} \left(2C \cdot \frac{l^2 \mu^3}{\varepsilon^2} d^{3/2} \log(\mu n d) \log n (\log\log n)^2 \right)^{2^{-i}} \\
&\leq 2 \cdot 8^{\sum_{i=0}^{l} i2^{-i}} \left(2C \cdot \frac{l^2 \mu^3}{\varepsilon^2} d^{3/2} \log(\mu n d) \log n (\log\log n)^2 \right)^{\sum_{i=0}^{l} 2^{-i}} \\
&\leq 2 \cdot 8^2 \left(2C \cdot \frac{l^2 \mu^3}{\varepsilon^2} d^{3/2} \log(\mu n d) \log n (\log\log n)^2 \right)^2 \\
&\leq 2 \cdot 64 \cdot 4C^2 \cdot \frac{l^4 \mu^6}{\varepsilon^4} d^3 \log^2(\mu n d) \log^2 n (\log\log n)^4
\end{aligned}$$

We conclude that for some constant $C' > C$

$$\begin{aligned}
n_l &\leq C' \cdot \frac{l^4 \mu^6}{\varepsilon^4} d^3 \log^2(\mu n d) \log^2 n (\log\log n)^4 \\
&\leq C' \cdot \frac{\mu^6}{\varepsilon^4} d^3 \log^2(\mu n d) \log^2 n (\log\log n)^8.
\end{aligned}$$

It remains to bound the failure probability. Note that we use a $\log n$ factor in the sampling sizes at all stages rather than $\log n_i$. The failure probability at each stage is thus bounded by $\frac{1}{n^{c'}}$ for $c' = c + 1 > 2$ by adjusting constants. We can thus take a union bound over the stages to get an error probability of at most

$$l \cdot \frac{1}{n^{c'}} = \frac{\log\log n}{n^{c'}} \leq \frac{1}{n^{c'-1}} \leq \frac{1}{n^c}.$$

Now recall from Theorem 4.3.7, that the running time was dominated by $O(n_i d^2)$ due to the singular value decomposition. We can thus bound the running time of the recursive algorithm for a sufficiently large absolute constant $C > 1$ by

$$Cd^2 \sum\nolimits_{i=0}^{l-1} n_i \leq Cd^2 n \log\log n \in O(d^2 n \log\log n). \qquad \square$$

5 Smallest enclosing ball for probabilistic data

We study the smallest enclosing ball problem for probabilistic data. Here, we assume that the dimension d is a constant. Our input consists of n finite discrete distributions of points in which a point can appear at z different locations in $\mathbb{R}^d \cup \{\perp\}$. We include the possibility that a point does not appear at all. Our aim is to find a center $c \in \mathbb{R}^d$ such that the *expected* maximum distance of c to points drawn from the input distributions is minimized. This is a natural extension of the smallest enclosing ball problem to the setting of probabilistic data. We develop the first $(1+\varepsilon)$-approximation algorithm for the probabilistic smallest enclosing ball problem. This is done by reduction to metric 1-median problems of exponential size that we reduce again via sampling techniques to approximate the resulting problems in polynomial time using only the small subsample, cf. Scheme (2.6). See also Section 2.6 or the exact definitions of the cost functions. We first describe our algorithm and provide the high-level ideas for its analysis. Pseudocode is given in Algorithm 1.

5.1 The algorithm

Our first step is to distinguish between the case that the total probability to realize an input point is too small and the case that with reasonable constant probability we realize at least one point when we draw from the input distributions. In both cases we can reduce the probabilistic smallest enclosing ball problem to closely related 1-median problems. Known techniques in this area enable us to take a small sample and compute an approximation of the optimal solution based only on this sample.

In the first case, we have no chance to gather sufficient information by sampling realizations of the probabilistic point set. We see only empty sets with high probability. Therefore we have to deal with this case differently. It turns out that a good approximation can be recovered if we focus on the event that a sample from the n input distributions contains exactly one point. We will see that in this case the expected cost can be approximated by a weighted 1-median problem, where each possible locations' weight is its probability, cf. [38]. More precisely (cf. Section 2.6),

$$\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,c)] \approx \mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X,c)] = \sum_{i\in[n],j\in[z]} p_{ij}\cdot \left\|q_{ij}-c\right\|_2 .$$

In the second case, the probability of sampling a non-empty realization from the input distributions is large enough. We can thus expand the expectation via its definition and thus rewrite the cost function as a weighted sum in terms of a metric distance measure on realizations drawn from the collection of n distributions. We will formally define the distance measure $m(A,B)=\max_{a\in A,b\in B}\left\|a-b\right\|_2$ for realizations of X or even more generally for the space of all finite non-empty point sets $A,B\subset\mathbb{R}^d$ and show that it is near-metric, but if one of the sets is a singleton comprising exactly one point then the distance measure m even satisfies all properties of a metric. Then we can rewrite

$$\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,c)] = \sum_P \mathbf{Pr}\left[X=P\right]\cdot \max_{p\in P}\left\|p-c\right\|_2 = \sum_P \mathbf{Pr}\left[X=P\right]\cdot m(P,c).$$

Similar to the first setting of our case distinction it remains to solve a weighted metric 1-median problem. The complication is that it is defined on the more complex space of realizations, which has exponential size in the number of input distributions.

Note that in the first case the elements are the $\Theta(nz)$ non-empty locations $q_{ij}\in\mathbb{R}^d$. In the second case, the elements are all possible realizations drawn from the input distributions, which can be as many as $\Theta(z^n)$.

In both cases, we will argue that a near-optimal solution derived from a sample of a constant number of elements will be a good approximation to the optimal solution. We remark that our algorithm cannot efficiently approximate the *cost* of the expected smallest enclosing ball, but only its approximate location.

5.2 Analysis

5.2.1 Reduction to 1-median problems

Now, we present our reductions to 1-median problems in more details. To this end let P denote a realization of $X=\{X_1,\ldots,X_n\}$ where $X\sim\mathcal{D}$. We begin our studies with the case in which the probability of drawing $P=\emptyset$ is large. More formally, we assume $\sum_{i=1}^n \mathbf{Pr}\left[X_i\neq\bot\right]\leq\varepsilon$.

```
Algorithm 1: ProbSmallestEnclosingBall($\mathcal{D}, \varepsilon, \delta$)
Input  : A set $\mathcal{D} = \{D_1, \dots, D_n\}$ of $n$ point distributions where $\forall i \in [n]$
         $D_i = \{(q_{i1}, p_{i1}), \dots, (q_{iz}, p_{iz})\} \subset (\mathbb{R}^d \cup \{\perp\}) \times [0,1]$ and $\sum_{j=1}^{z} p_{ij} = 1$,
         approximation parameter $0 < \varepsilon \leq \frac{1}{2}$, and
         a failure probability $0 < \eta \leq \frac{1}{2}$.
Output : $c \in \mathbb{R}^d$ that satisfies Theorem 5.2.8.
Set $s \in \Theta\left(\frac{1}{\varepsilon^2}\log\left(\frac{1}{\varepsilon\eta}\right)\right)$, $s_2 \in \Theta\left(\frac{s}{\varepsilon\eta}\right)$
$\varepsilon' := \varepsilon\eta/40$
Set $Q := \{q_{ij} \mid q_{ij} \neq \perp, i \in [n], j \in [z]\}$, the non-empty locations;
if $\sum_{q_{ij} \in Q} p_{ij} \leq \varepsilon$ then
    Pick a random sample $R$ of $s$ locations from $Q$, where for every $r \in R$ we have
      $r = q_{ij}$ with probability proportional to $p_{ij}$ for every $i \in [n]$ and $j \in [z]$.
    Compute a $(1+\varepsilon')$-approximation $c \in \mathbb{R}^d$ to the 1-median of $R$, i.e.,
      $\mathrm{cost}_{\mathrm{MED}}(R, c) \leq (1+\varepsilon) \min_{c' \in \mathbb{R}^d} \mathrm{cost}_{\mathrm{MED}}(R, c')$.
else
    for $t := 1$ to $s_2$ do
        Set $R_t := \emptyset$.
        for $i := 1$ to $n$ do
            Pick a random location $r_i$ such that $r_i = q_{ij}$ with probability $p_{ij}$.
            $R_t := R_t \cup \{r_i\}$.
        Compute the diameter $\Delta_t = \max_{r_1, r_2 \in R_t} \|r_1 - r_2\|_2$ of $R_t$.
        Set $G_t$ to be a $d$-dimensional grid whose side length is $\varepsilon\Delta_t/\sqrt{d}$.
        $S_t := \emptyset$.
        Add to $S_t$ the center of each cell in $G_t$ that contains at least one point of $R_t$
    Compute a center $c \in \mathbb{R}^d$ such that
      $\sum_{t=1}^{s_2} \mathrm{cost}_{\mathrm{SEB}}(S_t, c) \leq (1+\varepsilon') \min_{c' \in \mathbb{R}^d} \sum_{t=1}^{s_2} \mathrm{cost}_{\mathrm{SEB}}(S_t, c')$.
return $c$
```

In the first case our reduction connects the objective function of the probabilistic smallest enclosing ball problem to the probabilistic 1-median objective function. We establish this relationship in the following lemma inspired by Lemma 1 in [38].

Lemma 5.2.1. *If $\sum_{i=1}^{n} \mathbf{Pr}[X_i \neq \perp] \leq \varepsilon$ then for every $c \in \mathbb{R}^d$*

$$(1-\varepsilon)\,\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c)] \leq \mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, c)] \leq \mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c)]$$

where the expectation is taken over the randomness of $X \sim \mathcal{D}$.

Proof. Fix an arbitrary $c \in \mathbb{R}^d$. For every realization P of n locations in $\mathbb{R}^d \cup \{\bot\}$ we have $\mathrm{cost}_{\mathrm{SEB}}(P, c) \leq \mathrm{cost}_{\mathrm{MED}}(P, c)$. The right hand side of the proposition follows directly from this, since

$$\begin{aligned}\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, c)] &= \sum_{P} \mathbf{Pr}\left[X = P\right] \cdot \mathrm{cost}_{\mathrm{SEB}}(P, c) \\ &\leq \sum_{P} \mathbf{Pr}\left[X = P\right] \cdot \mathrm{cost}_{\mathrm{MED}}(P, c) = \mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c)]\,,\end{aligned}$$

where the sum is over every realization P of $X \sim \mathcal{D}$. For the left hand side of the proposition, we have

$$\begin{aligned}\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, c)] &= \sum_{P} \mathbf{Pr}\left[X = P\right] \cdot \mathrm{cost}_{\mathrm{SEB}}(P, c) \\ &\geq \sum_{i \in [n], j \in [z]} \mathbf{Pr}\left[X = \{q_{ij}\}\right] \cdot \|q_{ij} - c\|_2 \\ &= \sum_{i \in [n], j \in [z]} \mathbf{Pr}\left[X_i = q_{ij}\right] \cdot \prod_{k \in [n] \setminus \{i\}} \mathbf{Pr}\left[X_k = \bot\right] \cdot \|q_{ij} - c\|_2 \\ &\geq \sum_{i \in [n], j \in [z]} \mathbf{Pr}\left[X_i = q_{ij}\right] \cdot (1 - \sum_{k \in [n] \setminus \{i\}} \mathbf{Pr}\left[X_k \neq \bot\right]) \cdot \|q_{ij} - c\|_2 \\ &\geq (1 - \varepsilon) \sum_{i \in [n], j \in [z]} \mathbf{Pr}\left[X_i = q_{ij}\right] \cdot \|q_{ij} - c\|_2 \\ &= (1 - \varepsilon) \sum_{i \in [n]} \mathbb{E}[\|X_i - c\|_2] = (1 - \varepsilon)\, \mathbb{E}\left[\sum_{i \in [n]} \|X_i - c\|_2\right] \\ &= (1 - \varepsilon)\, \mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c)]\,,\end{aligned}$$

which concludes the proof. □

By Lemma 5.2.1 we can argue that any near-optimal solution to the probabilistic 1-median problem is a near-optimal solution to the probabilistic smallest enclosing ball problem.

Corollary 5.2.2. *Let $c_{\mathrm{MED}} \in \mathbb{R}^d$ be a center that minimizes $\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, \cdot)]$ and let $c_{\mathrm{SEB}} \in \mathbb{R}^d$ be a center that minimizes $\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, \cdot)]$ respectively. Let $\tilde{c} \in \mathbb{R}^d$ be a center that satisfies $\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, \tilde{c})] \leq (1 + \varepsilon)\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X, c_{\mathrm{MED}})]$. If $\sum_{i=1}^{n} \mathbf{Pr}\left[X_i \neq \bot\right] \leq \varepsilon$ then*

$$\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, \tilde{c})] \leq (1 + 4\varepsilon)\, \mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X, c_{\mathrm{SEB}})]\,.$$

Proof. By Lemma 5.2.1 we have

$$\begin{aligned}\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,\tilde{c})] &\leq \mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X,\tilde{c})]\\ &\leq (1+\varepsilon)\,\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X,c_{\mathrm{MED}})]\\ &\leq (1+\varepsilon)\,\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X,c_{\mathrm{SEB}})]\\ &\leq \frac{(1+\varepsilon)}{(1-\varepsilon)}\,\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,c_{\mathrm{SEB}})]\\ &\leq (1+4\varepsilon)\,\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,c_{\mathrm{SEB}})]\end{aligned}$$ □

By linearity of expectation it can be seen that any probabilistic 1-median instance is equivalent to a weighted instance of the non-probabilistic 1-median problem over the collection of all locations $q_{ij}, i \in [n], j \in [z]$. That is

$$\mathbb{E}[\mathrm{cost}_{\mathrm{MED}}(X,c)] = \mathbb{E}\left[\sum\nolimits_{i=1}^{n} \|X_i - c\|_2\right] = \sum\nolimits_{i=1}^{n} \mathbb{E}[\|X_i - c\|_2] = \sum_{i\in[n],j\in[z]} p_{ij} \cdot \|q_{ij} - c\|_2 \,.$$

Consequently it remains to compute a center c that will approximately minimize the 1-median instance on the right hand side. To this end and to reduce the number of points from $\Theta(nz)$ to a constant number of elements we are going to apply subsampling results from the theory of metric 1-median problems that will lead us to a good approximation.

Turning our focus to the second case, we have a reasonable probability that a realization of our probabilistic point set, i.e., a sample from the n input distributions contains at least one actual location in $\mathbb{R}^d$. More formally, in what follows, we assume that $\sum_{i=1}^{n} \mathbf{Pr}\left[X_i \neq \perp\right] > \varepsilon$ and again aim to reduce the probabilistic smallest enclosing ball problem to a 1-median problem defined on a metric space.

The idea behind our reduction becomes clear by rewriting the cost function in the following way

$$\mathbb{E}[\mathrm{cost}_{\mathrm{SEB}}(X,c)] = \sum_{P} \mathbf{Pr}\left[X = P\right] \cdot \max_{p\in P} \|p - c\|_2 \,.$$

The summation is over every possible non-empty realization P of the random variable $X \sim \mathcal{D}$ sampled from the n input distributions. $\mathbf{Pr}\left[X = P\right]$ denotes the probability of drawing the realization P. As in the first case, this is a weighted 1-median problem, where the distance function from a center c to an item P is $\mathrm{cost}_{\mathrm{SEB}}(P,c) = \max_{p\in P} \|p - c\|_2$ and this distance is weighted by $\mathbf{Pr}\left[X = P\right]$. However, it is not clear whether we still deal with a metric space as desired.

To this end, we can show that for finite non-empty sets $A, B \subset \mathbb{R}^d$ the distance measure

$$m(A, B) = \max_{a \in A, b \in B} \|a - b\|_2$$

is near-metric. Note that m cannot be a proper metric since $m(C, C) > 0$ holds for any non-singleton set $C \subset \mathbb{R}^d$. To cope with this problem, we simply define $m(A, B) = 0$ if and only if $A = B$ and note that this does not restrict our reduction since we actually need only the cases where $A = \{c\}$ comprises a single center and $B = P$ is an arbitrary non-empty set of points. This extends the above to a full metric space as we show in our next lemma.

Lemma 5.2.3. *Let $\mathcal{X}$ be the family of all finite non-empty subsets of $\mathbb{R}^d$. For arbitrary $A, B \in \mathcal{X}$ let*

$$m(A, B) = \begin{cases} \max_{a \in A, b \in B} \|a - b\|_2 & \text{if } A \neq B \\ 0 & \text{if } A = B. \end{cases}$$

Then $(\mathcal{X}, m)$ is a metric space.

Proof. The non-negativity and symmetry of m follow from the corresponding metric properties of Euclidean space $(\mathbb{R}^d, \|\cdot\|_2)$ and by definition. Further, if $A \neq B$, then there exist elements $a \in A, b \in B$, such that $a \neq b$. Thus, $m(A, B) \geq \|a - b\|_2 > 0$. But if $A = B$ then $m(A, B) = 0$ holds by definition. This proves the identity of indiscernible elements.

For proving the triangle inequality, first let $A, B, C \in \mathcal{X}$ be distinct. Let $a \in A$ and $c \in C$ such that $m(A, C) = \|a - c\|_2$. Hence for any $b \in B$,

$$m(A, C) = \|a - c\|_2 \leq \|a - b\|_2 + \|b - c\|_2 \leq m(A, B) + m(B, C),$$

where the first inequality is due to the triangle inequality in Euclidean space, and the last inequality is by the definition of m. Now, if $A = C$, then the claim follows as above by non-negativity. If $A = B$ the triangle inequality reduces to

$$m(A, C) = 0 + m(A, C) = m(A, B) + m(B, C).$$

The case $B = C$ is analogous. □

This result enables us to see our problem as a metric 1-median problem to which we can apply sampling results from the theory of 1-median problems to reduce the number of terms from $\Theta(z^n)$ to a constant number of realizations. At this point, our reduction is completed.

5.2.2 Reducing the number of points

In this section we will show that a sample of the elements of constant size is sufficient to compute a solution that is a good approximation to the optimal solution.

An additional complication lies in the second case of our reduction. The elements themselves consist of up to linearly many points drawn from the n input distributions. Therefore reducing the number of elements to a constant amount does not necessarily yield a sublinear number of points that have to be kept in memory. We will thus deal with this issue in the remainder of the section.

Subsampling the elements In the previous section we have performed a reduction of the probabilistic smallest enclosing ball problem to 1-median problems defined on two different metric spaces.

In the following we want to leverage the metric properties to derive a general subsampling result, i.e. applying to both cases. This will establish that we can compute a $(1+\varepsilon)$-approximation to an instance of the probabilistic smallest enclosing ball problem by computing an approximation to a subsampled instance of 1-median that only consists of a number of locations or realizations independent of n, the number of input distributions.

To this end, we will need a result from [2] that is based on a lemma from [86] published in the appendix of [134]. Lemma 5.2.4 states that, with high probability, any center that is far from being optimal for the original problem is also far from being optimal in a sub-sampled problem. We will leverage this result in its contrapositive form to conclude that a near optimal solution for the subsampled problem is also close to optimal for the original problem. In the paper by Ackermann et al. [2] the underlying space is any metric space with a metric distance measure.

Lemma 5.2.4 (Lemma 3.3 from [2])**.** *Let P be a finite non-empty subset of n items from $\mathcal{X}$; see Lemma 5.2.3. Let $c \in \operatorname{argmin}_{x\in\mathbb{R}^d}\mathrm{cost}_{\mathrm{MED}}(P,x)$. Let b be an arbitrary point in $\mathbb{R}^d$ where $\mathrm{cost}_{\mathrm{MED}}(P,b) > (1+\frac{4}{5}\varepsilon)\mathrm{cost}_{\mathrm{MED}}(P,c)$. Then any multiset $S \subseteq P$ of s i.i.d. uniformly sampled items from P satisfies*

$$\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{MED}}(S,b) \leq \mathrm{cost}_{\mathrm{MED}}(S,c) + \frac{\varepsilon s}{5n}\mathrm{cost}_{\mathrm{MED}}(P,c)\right] \leq \exp\left(-\frac{\varepsilon^2 s}{144}\right).$$

This result leads us to our next lemma. We show that a near-optimal solution to a sample of the elements of constant size is a $(1+\varepsilon)$-approximation for the original 1-median problem. Via the previous reductions this establishes that our algorithm returns a $(1+\varepsilon)$-approximation to the probabilistic smallest enclosing ball problem, given that it has access to an arbitrary algorithm that computes an approximation for the sample. The proof closely follows the argumentation conducted in Lemma 3.4 from [2].

Lemma 5.2.5. *Let P be a finite non-empty subset of n items from $\mathcal{X}$; see Lemma 5.2.3. Let $S \subseteq P$ be a uniform sample of s i.i.d. items from P. Let $c \in \operatorname{argmin}_{x\in\mathbb{R}^d}\mathrm{cost}_{\mathrm{MED}}(P,x)$. Let $c_S \in \operatorname{argmin}_{x\in\mathbb{R}^d}\mathrm{cost}_{\mathrm{MED}}(S,x)$ and let $\tilde{c}_S \in \mathbb{R}^d$ be a center that satisfies $\mathrm{cost}_{\mathrm{MED}}(S,\tilde{c}_S) \leq (1+\frac{\varepsilon\eta}{40})\,\mathrm{cost}_{\mathrm{MED}}(S,c_S)$. There exists a constant λ such that if $s \geq \frac{\lambda}{\varepsilon^2}\log\frac{1}{\eta\varepsilon}$ then*

$$\mathbf{Pr}\left[\mathrm{cost}_{\mathrm{MED}}(P,\tilde{c}_S) \leq (1+\varepsilon)\mathrm{cost}_{\mathrm{MED}}(P,c)\right] \geq 1-\frac{3\eta}{4}.$$

Proof. Let $\rho = \frac{4s}{\eta n}\mathrm{cost}_{\mathrm{MED}}(P,c)$. The average distance of c to the points in P equals $\frac{1}{n}\mathrm{cost}_{\mathrm{MED}}(P,c)$. By Markov's inequality we thus have $\mathbf{Pr}\left[\|x-c\|_2 > \rho\right] \leq \frac{\eta}{4s}$ for any $x \in S$. By the union bound we can infer that the sample is contained in a ball of radius ρ centered at c, i.e. $S \subseteq \mathrm{B}(c,\rho)$ holds with probability at least $1-\frac{\eta}{4}$. For every $p \in \mathrm{B}(c,\rho)$ and every $x \in \mathbb{R}^d \setminus \mathrm{B}(c,3\rho)$ we have $\|p-x\|_2 \geq 2\rho$. Thus, conditioned on the event $S \subseteq \mathrm{B}(c,\rho)$ and summing over the points in S we get $\mathrm{cost}_{\mathrm{MED}}(S,x) \geq 2\rho s$ but we also know that $\mathrm{cost}_{\mathrm{MED}}(S,c) \leq \rho s$. Therefore we have

$$\mathrm{cost}_{\mathrm{MED}}(S,x) \geq 2\mathrm{cost}_{\mathrm{MED}}(S,c) \geq 2\mathrm{cost}_{\mathrm{MED}}(S,c_S).$$

So, any solution that is a $(1+\varepsilon)$-approximation with respect to S is contained in $\mathrm{B}(c,3\rho)$. In particular this holds for $\tilde{c}_S$ since $\frac{\varepsilon\eta}{40} < \varepsilon$. We conclude that $\tilde{c}_S \in \mathrm{B}(c,3\rho)$ with probability at least $1-\frac{\eta}{4}$ since

$$\mathrm{cost}_{\mathrm{MED}}(S,\tilde{c}_S) \leq (1+\frac{\varepsilon\eta}{40})\,\mathrm{cost}_{\mathrm{MED}}(S,c_S) < 2\mathrm{cost}_{\mathrm{MED}}(S,c_S).$$

It is a well known fact (see e.g. [12, 69] or [78, Ch. 10]) that any ball $\mathrm{B}(\sigma,\tau) \subseteq \mathbb{R}^d$ can be covered by $2^{O(d)}$ balls of radius $\frac{\tau}{2}$. From this we can deduce that the ball $\mathrm{B}(c,3\rho)$ can be covered by at most $2^{jO(d)}$ balls of radius $\frac{3\rho}{2^j}$ for any $j \in \mathbb{N}$ by applying the construction recursively. We choose $j = \lceil\log\frac{120s}{\varepsilon\eta}\rceil$. This yields a set of $l \leq (\frac{240s}{\varepsilon\eta})^{O(d)}$ balls covering $\mathrm{B}(c,3\rho)$, each having radius at most

$$\frac{3\rho}{2^j} = \frac{12s}{\eta n 2^j}\mathrm{cost}_{\mathrm{MED}}(P,c) \leq \frac{\varepsilon}{10n}\mathrm{cost}_{\mathrm{MED}}(P,c).$$

Let $C = \{c_1,\ldots,c_l\}$ be the set of their centers. Furthermore define

$$C_{\mathrm{bad}} = \left\{ b \in C \;\middle|\; \mathrm{cost}_{\mathrm{MED}}(P,b) > \left(1+\frac{4}{5}\varepsilon\right)\mathrm{cost}_{\mathrm{MED}}(P,c)\right\}.$$

Now, by the union bound and using Lemma 5.2.4 it follows that there is $s \in \Theta\left(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon\eta}\right)$ that depends linearly on the constant dimension d, such that

$$\begin{aligned}\mathbf{Pr}\left[\forall b \in C_{\text{bad}} \colon \operatorname{cost}_{\text{MED}}(S, b) \le \operatorname{cost}_{\text{MED}}(S, c) + \frac{\varepsilon s}{5n}\operatorname{cost}_{\text{MED}}(P, c)\right]& \\ \le \left(\frac{240 s}{\varepsilon\eta}\right)^{O(d)} \exp\left(-\frac{\varepsilon^2 s}{144}\right) \le \frac{\eta}{4}&.\end{aligned}$$

Thus, we conclude with probability at least $1 - \frac{\eta}{4}$, that $\operatorname{cost}_{\text{MED}}(S, b) > \operatorname{cost}_{\text{MED}}(S, c) + \frac{\varepsilon s}{5n}\operatorname{cost}_{\text{MED}}(P, c)$ holds simultaneously for all $b \in C_{\text{bad}}$.

Recall that $\tilde{c}_S$ is a $(1 + \frac{\varepsilon\eta}{40})$-approximation to S by assumption. Let $q \in C$ be the closest point to $\tilde{c}_S$. Since $\tilde{c}_S \in \mathrm{B}(c, 3\rho)$ we have

$$\|q - \tilde{c}_S\|_2 \le \frac{\varepsilon}{10n}\operatorname{cost}_{\text{MED}}(P, c).$$

Moreover, by the triangle inequality and the (near)-optimal properties of $\tilde{c}_S$ and c_S respectively, it follows that

$$\begin{aligned}\operatorname{cost}_{\text{MED}}(S, q) &\le \operatorname{cost}_{\text{MED}}(S, \tilde{c}_S) + \frac{\varepsilon s}{10n}\operatorname{cost}_{\text{MED}}(P, c)\\ &\le \left(1 + \frac{\varepsilon\eta}{40}\right)\operatorname{cost}_{\text{MED}}(S, c_S) + \frac{\varepsilon s}{10n}\operatorname{cost}_{\text{MED}}(P, c)\\ &\le \left(1 + \frac{\varepsilon\eta}{40}\right)\operatorname{cost}_{\text{MED}}(S, c) + \frac{\varepsilon s}{10n}\operatorname{cost}_{\text{MED}}(P, c) \qquad (5.1)\end{aligned}$$

Note that $\mathbb{E}[\operatorname{cost}_{\text{MED}}(S, c)] = \frac{s}{n}\operatorname{cost}_{\text{MED}}(P, c)$ and therefore, again using Markov's inequality, we see that, with probability at least $1 - \frac{\eta}{4}$, we can bound $\operatorname{cost}_{\text{MED}}(S, c)$ by $\frac{4s}{\eta n}\operatorname{cost}_{\text{MED}}(P, c)$ from above. Assuming this bound to hold, we can continue our derivation

$$\begin{aligned}(5.1) &\le \operatorname{cost}_{\text{MED}}(S, c) + 2 \cdot \frac{\varepsilon s}{10n}\operatorname{cost}_{\text{MED}}(P, c)\\ &= \operatorname{cost}_{\text{MED}}(S, c) + \frac{\varepsilon s}{5n}\operatorname{cost}_{\text{MED}}(P, c)\end{aligned}$$

to deduce that, again with probability at least $1 - \frac{\eta}{4}$, we have $\operatorname{cost}_{\text{MED}}(S, q) < \operatorname{cost}_{\text{MED}}(S, b)$ for every $b \in C_{\text{bad}}$. This means that $q \notin C_{\text{bad}}$ and therefore we have $\operatorname{cost}_{\text{MED}}(P, q) \le (1 + \frac{4}{5}\varepsilon)\operatorname{cost}_{\text{MED}}(P, c)$. Finally, again leveraging the triangle inequality we can conclude

$$\begin{aligned}\operatorname{cost}_{\text{MED}}(P, \tilde{c}_S) &\le \operatorname{cost}_{\text{MED}}(P, q) + n \cdot \|q - \tilde{c}_S\|_2\\ &\le \left(1 + \frac{4}{5}\varepsilon\right)\operatorname{cost}_{\text{MED}}(P, c) + \frac{\varepsilon}{10}\operatorname{cost}_{\text{MED}}(P, c)\\ &\le (1 + \varepsilon)\operatorname{cost}_{\text{MED}}(P, c).\end{aligned}$$

Note, that all events that we assumed occur simultaneously with probability $1 - \frac{3\eta}{4}$ by the union bound. This yields the proposition. □

Discretizing the realizations We want to solve the optimization problem in Line 17 of the algorithm efficiently. To this end, note that a realization denoted by R_t may have linear size in n. In order to get rid of the dependency on n, we discretize R_t by covering the points from R_t using a grid to size $O(1/\varepsilon^d)$. In the following we show that for arbitrary center, evaluating the distance measure m on R_t and on its discretization S_t, as computed in line 16 of the algorithm, differs only by a $(1 \pm \varepsilon)$-factor. This, finally implies that an optimal solution regarding the discretized realizations S_t is a near-optimal solution to the same objective when evaluating with respect to R_t. Hence, a subset S_t that preserves the maximum distance for any $c \in \mathbb{R}^d$, up to a factor $(1 \pm \varepsilon)$ will suffice. Such a subset is known as strong coreset for 1-center. See e.g. [5] for examples of alternative constructions.

Lemma 5.2.6. *For every integer t, $1 \leq t \leq s_2$ and center $c \in \mathbb{R}^d$, we have*

$$(1-\varepsilon)\operatorname{cost}_{\mathrm{SEB}}(R_t, c) \leq \operatorname{cost}_{\mathrm{SEB}}(S_t, c) \leq (1+\varepsilon)\operatorname{cost}_{\mathrm{SEB}}(R_t, c).$$

Proof. Let $\Delta_t = \max_{r_1, r_2 \in R_t} \|r_1 - r_2\|_2$ denote the diameter of R_t. Note that by construction, for every point $x \in R_t$ there exists a point $y \in S_t$ at distance at most $\|x - y\|_2 \leq \frac{\varepsilon \Delta_t}{2}$ located in the same cell and vice versa. Fix an arbitrary $c \in \mathbb{R}^d$. By triangle inequality we have for some $y \in S_t$ and its associated point $x \in R_t$ that

$$\begin{aligned}
\operatorname{cost}_{\mathrm{SEB}}(S_t, c) &= \|c - y\|_2 \\
&\leq \|c - x\|_2 + \|x - y\|_2 \\
&\leq \operatorname{cost}_{\mathrm{SEB}}(R_t, c) + \frac{\varepsilon \Delta_t}{2} \\
&\leq \operatorname{cost}_{\mathrm{SEB}}(R_t, c) + \varepsilon \operatorname{cost}_{\mathrm{SEB}}(R_t, c).
\end{aligned}$$

To verify the last inequality, consider $a, b \in R_t$ attaining $\|a - b\|_2 = \Delta_t$. By the pigeonhole principle and triangle inequality we have

$$\Delta_t = \|a - b\|_2 \leq 2 \max\{\|a - c\|_2, \|b - c\|_2\} \leq 2 \operatorname{cost}_{\mathrm{SEB}}(R_t, c).$$

The lower bound follows from $\operatorname{cost}_{\mathrm{SEB}}(R_t, c) \leq \operatorname{cost}_{\mathrm{SEB}}(S_t, c) + \varepsilon \operatorname{cost}_{\mathrm{SEB}}(R_t, c)$ and can be derived similarly. □

Since every term $\text{cost}_{\text{SEB}}(R_t, c)$ in the sum over t is approximated up to a factor of $1 \pm \varepsilon$, the sum is also approximated as stated in the following corollary.

Corollary 5.2.7. *Let $R_1, \ldots, R_{s_2} \subset \mathbb{R}^d$ and let $S_1, \ldots, S_{s_2}$ be their corresponding discretizations as computed by Algorithm 1. Let c_{opt} minimize $R(c) = \sum_{t=1}^{s_2} \text{cost}_{\text{SEB}}(R_t, c)$ over $c \in \mathbb{R}^d$, and let c^* minimize $S(c) = \sum_{t=1}^{s_2} \text{cost}_{\text{SEB}}(S_t, c)$ over $c \in \mathbb{R}^d$. Then*

$$R(c^*) \leq (1 + 4\varepsilon)\, R(c_{opt}).$$

Proof. By applying Lemma 5.2.6 term-wise it holds that $R(c^*) \leq S(c^*)/(1 - \varepsilon)$. Furthermore $S(c^*) \leq S(c_{opt})$ since c^* minimizes S by definition. Another application of Lemma 5.2.6 yields $S(c_{opt}) \leq (1 + \varepsilon)\, R(c_{opt})$. Now putting all together results in

$$R(c^*) \leq \frac{(1+\varepsilon)}{(1-\varepsilon)}\, R(c_{opt}) \leq (1 + 4\varepsilon)\, R(c_{opt})$$

which concludes the proof. □

5.2.3 Main result

In the previous sections we have gathered all the results that we need to combine in order to prove our main result concerning Algorithm 1.

Theorem 5.2.8. *Let $0 < \varepsilon, \eta \leq \frac{1}{2}$. Let $\mathcal{D}$ be a set of n discrete distributions, where each distribution is over z locations in $\mathbb{R}^d \cup \{\perp\}$. Let $c \in \mathbb{R}^d$ denote the output of a call to* ProbSmallestEnclosingBall$(\mathcal{D}, \varepsilon, \delta)$*, see Algorithm 1. Then, with probability at least $1 - \eta$ over the randomness of the algorithm, c satisfies*

$$\mathbb{E}[\text{cost}_{\text{SEB}}(X, c)] \leq (1 + \varepsilon) \min_{c' \in \mathbb{R}^d} \mathbb{E}\left[\text{cost}_{\text{SEB}}(X, c')\right],$$

where the expectations are taken over the randomness of $X \sim \mathcal{D}$. Moreover, the center c can be computed in time

$$O\left(\left(\frac{nz}{\varepsilon^3 \eta} + \frac{1}{\varepsilon^{4d+2}\eta^d}\right) \log^{d+1}\left(\frac{1}{\varepsilon\eta}\right)\right).$$

Proof. We begin with the correctness. We have $\sum_{i=1}^{n} \mathbf{Pr}\left[X_i \neq \perp\right] \leq \varepsilon$ in the low-probability case. By Corollary 5.2.2 we can focus on computing a $(1 + \varepsilon)$-approximation to the probabilistic 1-median problem while loosing only a factor of $1 + 4\varepsilon$. Our further reduction to a deterministic version of the 1-median problem brings no additional loss and enables us to apply the sampling result formalized in Lemma 5.2.5 directly. Sampling elements proportional to their weights corresponds to uniform sampling from a multiset where each element occurs a number of times that has the same proportion to the size of the multiset as its weight to the total weight added over all elements. The assumptions of Lemma 5.2.5 are thus satisfied. Now a $(1+\varepsilon')$-approximation for $\varepsilon' = \frac{\varepsilon\eta}{40}$ to the subsampled instance as computed by Algorithm 1, is with probability at least $(1-\frac{3\eta}{4}) \geq 1-\eta$ a $(1+\varepsilon)$-approximation to the 1-median problem. Thus, by Corollary 5.2.2, we can conclude that we have a $(1+4\varepsilon)$-approximation to the original probabilistic smallest enclosing ball instance.

In the case $\sum_i \mathbf{Pr}\left[X_i \neq \perp\right] > \varepsilon$ the reduction involves a metric 1-median problem on the set of possible realizations with the distance function m. We can thus apply again Lemma 5.2.5 where the sets R_t are samples from the family of all possible realizations of the probabilistic point set. Also, we have at least s non-empty realizations with probability at least $1-\frac{\eta}{4}$ by choosing $s_2 = \frac{4s}{\varepsilon\eta}$. This holds since the probability of sampling a non-empty set is bounded below by ε. Consequently the expected number of samples L that we need to collect s non-empty sets is bounded by $\frac{s}{\varepsilon}$. Now by an application of Markov's inequality we have that $\mathbf{Pr}\left[L \geq \frac{4s}{\varepsilon\eta}\right] \leq \frac{\eta}{4}$. Corollary 5.2.7 states that by discretizing the sets to keep them small we loose only a factor of $(1+4\varepsilon)$. Therefore we can conclude that by our assumptions we end up with an approximation factor of at most $(1+\varepsilon)(1+4\varepsilon) \leq 1+7\varepsilon$ with probability at least $(1-\frac{3\eta}{4}-\frac{\eta}{4}) = 1-\eta$. Rescaling ε concludes the correctness of our proposition.

We proceed with the running time: Line 5 takes $O(|Q|\cdot s) \subseteq O(nzs)$ time by partitioning the interval $[0,1]$ according to the probabilities p_{ij} and then sample numbers in this interval. The sets R_t in Line 11 can be sampled similarly to Line 5 in $O(zs)$ time for each of the n input distributions $D_i \in \mathcal{D}$. This sums up again to $O(nzs)$ time. The grid G_t can be computed in $O(n+\frac{1}{\varepsilon^d})$ time using hashing where collisions are simply ignored. The set S_t has $O(\frac{1}{\varepsilon^d})$ points. In order to sample at least s non-empty realizations with probability at least $1-\frac{\eta}{4}$ we chose to make $s_2 \in \Theta(\frac{s}{\varepsilon\eta})$ trials in the loop of line 8. We can proceed with only s of them and abort the loop when we have collected sufficiently many. Excluding Lines 6 and 17 the running time thus sums up to $O(\frac{nzs}{\varepsilon\eta}+\frac{s}{\varepsilon^d})$.

Note that we can solve the metric 1-median optimization problems in Lines 6 and 17 in a similar way. We stress that the running time is dominated by Line 17 due to the more complex type of elements. We thus focus on that case. Let $S = \bigcup_t S_t$ be the input set. We first compute a 2-approximation $\tilde{\Delta}$ to the diameter Δ of S by fixing an arbitrary point $p \in S$ and taking $\tilde{\Delta} = \max_{q\in S} \|p-q\|_2$. This is a 2-approximation because of the triangle inequality. Now put a grid that covers all points with cells of side length $\varepsilon'\tilde{\Delta}/(4s\sqrt{d})$. Let $c \in \mathbb{R}^d$ denote the optimal center. Let x be the closest grid point to c. By construction it holds that $\|x-c\|_2 \le \varepsilon'\tilde{\Delta}/(4s) \le \varepsilon'\Delta/(2s)$. Now, since the sum comprises the maximum element and by the triangle inequality it follows that

$$\sum\nolimits_{t=1}^{s} \mathrm{cost}_{\mathrm{SEB}}(S_t, c) \ge \max_{t\in[s]}\{\mathrm{cost}_{\mathrm{SEB}}(S_t, c)\} = \mathrm{cost}_{\mathrm{SEB}}\,(S, c) \ge \Delta/2.$$

Therefore we can conclude that

$$\begin{aligned}\sum\nolimits_{t=1}^{s} \mathrm{cost}_{\mathrm{SEB}}(S_t, x) &\le \sum\nolimits_{t=1}^{s} \mathrm{cost}_{\mathrm{SEB}}(S_t, c) + s\cdot\|x-c\|_2 \\ &\le \sum\nolimits_{t=1}^{s} \mathrm{cost}_{\mathrm{SEB}}(S_t, c) + s\cdot\varepsilon'\Delta/(2s) \\ &\le (1+\varepsilon')\sum\nolimits_{t=1}^{s} \mathrm{cost}_{\mathrm{SEB}}(S_t, c).\end{aligned}$$

This means that by trying all grid points we will end up with a $(1+\varepsilon')$-approximation as desired. The grid has $O(\frac{s^d}{\varepsilon'^d})$ points and for each of these, we have to sum up distances to s sets. The distances have to be determined by maximization over $O(\frac{1}{\varepsilon^d})$ points each.

We can conclude that the total running time is thus

$$\begin{aligned}O\left(\frac{nzs}{\varepsilon\eta} + \frac{s^d}{\varepsilon'^d}\cdot s\cdot\frac{1}{\varepsilon^d}\right) &\subseteq O\left(\frac{nzs}{\varepsilon\eta} + \frac{s^{d+1}}{\varepsilon^{2d}\eta^d}\right) \\ &\subseteq O\left(\left(\frac{nz}{\varepsilon^3\eta} + \frac{1}{\varepsilon^{4d+2}\eta^d}\right)\log^{d+1}\left(\frac{1}{\varepsilon\eta}\right)\right).\end{aligned} \qquad \square$$

5.3 Extensions to the streaming setting

In the streaming setting, our space requirements are limited to be of order at most $(\log n)^{O(1)}$ and we are allowed only a single pass over the input data. We assume that the input distributions $D_i = \{(q_{i1}, p_{i1}), \ldots, (q_{iz}, p_{iz})\} \subset (\mathbb{R}^d \cup \{\perp\}) \times [0,1]$, for $i \in [n]$ arrive one by one and are only inserted but never deleted. In order to translate our algorithm into a single-pass algorithm with a memory bound independent of n (though exponential in d), note that by Lemma 5.2.5 in both of the cases in which our algorithm operates, we only need a constant size sample of the elements in order to get a good approximation. In the first case we need to sample $s \in \Theta(\frac{1}{\varepsilon^2}\log\frac{1}{\varepsilon\eta})$ of the locations $q_{ij} \neq \perp$ proportional to their probabilities p_{ij} with repetition which can be done by running s independent copies of the weighted sampling algorithm from [28] which is a straightforward generalization of the well-known reservoir sampling approach [140] to the weighted case; see also [50]. At the same time we also sample everything we need for the second case. That is, we sample

$\Theta(\frac{s}{\varepsilon\eta})$ times independently from the n distributions one by one as they arrive in the data stream. By our reasoning from Section 5.2.2, for every independent non-empty copy we can store a coreset of size $O(\frac{1}{\varepsilon^d})$ that allows us to efficiently approximate the distance measure defined on the original realizations but its construction assumes that the diameter Δ of a realization is known in advance. While streaming the input in one pass, we could re-insert all the points whenever the diameter of a realization grows significantly. However, this can happen quite often and therefore may require many re-computations and re-insertions of the discretized point set.

We can get around this problem by replacing the grid construction by a concentric exponential grid that we can maintain efficiently and dynamically in the streaming setting. We note that our construction is inspired by the grids used in [74] to obtain coresets for the k-median problem, though extended here to handle insertions. The first point x_0 will be the center of the grid. When the second point arrives, we put a d-dimensional axis-parallel cube centered at the first point and large enough to cover both points. Let the side length of the cube be l. Then we subdivide it into equal cells of side length $\frac{\varepsilon l}{4\sqrt{d}}$. Whenever a point is inserted outside the current range of the grid, we double its side length until the new point is covered. This results in concentric levels L_i of side length $l2^i$ where we subdivide the space $L_i \setminus L_{i-1}$ into cells of side length $\frac{\varepsilon l 2^i}{4\sqrt{d}}$. Thus, the cells also become coarser and coarser with increasing distance from the center of the grid.

We need to ensure two properties. The approximation guarantee of Lemma 5.2.6 needs to be satisfied and the number of cells should not grow too large.

For the first issue, recall Lemma 5.2.6 and note that for any point x and its closest point in the grid y we need to bound $\|x-y\|_2$ by $\frac{\varepsilon\Delta}{2}$, such that it continues to hold. To see that this is indeed true, let $\lambda = l2^{i^*}$ be the side length of the final cube covering all input points. The coarsest cells thus have a side length of $\frac{\varepsilon l 2^{i^*}}{4\sqrt{d}}$. Also there must exist a point at distance more than $\frac{\lambda}{4}$ from the center of our grid. Therefore we can deduce $\Delta > \frac{\lambda}{4}$ and consequently $\|x-y\|_2 \leq \sqrt{d}\,\|x-y\|_\infty \leq \frac{\varepsilon\lambda}{8} < \frac{\varepsilon\Delta}{2}$ which means that our reasoning from Section 5.2.2 still applies.

Now we address the second complication. Our grid construction may have $\Omega(\log\frac{\Delta}{l})$ levels since, at the end of the stream, our grid construction is supposed to cover all the sampled input points. To reduce this factor, note that since our cells become coarser and coarser, after at most $j = i + \lceil \log \frac{4\sqrt{d}}{\varepsilon} \rceil$ levels it holds that

$$\frac{\varepsilon l 2^{i+\left\lceil \log\frac{4\sqrt{d}}{\varepsilon}\right\rceil}}{4\sqrt{d}} \geq l2^i,$$

i.e., the coarsest cells are at least as large as the entire grid at the ith level. Thus, we can collapse everything up to the ith level and store only the center point as a representative and, more importantly, we only need to store $O(\log \frac{1}{\varepsilon})$ levels instead of $\Omega(\log \frac{\Delta}{l})$.

As a consequence of the alternative construction, our space requirements are thus a little larger in the streaming setting but the insertion of a point x can be done in time $O(d + \log \frac{1}{\varepsilon}) \subseteq O(\log \frac{1}{\varepsilon})$. Deciding at which level we need to insert can be achieved in $O(d)$ by computing $||x - x_0||_\infty$ but we possibly need to inspect all $O(\log \frac{1}{\varepsilon})$ grids to find the one corresponding to the level. Deciding into which actual cell the point is inserted can be done by inspecting every coordinate one by one. The actual insertion can be achieved in time $O(1)$ using arrays or hashing for each grid. All in all we can maintain all information that we need in one single pass over the probabilistic data by storing a summary of $O(s + \frac{s}{\varepsilon^d} \log \frac{1}{\varepsilon}) \subseteq O(\frac{1}{\varepsilon^{d+2}} \log^2 \frac{1}{\varepsilon\eta})$ points from $\mathbb{R}^d$. We summarize this discussion in the following corollary.

Corollary 5.3.1. *For any $0 < \varepsilon, \eta \leq \frac{1}{2}$ there exists a single-pass insertion-only streaming algorithm that returns a $(1 + \varepsilon)$-approximation to the probabilistic smallest enclosing ball problem with probability at least $1 - \eta$. The algorithm stores*

$$O\left(\frac{1}{\varepsilon^{d+2}} \log^2 \frac{1}{\varepsilon\eta}\right)$$

points from $\mathbb{R}^d$ and has an update time of

$$O\left(\frac{z}{\varepsilon^3\eta} \log^2 \frac{1}{\varepsilon\eta}\right)$$

as the input distributions $D_i = \{(q_{i1}, p_{i1}), \dots, (q_{iz}, p_{iz})\} \subset (\mathbb{R}^d \cup \{\perp\}) \times [0, 1]$, for $i \in [n]$ arrive one by one. The post-processing after reading the input stream can be done in time

$$O\left(\frac{1}{\varepsilon^{4d+2}\eta^d} \log^{d+2} \frac{1}{\varepsilon\eta}\right).$$

Proof. Using the modifications to Algorithm 1 discussed above, the proposition follows by reasoning similarly as in the proof of Theorem 5.2.8 in the off-line setting. □

6 Conclusions and open problems

We summarize the main results of the present manuscript and pose related open questions.

Bayesian regression via subspace embeddings We have introduced ε-subspace embeddings as a data reduction technique for Bayesian regression. Furthermore, we have surveyed their useful properties when the computations are performed in sequential streaming as well as in distributed environments. These scenarios are highly desirable when dealing with massive data sets, [142]. The size of the reduced data set is $k \in O((d/\varepsilon)^{O(1)})$. which is in particular independent of the number n of observations in the original data set. Therefore, subsequent computations can operate within time and space bounds that are also independent of n, regardless of which algorithm is actually used.

Under mild assumptions, we prove that evaluating a Gaussian likelihood function based on the embedded data instead of the original data yields a good approximation in terms of the ℓ_2 Wasserstein distance. Our main result shows that the posterior distribution of Bayesian linear regression is approximated up to little error depending on only an ε-fraction of its defining parameters. This holds when using arbitrary normal priors or the degenerate case of a uniform distribution over $\mathbb{R}^d$. We also showed how to extend these results to general ℓ_p spaces, leading to similar but weaker results for p-generalized normal distributions. These are interesting for modeling different sensitivities to outlying observations via differently shaped tails, depending on the value of $p \in [1, \infty)$

Open problems lie in generalizing our results to other classes of distributions for the likelihood and to more general priors, like hierarchical priors. It seems likely to generalize to arbitrary priors via simple convexity arguments given in [84]. However, these depend on the normalization constants of the involved distributions, which can be large. But as we have seen in our results, these constants do not play a crucial role at least for bounding the Wasserstein distance.

Another interesting direction towards generalizing the likelihood component would be to consider generalized linear models in the Bayesian setting or normal mixtures, since the latter allow to approximate any continuous distribution in the limit of mixture components to infinity.

Core dependency networks Inspired by the question of how we can train probabilistic graph models on a large dataset, we have studied coresets for learning the structure of dependency networks. We established the first rigorous guarantees for obtaining compressed $(1+\varepsilon)$-approximations of Gaussian dependency networks on massive data sets. Core dependency networks provide several interesting directions for future research. The conditional local structure opens the door to explore hybrid multivariate models on massive data sets, where each variable can potentially come from a different generalized linear model. This can further be used to hint at independencies among variables in the multivariate setting, making them useful in many other large data applications. Our results may pave the way to establish coresets for deep models using the close connection between dependency networks and deep generative stochastic networks [18], sum-product networks [112, 123], as well as other statistical models that build multivariate distributions from univariate ones [147].

Generalized linear models We also studied coresets for generalized linear models such as Poisson or logistic regression. This was motivated not only by their natural application in dependency networks, but also since they build one of the most interesting family of regression models in statistics, machine learning and computer science.

Unfortunately, we proved worst-case impossibility results on coresets of sublinear size for Poisson regression. A review of the Poisson log-normal model for count data provided insights into why sampling based coreset constructions for ℓ_2 are able to recover close approximations of the maximum likelihood estimator, although they are not able to approximate the cost for all solutions up to small factors.

Similarly, we showed that sublinear coresets for logistic regression do not exist in general. To overcome this situation we introduced a new complexity measure μ that quantifies the amount of overlap of positive and negative observations in the data and the balance in their statistical model. We developed the first rigorously sublinear $(1 \pm \varepsilon)$-coresets for logistic regression, given that the original data has small μ-complexity. Future research may focus on translating these techniques to the streaming setting. It is non-trivial to maintain a sample of our distribution in one pass over the data. An interesting direction is to focus on sketching algorithms [144], rather than sampling, to achieve one-pass. A promising technique is given in [33], but it seems that in its original form it is limited to a $\Theta(1)$-approximation; $\Theta(\log n)$ if the sketch needs to form a coreset, even for mild μ-complex data. Another interesting question is whether we can give lower bounds in terms of μ and study the trade-off between the dependency on μ and n.

Smallest enclosing ball for probabilistic data We studied the smallest enclosing ball problem for probabilistic data, where every input point follows a discrete probability distribution on several locations including the case that a point is not realized at all. Our result is the first randomized $(1 + \varepsilon)$-approximation algorithm for this problem. Our main technical contribution is a reduction to different 1-median problems, one of which is defined on a more complex metric space. Our result improves upon the polynomial time $O(1)$-approximation algorithms for metric k-center problems from [38, 68] for the special case of the Euclidean space and $k = 1$.

We hope that our techniques will help to extend machine learning problems like support vector data description and support vector machines to their probabilistic versions. This seems promising due to their connection to the smallest enclosing ball problem [133, 137]. However, finding a reduction to a small finite dimension is crucial, since these methods usually lift the points into very large dimensions via kernel functions.

An important open problem is thus to reduce the exponential dependency on the dimension d. This could be tackled for example, by computing the d-dimensional grids after projecting the data on low-dimensional space using random projections [1, 26], PCA [55], or considering the low dimensional subspace spanned by a small random sample [100]. Another interesting direction is to combine the latter approaches within the framework of stochastic gradient descent, cf. [114].

Finally, our result assumes distributions over finite possible locations. A generalization to families of continuous distributions seems promising, given an efficient algorithm to sample from these distributions.

Bibliography

[1] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *Journal of Computer and System Sciences*, **66**(4):671–687, 2003.

[2] M. R. Ackermann, J. Blömer, and C. Sohler. Clustering for metric and nonmetric distance measures. *ACM Transactions on Algorithms*, **6**(4):59:1–59:26, 2010.

[3] P. K. Agarwal and R. Sharathkumar. Streaming algorithms for extent problems in high dimensions. *Algorithmica*, **72**(1):83–98, 2015.

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

[5] P. K. Agarwal, S. Har-Peled, and K. R. Varadarajan. Geometric approximations via coresets. *Combinatorial and Computational Geometry - MSRI Publications*, **52**:1–30, 2005.

[6] N. Ailon and E. Liberty. Fast dimension reduction using Rademacher series on dual BCH codes. *Discrete & Computational Geometry*, **42**(4):615–630, 2009.

[7] G. I. Allen and Z. Liu. A local poisson graphical model for inferring networks from sequencing data. *IEEE Transactions on Nanobioscience*, **12**(3):189–198, 2013.

[8] N. Alon, L. Babai, and A. Itai. A fast and simple randomized parallel algorithm for the maximal independent set problem. *Journal of Algorithms*, **7**(4):567–583, 1986.

[9] N. Alon, Y. Matias, and M. Szegedy. The space complexity of approximating the frequency moments. *Journal of Computer and System Sciences*, **58**(1):137–147, 1999.

[10] A. Andoni. High frequency moments via max-stability. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6364–6368, 2017.

[11] A. Andoni, H. L. Nguyên, Y. Polyanskiy, and Y. Wu. Tight lower bound for linear sketches of moments. In *Proceedings of the 40th International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 25–32, 2013.

[12] P. Assouad. Plongements Lipschitziens dans $\mathbb{R}^n$. *Bulletin de la Société Mathématique de France*, **111**(4):429–448, 1983.

[13] M. Badoiu and K. L. Clarkson. Smaller core-sets for balls. In *Proceedings of the 14th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 801–802, 2003.

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

[15] M. Badoiu, S. Har-Peled, and P. Indyk. Approximate clustering via core-sets. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 250–257, 2002.

[16] M.-F. Balcan, B. Manthey, H. Röglin, and T. Roughgarden. Analysis of algorithms beyond the worst case (Dagstuhl seminar 14372). *Dagstuhl Reports*, **4**(9):30–49, 2015.

[17] J. D. Batson, D. A. Spielman, and N. Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, **41**(6):1704–1721, 2012.

[18] Y. Bengio, E. Laufer, G. Alain, and J. Yosinski. Deep generative stochastic networks trainable by backprop. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, pages 226–234, 2014.

[19] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **36**(2):192–236, 1974.

[20] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society, Series D (The Statistician)*, **24**(3):179–195, 1975.

[21] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

[22] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**(4):929–965, 1989.

[23] C. Boutsidis and A. Gittens. Improved matrix algorithms via the Subsampled Randomized Hadamard Transform. *SIAM Journal on Matrix Analysis and Applications*, **34**(3):1301–1340, 2013.

[24] C. Boutsidis, P. Drineas, and M. Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Transactions on Information Theory*, **59**(10):6880–6892, 2013.

[25] V. Braverman, D. Feldman, and H. Lang. New frameworks for offline and streaming coreset constructions. *CoRR*, abs/1612.00889, 2016.

[26] E. J. Candes and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, **52**(12): 5406–5425, 2006.

[27] J. M. Carlson, Z. L. Brumme, C. M. Rousseau, C. J. Brumme, P. Matthews, C. M. Kadie, J. I. Mullins, B. D. Walker, P. R. Harrigan, P. J. R. Goulder, and D. Heckerman. Phylogenetic dependency networks: Inferring patterns of CTL escape and codon covariation in HIV-1 gag. *PLoS Computational Biology*, **4**(11):1–23, 2008.

[28] M. T. Chao. A general purpose unequal probability sampling plan. *Biometrika*, **69** (3):653–656, 1982.

[29] M. Charikar, K. C. Chen, and M. Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, **312**(1):3–15, 2004.

[30] K. L. Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM symposium on Discrete algorithms (SODA)*, pages 257–266, 2005.

[31] K. L. Clarkson and D. P. Woodruff. Numerical linear algebra in the streaming model. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 205–214, 2009.

[32] K. L. Clarkson and D. P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2013.

[33] K. L. Clarkson and D. P. Woodruff. Sketching for M-estimators: A unified approach to robust regression. In *Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 921–939, 2015.

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

[35] K. L. Clarkson, P. Drineas, M. Magdon-Ismail, M. W. Mahoney, X. Meng, and D. P. Woodruff. The fast Cauchy transform and faster robust linear regression. *SIAM Journal on Computing*, **45**(3):763–810, 2016.

[36] M. B. Cohen, J. Nelson, and D. P. Woodruff. Optimal approximate matrix product in terms of stable rank. In *Proceedings of the 43rd International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 11:1–11:14, 2016.

[37] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, **19**(1):15–18, 1977.

[38] G. Cormode and A. McGregor. Approximation algorithms for clustering uncertain data. In *Proceedings of the 27th Symposium on Principles of Database Systems (PODS)*, pages 191–200, 2008.

[39] K. Csillery, M. Blum, O. Gaggiotti, and O. Francois. Approximate Bayesian Computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**(7):410–418, 2010.

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

[41] J. Dean and S. Ghemawat. MapReduce: a flexible data processing tool. *Communications of the ACM*, **53**(1):72–77, 2010.

[42] M. Dietzfelbinger. Universal hashing and k-wise independent random variables via integer arithmetic without primes. In *Proceedings of the 13th Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 569–580, 1996.

[43] P. A. M. Dirac. *The Principles of Quantum Mechanics.* International Series of Monographs on Physics. Oxford University Press, 1981.

[44] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1127–1136, 2006.

[45] P. Drineas, M. W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, **30**(2):844–881, 2008.

[46] P. Drineas, M. W. Mahoney, S. Muthukrishnan, and T. Sarlós. Faster least squares approximation. *Numerische Mathematik*, **117**(2):219–249, 2011.

[47] P. Drineas, M. Magdon-Ismail, M. W. Mahoney, and D. P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *Journal of Machine Learning Research*, **13**:3475–3506, 2012.

[48] W. DuMouchel and D. K. Agarwal. Applications of sampling and fractional factorial designs to model-free data squashing. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 511–516, 2003.

[49] W. DuMouchel, C. Volinsky, T. Johnson, C. Cortes, and D. Pregibon. Squashing flat files flatter. In *Proceedings of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 6–15, 1999.

[50] P. S. Efraimidis. Weighted random sampling over data streams. In *Algorithms, Probability, Networks, and Games*, pages 183–195. Springer International, 2015.

[51] H. Fanaee-T and J. Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, **2**(2-3):113–127, 2014.

[52] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC)*, pages 434–444, 1988.

[53] D. Feldman and M. Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd ACM Symposium on Theory of Computing (STOC)*, pages 569–578, 2011.

[54] D. Feldman, M. Faulkner, and A. Krause. Scalable training of mixture models via coresets. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2142–2150, 2011.

[55] D. Feldman, M. Schmidt, and C. Sohler. Turning Big Data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. In *Proceedings of the 24th*

Annual ACM-SIAM Symposium on Discrete Algorithms, (SODA), pages 1434–1453, 2013.

[56] D. Feldman, A. Munteanu, and C. Sohler. Smallest enclosing ball for probabilistic data. In *Proceedings of the 30th Annual Symposium on Computational Geometry (SoCG)*, pages 214–223, 2014.

[57] D. Feldman, M. Schmidt, and C. Sohler. Turning Big Data into tiny data: Constant-size coresets for k-means, PCA and projective clustering. *Unpublished journal version*, 2018. Via personal communication.

[58] A. M. Frieze, R. Kannan, and S. Vempala. Fast Monte-Carlo algorithms for finding low-rank approximations. *Journal of the ACM*, **51**(6):1025–1041, 2004.

[59] A. Gelman and M. D. Hoffman. The No-U-turn sampler: adaptively setting path lengths in Hamiltonian monte carlo. *Journal of Machine Learning Research*, **15**: 1593–1623, 2014.

[60] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, **2**(4):1360–1383, 2008.

[61] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, 3rd edition, 2014.

[62] L. N. Geppert, K. Ickstadt, A. Munteanu, J. Quedenfeld, and C. Sohler. Random projections for Bayesian regression. *Statistics and Computing*, **27**(1):79–101, 2017.

[63] F. Gessert, W. Wingerath, S. Friedrich, and N. Ritter. NoSQL database systems: a survey and decision guidance. *Computer Science - Research and Development*, 32 (3-4):353–365, 2017.

[64] C. R. Givens and R. M. Shortt. A class of Wasserstein metrics for probability distributions. *The Michigan Mathematical Journal*, **31**(2):231–240, 1984.

[65] G. H. Golub and C. F. van Loan. *Matrix computations*. Johns Hopkins University Press, Baltimore, 4th edition, 2013.

[66] I. R. Goodman and S. Kotz. Multivariate θ-generalized normal distributions. *Journal of Multivariate Analysis*, **3**(2):204–219, 1973.

[67] J. Groß. *Linear Regression.* Springer, Berlin, Heidelberg, 2003.

[68] S. Guha and K. Munagala. Exceeding expectations and clustering uncertain data. In *Proceedings of the 28th Symposium on Principles of Database Systems (PODS)*, pages 269–278, 2009.

[69] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. In *Proceedings of the 44th Symposium on Foundations of Computer Science (FOCS)*, pages 534–543, 2003.

[70] L. Habel, A. Molina, T. Zaksek, K. Kersting, and M. Schreckenberg. Traffic simulations with empirical data: How to replace missing traffic flows? In *Proceedings of Traffic and Granular Flow '15*, pages 491–498, 2016.

[71] F. Hadiji, A. Molina, S. Natarajan, and K. Kersting. Poisson dependency networks: Gradient boosted models for multivariate count data. *Machine Learning*, **100**(2-3): 477–507, 2015.

[72] N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**(2):217–288, 2011.

[73] S. Har-Peled. A simple algorithm for maximum margin classification, revisited. *CoRR*, abs/1507.01563, 2015.

[74] S. Har-Peled and S. Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 291–300, 2004.

[75] S. Har-Peled, D. Roth, and D. Zimak. Maximum margin coresets for active and noise tolerant learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 836–841, 2007.

[76] H. He and E. A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, **21**(9):1263–1284, 2009.

[77] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. M. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, **1**:49–75, 2000.

[78] J. Heinonen. *Lectures on analysis on metric spaces.* Universitext. Springer, New York, 2001.

[79] G. Heinze and M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in Medicine*, **21**(16):2409–2419, 2002.

[80] J. Hilbe. *Logistic Regression Models.* Texts in Statistical Science. Chapman & Hall/CRC, Boca Raton, 2009.

[81] R. Horn and C. Johnson. *Matrix Analysis.* Cambridge University Press, 2nd edition, 1990.

[82] L. Huang and J. Li. Stochastic k-center and j-flat-center problems. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 110–129, 2017.

[83] L. Huang, J. Li, J. M. Phillips, and H. Wang. ε-kernel coresets for stochastic points. In *Proceedings of the 24th Annual European Symposium on Algorithms (ESA)*, pages 50:1–50:18, 2016.

[84] J. H. Huggins, T. Campbell, and T. Broderick. Coresets for scalable Bayesian logistic regression. In *Proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 4080–4088, 2016.

[85] P. Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *Journal of the ACM*, **53**(3):307–323, 2006.

[86] P. Indyk and M. Thorup. Approximate 1-medians. Unpublished manuscript, 2000.

[87] T. S. Jayram, R. Kumar, and D. Sivakumar. The one-way communication complexity of Hamming distance. *Theory of Computing*, **4**(1):129–135, 2008.

[88] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions, Volume 1.* Wiley & Sons, New York, 2nd edition, 1994.

[89] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Multivariate Distributions, Volume 1: Models and Applications.* Wiley & Sons, New York, 2nd edition, 2004.

[90] N. L. Johnson, S. Kotz, and A. W. Kemp. *Univariate Discrete Distributions.* Wiley & Sons, New York, 3rd edition, 2005.

[91] W. B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary Mathematics*, **26**(1):189–206, 1984.

[92] R. Kannan and S. Vempala. Spectral algorithms. *Foundations and Trends in Theoretical Computer Science*, **4**(3-4):157–288, 2009.

[93] R. Kannan, S. Vempala, and D. P. Woodruff. Principal component analysis and higher correlations for distributed data. In *Proceedings of the 27th Conference on Learning Theory (COLT)*, pages 1040–1057, 2014.

[94] M. J. Kearns and U. V. Vazirani. *An Introduction to Computational Learning Theory.* MIT Press, Cambridge, 1994.

[95] K. Kersting, A. Molina, and A. Munteanu. Core dependency networks. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 3820–3827, 2018.

[96] C. Kleiber and S. Kotz. *Statistical Size Distributions in Economics and Actuarial Sciences.* Wiley, Hoboken, 2003.

[97] D. Koller and N. Friedman. *Probabilistic Graphical Models.* MIT Press, Cambridge, 2009.

[98] I. Kremer, N. Nisan, and D. Ron. On randomized one-round communication complexity. *Computational Complexity*, **8**(1):21–49, 1999.

[99] M. Kubat and S. Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *In Proceedings of the 14th International Conference on Machine Learning (ICML)*, pages 179–186, 1997.

[100] A. Kumar, Y. Sabharwal, and S. Sen. Linear-time approximation schemes for clustering problems in any dimensions. *Journal of the ACM*, **57**(2):5:1–5:32, 2010.

[101] E. Kushilevitz and N. Nisan. *Communication complexity.* Cambridge University Press, 1997.

[102] C. Lammersen, M. Schmidt, and C. Sohler. Probabilistic k-median clustering in data streams. In *Proceedings of the 10th International Workshop on Approximation and Online Algorithms (WAOA)*, pages 70–81, 2012.

[103] M. Langberg and L. J. Schulman. Universal ε-approximators for integrals. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 598–607, 2010.

[104] Y. Liang, M. Balcan, V. Kanchanapally, and D. P. Woodruff. Improved distributed principal component analysis. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3113–3121, 2014.

[105] M. Lucic, O. Bachem, and A. Krause. Strong coresets for hard and soft Bregman clustering with applications to exponential family mixtures. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9, 2016.

[106] P. Ma, M. W. Mahoney, and B. Yu. A statistical perspective on algorithmic leveraging. *Journal of Machine Learning Research*, **16**:861–911, 2015.

[107] D. Madigan, N. Raghavan, W. DuMouchel, M. Nason, C. Posse, and G. Ridgeway. Likelihood-based data squashing: A modeling approach to instance construction. *Data Mining and Knowledge Discovery*, **6**(2):173–190, 2002.

[108] P. McCullagh and J. A. Nelder. *Generalized Linear Models.* Chapman & Hall, London, 1989.

[109] C. R. Mehta and N. R. Patel. Exact logistic regression: Theory and examples. *Statistics in Medicine*, **14**(19):2143–2160, 1995.

[110] X. Meng and M. W. Mahoney. Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 91–100, 2013.

[111] H. Minkowski. *Geometrie der Zahlen.* B.G. Teubner, Leipzig, Berlin, 1910.

[112] A. Molina, S. Natarajan, and K. Kersting. Poisson sum-product networks: A deep architecture for tractable multivariate Poisson distributions. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence (AAAI)*, pages 2357–2363, 2017.

[113] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[114] A. Munteanu and C. Schwiegelshohn. Coresets – methods and history: A theoreticians design pattern for approximation and streaming algorithms. *Künstliche Intelligenz*, **32**(1):37–53, 2018.

[115] A. Munteanu, C. Schwiegelshohn, C. Sohler, and D. P. Woodruff. On coresets for logistic regression. *CoRR*, abs/1805.08571, 2018.

[116] S. Muthukrishnan. Data streams: Algorithms and applications. *Foundations and Trends in Theoretical Computer Science*, **1**(2):117–236, 2005.

[117] J. Nelson and H. L. Nguyen. OSNAP: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 117–126, 2013.

[118] J. Nelson and H. L. Nguyen. Sparsity lower bounds for dimensionality reducing maps. In *Proceedings of the 45th Annual ACM Symposium on Theory of Computing (STOC)*, pages 101–110, 2013.

[119] J. Nelson and H. L. Nguyên. Lower bounds for oblivious subspace embeddings. In *Proceedings of the 41st International Colloquium on Automata, Languages, and Programming (ICALP)*, pages 883–894, 2014.

[120] S. Paul, C. Boutsidis, M. Magdon-Ismail, and P. Drineas. Random projections for linear support vector machines. *ACM Transactions on Knowledge Discovery from Data*, **8**(4):22:1–22:25, 2014.

[121] A. Phatak, H. T. Kiiveri, L. H. Clemmensen, and W. J. Wilson. NetRaVE: constructing dependency networks using sparse linear regression. *Bioinformatics*, **26** (12):1576–1577, 2010.

[122] J. M. Phillips. Coresets and sketches. In *Handbook of Discrete and Computational Geometry*, pages 1269–1288. Chapman and Hall/CRC, Boca Raton, 3rd edition, 2017.

[123] H. Poon and P. M. Domingos. Sum-product networks: A new deep architecture. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 337–346, 2011.

[124] S. J. Reddi, B. Póczos, and A. J. Smola. Communication efficient coresets for empirical loss minimization. In *Proceedings of the 31st Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 752–761, 2015.

[125] T. Roughgarden. Beyond worst-case analysis. *Invited talk held at the Highlights of Algorithms conference (HALG),* 2017.

[126] M. Rudelson and R. Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, **54**(4):21, 2007.

[127] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, **71**(2):319–392, 2009.

[128] F. Rusu and A. Dobra. Pseudo-random number generation for sketch-based estimations. *ACM Transactions on Database Systems*, **32**(2):1–48, 2007.

[129] T. Sarlós. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 143–152, 2006.

[130] A. Siddiqa, A. Karim, and A. Gani. Big Data storage technologies: A survey. *Frontiers of Information Technology & Electronic Engineering*, **18**(8):1040–1070, 2017.

[131] C. Sohler and D. 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.

[132] M. T. Subbotin. On the law of frequency of error. *Matematicheskiĭ Sbornik*, **31**(2): 296–301, 1923.

[133] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, **54**(1):45–66, 2004.

[134] M. Thorup. Quick k-median, k-center, and facility location for sparse graphs. *SIAM Journal on Computing*, **34**(2):405–432, 2005.

[135] E. Tolochinsky and D. Feldman. Coresets for monotonic functions with applications to deep learning. *CoRR*, abs/1802.07382, 2018.

[136] J. A. Tropp. Improved analysis of the subsampled randomized hadamard transform. *Advances in Adaptive Data Analysis*, **3**(1-2):115–126, 2011.

[137] I. W. Tsang, J. T. Kwok, and P. Cheung. Core vector machines: Fast SVM training on very large data sets. *Journal of Machine Learning Research*, **6**:363–392, 2005.

[138] V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, 1995.

[139] C. Villani. *Optimal transport: Old and new.* Springer, Berlin, 2009.

[140] J. S. Vitter. Random sampling with a reservoir. *ACM Transactions on Mathematical Software*, **11**(1):37–57, 1985.

[141] R. Wang and D. P. Woodruff. Tight bounds for ℓ_p oblivious subspace embeddings. *CoRR*, abs/1801.04414, 2018.

[142] M. Welling, Y. W. Teh, C. Andrieu, J. Kominiarczuk, T. Meeds, B. Shahbaba, and S. Vollmer. Bayesian inference & Big Data: A snapshot from a workshop. *Bulletin of the International Society for Bayesian Analysis*, **21**(4):8–11, 2014.

[143] R. Winkelmann. *Econometric Analysis of Count Data.* Springer, Berlin, 5th edition, 2008.

[144] D. P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, **10**(1-2):1–157, 2014.

[145] D. P. Woodruff. Sketching as a tool for geometric problems. Invited talk held at TU Dortmund, 2017.

[146] D. P. Woodruff and Q. Zhang. Subspace embeddings and ℓ_p-regression using exponential random variables. In *Proceedings of the 26th Annual Conference on Learning Theory (COLT)*, pages 546–567, 2013.

[147] E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. On graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, **16**:3813–3847, 2015.

[148] J. Yang, X. Meng, and M. W. Mahoney. Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, **104**(1):58–92, 2016.

[149] M. Zhou, L. Li, D. B. Dunson, and L. Carin. Lognormal and gamma mixed negative binomial regression. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.