UNIVERSITY OF CALIFORNIA
SANTA CRUZ

VOLUME SAMPLING FOR LINEAR REGRESSION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Michał Dereziński

June 2018

The Dissertation of Michał Dereziński
is approved:

Professor Manfred K. Warmuth, Chair

Professor S.V.N. Vishwanathan

Professor David P. Helmbold

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Volume sampling for linear regression

by

Michał Dereziński

In this thesis we study the following basic machine learning task: Given a fixed set of n input points in a d-dimensional linear regression problem, we wish to predict a hidden response value for each of the points. We can only afford to attain the responses for a small subset of the points that are then used to construct linear predictions for all points in the dataset. The performance of the predictions is evaluated by the total square loss on all responses. We show that a good approximate solution to this least squares problem can be obtained from just dimension d many responses by using a joint sampling technique called volume sampling. Moreover, the least squares solution obtained for the volume sampled subproblem is an unbiased estimator of optimal solution based on all n responses. This unbiasedness is a desirable property that is not shared by standard subset selection techniques.

Motivated by these basic properties, we develop a theoretical framework for studying volume sampling, which leads to a number of new expectation formulas and statistical guarantees which are of importance not only to least squares regression but also numerical linear algebra in general. Our methods lead to several novel extensions of volume sampling, including a regularized variant, and we propose the first efficient

algorithms which make this technique a practical tool in the machine learning toolbox. Finally, we provide experimental evidence which confirms our theoretical findings.

Acknowledgments

I would first and foremost like to thank my adviser, Manfred Warmuth. His passion for research and remarkable sense of curiosity motivated me to pursue this work from its very beginning.

I had the privilege of taking part in valuable conversations and collaborations with many excellent researchers including S.V.N. Vishwanathan, David Helmbold, S. Sathiya Keerthi, Dhruv Mahajan, Wojciech Kotłowski, Khashayar Rohanimanesh, Suju Rajan and Badri Narayan Bhaskar. I thank them for their help and feedback during my various projects. I would also like to especially thank Daniel Hsu, with whom I was fortunate to have the opportunity to collaborate and whose insights benefited this work.

My friends were an invaluable source of inspiration, and I would like to thank everyone at UC Santa Cruz for going through this journey with me. In particular, I'm grateful to everyone in E2-489 who provided so much support from brainstorming to editing and everything in between.

Finally, my deepest gratitude goes to my parents, as they supported and believed in me throughout all of my studies.

Chapter 1

Introduction

Least squares regression is one of the oldest and most basic learning methods in all of machine learning and statistics, and yet it is still extensively used to this day. We focus on the case when all of the input points are given but obtaining the response values for the points is expensive. As a motivating example, consider the task of optimizing the "activity" of an enzyme [LWG+07] (such as the efficacy at breaking down a certain compound). A large number of variants of an enzyme are considered, and we wish to predict the activity for each of these variants as efficiently as possible. Each variant is described by a feature vector and the simplest model is to assume that the activity (or the response variable) can be modeled as a linear combination of the features. However, obtaining the response value for a variant often involves expensive and lengthy experiments. Thus we ask the following basic question: Is it possible to estimate the least squares predictions for all variants after sampling the responses of only a small number of them? What is the smallest number of responses yielding useful results? Intuitively

the subset of chosen variants for which we will measure the response values should be "diverse". In answering these questions, we will demonstrate a fundamental new connection between linear least squares and a joint sampling distribution for producing diverse subsets called "volume sampling".

1.1 From one-dimensional regression to volume sampling

As an introductory case, consider linear regression in one dimension. We are given n points x_i. Each point has a hidden real response (or target value) y_i. Assume that obtaining the responses is expensive and the learner can afford to request the responses y_i for only a small number of indices i. After receiving the requested responses, the learner determines an approximate linear least squares solution. In the one dimensional case this is just a single weight (for simplicity, we omit the additional bias term). How many response values does the learner need to request so that the total square loss of its approximate solution on all n points is "close" to the total loss of the optimal linear least squares solution found with the knowledge of all responses? We will show that one response suffices if the index i is chosen proportional to x_i^2. When the learner uses the approximate solution $w_i^* = \frac{y_i}{x_i}$, then its expected loss

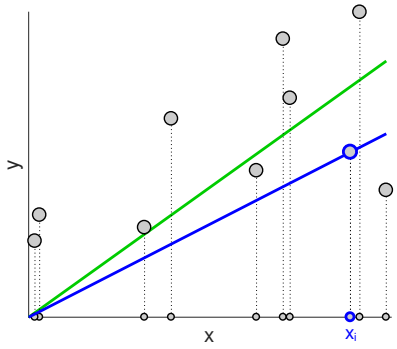

Figure 1.1: The expected loss of $w_i^* = \frac{y_i}{x_i}$ (blue line) based on one response y_i is twice the loss of the optimum w^* (green line).

equals 2 times the loss of the optimum w^* that is computed based on all responses (See Figure 1.1). Moreover, the approximate solution w_i^* is an unbiased estimator for the optimum w^*:

$$\mathbb{E}_i\left[\sum_j (x_j \frac{y_i}{x_i} - y_j)^2\right] = 2\sum_j (x_j w^* - y_j)^2 \quad \text{and} \quad \mathbb{E}_i\left[\frac{y_i}{x_i}\right] = w^*, \quad \text{when } P(i) \sim x_i^2.$$

Note that there are no range restrictions on the points and response values. Also, randomization is necessary to achieve this loss equation because for any deterministic algorithm, the total loss based on a single response can be up to n times the optimum: An instance of this occurs when all n points x_i are equal 1 and all responses are also 1, except for the response of index picked by the deterministic algorithm which is set to 0.

Both of the above equations generalize to the case when the points $\mathbf{x}_i$ lie in $\mathbb{R}^d$. Let $\mathbf{X}$ denote the $n \times d$ matrix that has the n transposed points $\mathbf{x}_i^\top$ as rows, and let $\mathbf{y} \in \mathbb{R}^n$ be the vector of responses. Now the goal is to minimize the (total) square loss

$$L(\mathbf{w}) = \sum_{i=1}^n (\mathbf{x}_i^\top \mathbf{w} - y_i)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2,$$

over all linear weight vectors $\mathbf{w} \in \mathbb{R}^d$. Let $\mathbf{w}^*$ denote the optimal such weight vector. We want to minimize the square loss based on a small number of responses we attained for a subset of rows. Again, the learner is initially given the fixed set of n rows (i.e. fixed design), but none of the responses. It is then allowed to choose a random subset of d indices, $S \subseteq \{1..n\}$, and obtains the responses for the corresponding d rows. The learner proceeds to find the optimal linear least squares solution $\mathbf{w}^*(S)$ for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. where $\mathbf{X}_S$ is the subset of d rows of $\mathbf{X}$ indexed by S and $\mathbf{y}_S$ the corresponding d responses from the response vector $\mathbf{y}$. As a generalization of the one-dimensional

distribution that chooses an index based on the squared length, set S of size d is chosen proportional to the squared volume of the parallelepiped spanned by the rows of $\mathbf{X}_S$. This squared volume equals $\det(\mathbf{X}_S^\top \mathbf{X}_S)$. Using elementary linear algebra, we will show that *volume sampling* the set S assures that $\mathbf{w}^*(S)$ is a good approximation to $\mathbf{w}^*$ in the following sense: In expectation, the square loss (on all n row response pairs) of $\mathbf{w}^*(S)$ is equal $d+1$ times the square loss of $\mathbf{w}^*$, and moreover, the unbiasedness property of estimator $\mathbf{w}^*(S)$ is retained:

$$\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*) \quad \text{and} \quad \mathbb{E}[\mathbf{w}^*(S)] = \mathbf{w}^*, \quad \text{when } P(S) \sim \det(\mathbf{X}_S^\top \mathbf{X}_S).$$

The discovery of these fundamental matrix expectation formulas was our primary motivation for studying volume sampling in the context of linear regression. In this work, we show many other results which offer a direct connection of volume sampling to not just least squares, but also other even more basic concepts in linear algebra such as the matrix pseudoinverse. We also develop several extensions of this joint sampling distribution which have additional useful properties, such as sampling sets S of arbitrary fixed size, introducing regularization, and reweighting the instances. Finally, we develop several efficient algorithms for volume sampling which make this procedure a practical tool for machine learning for the first time.

1.2 Related work

Determinantal sampling methods and applications. Volume sampling is a type of determinantal point process (DPP) [KT12]. DPP's have been given a lot of attention

in the literature due to their ability to produce diverse subsets of data, with many applications to machine learning, including recommendation systems [KT11, GPK16], clustering [Kan13], computer vision [KT10], matrix approximation [DRVW06, DR10, AB13], fairness [CDKV16, CKS+18] and optimization [ZKM17]. Those methods are typically concerned with sampling sets of size no more than the dimension d, of either fixed size or variable size.

Efficiency of volume sampling. Two primary types of volume sampling have been considered in the literature. The first one, proposed by [DRVW06], samples sets of fixed size $s \leq d$ proportionally to the squared *row*-volume of the submatrix $\mathbf{X}_S$. The second one, proposed by [AB13], samples sets of size $s \geq d$ proportionally to the squared *column*-volume of $\mathbf{X}_S$. Note that when $s = d$, those two definitions coincide. Our primary interest is concentrated on the sampling of sets of size $s \geq d$, i.e. of *at least* dimension many rows. However, due to the important overlapping case of $s = d$, we discuss algorithms for both settings. The first polynomial time algorithm for $s \leq d$ volume sampling was given by [DR10], and then improved by [GS12], running in time $O(nd^2s)$. An approximate sampling procedure was also developed by [DR10], however it is only useful for $s \ll d$. Our algorithms apply to this line of work only for the case of $s = d$. In this case, they do offer significant improvement over state-of-the-art (by a factor of d), with running time $O(nd^2)$.

In this thesis, we focus on volume sampling sets of size $s \geq d$, which was proposed by [AB13] and motivated with applications in graph theory, linear regression,

matrix approximation and more. Until very recently, there was no polynomial time algorithm for this type of volume sampling (apart from the case of $s = d$). The only known polynomial time algorithm for size $s > d$ volume sampling developed prior to this work was proposed by [LJS17] with time complexity $O(n^4 s)$. In Chapter 3 we propose an algorithm which runs in time $O(nd^2)$ (independent of the choice of s), which is faster by a factor of at least n^2. In Chapter 4, we develop a new rescaled variant of volume sampling, which after a preprocessing time of $\widetilde{O}(nd + d^3)$ produces a sample in time $\widetilde{O}((d^2 + s)d^2)$, which is considerably faster than $O(nd^2)$ for large enough n.

Volume sampling for matrix approximation. In the field of computational geometry a variant of volume sampling was used to obtain optimal bounds for low-rank matrix approximation. In this task, the goal is to select a small subset of rows of a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ (much fewer than the rank of $\mathbf{X}$, which is bounded by d), so that a good low-rank approximation of $\mathbf{X}$ can be constructed from those rows. [DRVW06] showed that volume sampling of size $s < d$ index sets obtains optimal multiplicative bounds for this task and [GS12] used that result to obtain even more effective matrix approximation algorithms. We show in this paper that for linear regression, fewer than rank many rows do not suffice to obtain multiplicative bounds. This is why we focus on volume sampling sets of size $s \geq d$.

Subset selection for linear regression. The problem of selecting a subset of the rows of the input matrix for solving a linear regression task has been extensively studied in statistics literature under the terms *optimal design* [Fed72] and *pool-based active learn-*

ing [SN09]. Various criteria for subset selection have been proposed, like A-optimality and D-optimality. For example, A-optimality seeks to minimize $\mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$, which is combinatorially hard to optimize exactly. Volume sampling was recently found to be a useful technique for approximately satisfying the A-optimality criterion [AZLSW17, NST18], motivated by the result of [AB13], bounding the expectation of the trace $\mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1})$ under volume sampling. In Chapter 2, we generalize this result to a matrix expectation formula using a new proof technique, obtaining that $\mathbb{E}[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}] = \frac{n-d+1}{s-d+1} (\mathbf{X}^\top \mathbf{X})^{-1}$. This expectation formula provides an approximate randomized solution for the sampled inverse covariance matrix rather than just its trace.

Approximately solving linear regression. Computing approximate solutions to linear regression has been explored in the domain of numerical linear algebra (see [Mah11] for an overview). Here, multiplicative bounds on the loss of the approximate solution can be achieved via two approaches. The first approach relies on "sketching" the input matrix $\mathbf{X}$ and the response vector $\mathbf{y}$ by multiplying both by the same suitably chosen random matrix. Algorithms which use sketching to generate a smaller input matrix for a given linear regression problem are computationally efficient [Sar06, CW13], but they require all of the responses from the original problem to generate the sketch and are thus not suitable for the goal of using as few response values as possible. The second approach is based on subsampling the rows of the input matrix and only asking for the responses of the sampled rows. The learner optimally solves the sampled subproblem and then uses the obtained weight vector for its prediction on all rows.

The selected subproblem is known under the term "**b**-agnostic minimal coreset" in [BDM13, DMM08], since it is selected without knowing the response vector (denoted as the vector **b**). This line of work is mostly based on i.i.d. sampling using the statistical leverage scores [DMIMW12]. Unlike volume sampling, i.i.d. leverage score sampling does not produce unbiased estimators for linear regression. Moreover, it suffers from coupon collector problem which prohibits any effective approximation guarantees for samples size smaller than $d \log d$. In this work we provide experimental results showing that for small sample sizes volume sampling is much more effective than leverage score sampling. A different and more elaborate sampling technique based on spectral sparsification [BSS12, LS15] was recently shown to be effective for linear regression [CP17], however this method also does not produce unbiased estimates, which is a primary concern of this work and desirable in many settings. Unbiasedness seems to require delicate control of the sampling probabilities, which we achieve using volume sampling.

Unbiased estimates for linear regression. First, we should address a potential confusion with the statistical term *best linear unbiased estimator* (BLUE). According to Gauss-Markov theorem, the least squares estimator $\mathbf{w}^* = \mathbf{X}^+\mathbf{y}$ is the BLUE estimator. However, it is unbiased *only* in the following narrow sense: if we assume that the responses are generated from a hidden linear transformation plus random noise which has mean zero, then the expectation of $\mathbf{w}^*$ w.r.t. this response noise is equal to the hidden weight vector defining the linear transformation. On the other hand, in the context of our work, unbiasedness refers to an estimator $\mathbf{w}(S)$ produced based on a

random subset $S \subseteq \{1..n\}$ of responses (the randomness coming from the sampling distribution for choosing S), with the response vector being *arbitrary and fixed*. This notion of unbiasedness demands that $\mathbb{E}_S[\mathbf{w}(S)] = \mathbf{w}^*$ and it is much harder to obtain. To our knowledge, volume sampling is the only known non-trivial method of producing such unbiased estimators. For the case of $s = d$, this fact was known in the linear algebra community [BTT90, BI92] as a mathematical identity (long before volume sampling was considered as a sampling procedure). We independently showed this result using a new proof technique (Chapter 2), extending it to volume sampling of any size $s \geq d$, and also introducing new unbiased estimators with strong loss bounds (Chapter 4). This is achieved by a new rescaled variant of volume sampling.

1.3 Overview of the chapters

We now sketch the contents of the main chapters of this dissertation (each chapter will have a separate more detailed introduction). The main results of Chapters 2 and 3 were published at NIPS'17 and AISTATS'18 conferences [DW17, DW18b] and will appear together as a JMLR paper [DW18a]. The results of Chapter 4 are based on a manuscript that is currently in submission [DWH18].

1.3.1 Chapter 2: Unbiased pseudoinverse estimator

We propose a matrix estimator for the pseudoinverse $\mathbf{X}^+$, computed from a small subset of rows of the matrix $\mathbf{X}$. When the subset is sampled according volume sampling, the estimator is unbiased and its covariance also has a closed form: It equals a

specific factor times $\mathbf{X}^{+}\mathbf{X}^{+\top}$. Our analysis for computing matrix expectations is based on a general framework we call reverse iterative sampling, which is developed in this chapter.

These new formulas establish a fundamental connection between volume sampling and linear least squares, because the least squares solution obtained for the volume sampled subproblem is an unbiased estimator of optimal solution based on all n responses. Moreover, a good approximate solution can be obtained from just dimension d many responses. Concretely, we show that if the rows are in general position and if a subset of d rows is chosen proportional to the squared volume spanned by those rows, then the expected total square loss (on all n rows) of the least squares solution found for the subset is exactly $d+1$ times the minimum achievable total loss. We provide lower bounds showing that the factor of $d+1$ is optimal, and any i.i.d. row sampling procedure requires $\Omega(d \log d)$ responses to achieve a finite factor guarantee.

1.3.2 Chapter 3: Regularized volume sampling

Given n vectors $\mathbf{x}_i \in \mathbb{R}^d$, we want to fit a linear regression model for noisy responses $y_i \in \mathbb{R}$. The ridge estimator is a classical solution to this problem. We propose a new regularized variant of volume sampling and show that the ridge estimator obtained from a subset selected with this procedure offers strong statistical guarantees in terms of the mean squared prediction error over the entire dataset of n vectors. The number of responses needed is proportional to the statistical dimension of the problem which is often much smaller than d. A second major contribution is that we speed up volume

sampling so that it is essentially as efficient as leverage scores, which is the main i.i.d. subsampling procedure for this task. Finally, we show theoretically and experimentally that volume sampling outperforms any i.i.d. sampling when responses are expensive.

1.3.3 Chapter 4: Leveraged volume sampling

Volume sampling has a unique and desirable property that the least squares weight vector it produces is an unbiased estimate of the optimum. It is therefore natural to ask if this method offers the best unbiased estimator in terms of the number of responses s needed to achieve a $1+\epsilon$ loss approximation. In this chapter, we show that standard volume sampling can have poor behavior when we require a very accurate approximation of a linear least squares problem – indeed worse than i.i.d. leverage score sampling, whose estimates are biased. We then develop a new rescaled variant of volume sampling that produces an unbiased estimator which avoids this bad behavior and has at least as good a tail bound as leverage score sampling: sample size $s = O(d \log d + d/\epsilon)$ suffices to guarantee total loss at most $1+\epsilon$ times the minimum with high probability. Thus, we improve on the best known sample size for an unbiased estimator, $s = O(d^2/\epsilon)$, constructed in Chapter 2 using standard volume sampling.

Our rescaling procedure leads to a new efficient algorithm for volume sampling which is based on a *determinantal rejection sampling* technique with potentially broader applications to determinantal point processes. Other contributions include introducing the combinatorics needed for rescaled volume sampling and developing tail bounds for sums of dependent random matrices which arise in the process.

Chapter 2

Unbiased pseudoinverse estimator

2.1 Introduction

Given a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, with $n \geq d$, suppose that our goal is to estimate the pseudoinverse $\mathbf{X}^+$ of $\mathbf{X}$ based on the pseudoinverse of a subset of rows. For a subset $S \subseteq \{1..n\}$ of s row indices (where the size s is fixed and $s \geq d$), we let $\mathbf{X}_S$ be the submatrix of the s rows indexed by S (see Figure 2.1). Consider a version of $\mathbf{X}$ in which all but the rows of S are zero. This matrix equals $\mathbf{I}_S\mathbf{X}$ where $\mathbf{I}_S$ is an n-dimensional diagonal matrix with $(\mathbf{I}_S)_{ii} = 1$ if $i \in S$ and 0 otherwise. We show a number of expectation formulas related to matrix $(\mathbf{I}_S\mathbf{X})^+$, treated as an estimator of pseudoinverse $\mathbf{X}^+$, when S is sampled from

$$\text{size } s \text{ volume sampling:} \quad P(S) \sim \det(\mathbf{X}_S^\top \mathbf{X}_S) \quad \text{where} \quad |S| = s.$$

For this type of sampling of the set S, we will prove that:

$$\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+] = \mathbf{X}^+ \quad \text{and} \quad \mathbb{E}[\underbrace{(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}}_{(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}}] = \frac{n-d+1}{s-d+1}\underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathbf{X}^+\mathbf{X}^{+\top}}.$$

Note that $(\mathbf{I}_S\mathbf{X})^+$ has the $d \times n$ shape of $\mathbf{X}^+$ where the s columns indexed by S contain $(\mathbf{X}_S)^+$ and the remaining $n-s$ columns are zero. The expectation of this matrix is $\mathbf{X}^+$ even though $(\mathbf{X}_S)^+$ is clearly not a submatrix of $\mathbf{X}^+$. The second expectation formula can be viewed as a second moment of the pseudoinverse estimator $(\mathbf{I}_S\mathbf{X})^+$, and it can be used to compute a useful notion of matrix variance with applications in random matrix theory:

$$\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}] - \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]^\top = \frac{n-s}{s-d+1}\mathbf{X}^+\mathbf{X}^{+\top}.$$

We prove the above expectation formulas using a general framework of reverse iterative sampling which we develop in this chapter. This technique also leads to efficient volume sampling algorithms, presented in Chapter 3, which beat the state-of-the-art by a factor of n^2 in time complexity, and make volume sampling nearly as efficient as the comparable i.i.d. sampling technique called leverage score sampling.

There is a direct connection between the pseudoinverse and solving linear least squares problems: recall that for an $n-$dimensional response vector $\mathbf{y}$, the optimal solution to the least squares problem $(\mathbf{X}, \mathbf{y})$ is $\mathbf{w}^* = \text{argmin}_{\mathbf{w}} ||\mathbf{X}\mathbf{w} - \mathbf{y}||^2 = \mathbf{X}^+\mathbf{y}$. Similarly $\mathbf{w}^*(S) = (\mathbf{X}_S)^+\mathbf{y}_S$ is the solution for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. The above expectation formula now implies that for any size $s \geq d$, if S of size s is drawn by

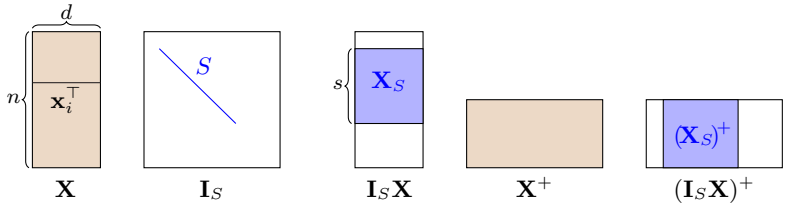

Figure 2.1: Shapes of the matrices. The indices of S may not be consecutive.

volume sampling, then $\mathbf{w}^*(S)$ is an unbiased estimator[1] for $\mathbf{w}^*$, i.e.

$$\mathbb{E}[\mathbf{w}^*(S)] = \mathbb{E}[(\mathbf{X}_S)^+\mathbf{y}_S] = \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+\mathbf{y}] = \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]\,\mathbf{y} = \mathbf{X}^+\mathbf{y} = \mathbf{w}^*.$$

Moreover, under the additional assumption that response vector $\mathbf{y}$ is generated by a linear transformation distorted with i.i.d. white noise (see Section 2.2.3 for details), the expectation formula for $(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}$ leads to an exact formula for the covariance matrix of the estimator $\mathbf{w}^*(S)$, as well as an approximate solution to the classical A-optimality criterion in optimal design.

For volume sampling of size $s = d$ we show an additional formula which relates the expected loss of $\mathbf{w}^*(S)$ to the loss of the best for a fixed hidden response vector $\mathbf{y}$. Namely, when matrix $\mathbf{X}$ is in general position and set S is volume sampled with $s = d$, we have

$$\mathbb{E}\big[L(\mathbf{w}^*(S))\big] = (d+1)L(\mathbf{w}^*), \qquad \text{where} \quad L(\mathbf{w}) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

We show that the above formula gives the best achievable multiplicative approximation factor for the least squares loss when using only dimension many responses. Our other lower bounds suggest that by its joint nature volume sampling is uniquely well suited for linear regression when a small number of responses is desired.

Outline of the chapter. In the next section, we define volume sampling as an instance of a more general procedure we call reverse iterative sampling, and we use this methodology to prove closed form matrix expressions for the expectation of the pseudoinverse estimator $(\mathbf{I}_S\mathbf{X})^+$ and its square $(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}$, when S is sampled by

[1]For size $s = d$ volume sampling, the fact that $\mathbb{E}[\mathbf{w}^*(S)] = \mathbf{w}^*$ can be found in an early paper [BTT90]. They give a direct proof based on Cramer's rule.

volume sampling. Central to volume sampling is the Cauchy-Binet formula for determinants. As a side, we produce a number of short self-contained proofs for this formula and show that leverage scores are the marginals of volume sampling. Then in Section 2.3 we formulate the problem of solving linear regression from a small number of responses, and state the upper bound for the expected square loss of the volume sampled least squares estimator (Theorem 2.6), followed by a discussion and related lower-bounds. In Section 2.4, we prove Theorem 2.6 and an additional related matrix expectation formula. As a side note, Section 2.5 presents an alternate proof technique which utilizes matrix differentials to show the unbiasedness of our pseudoinverse estimator (Theorem 2.3). Finally, Section 2.6 concludes the chapter by framing a number of open problems about unbiased estimators for linear regression.

2.2 Reverse iterative sampling

Let n be an integer dimension. For each subset $S \subseteq \{1..n\}$ of size s we are given a matrix formula $\mathbf{F}(S)$. Our goal is to sample set S of size s using some sampling process and then develop concise expressions for $\mathbb{E}_{S:|S|=s}[\mathbf{F}(S)]$. Examples of formula classes $\mathbf{F}(S)$ will be given below.

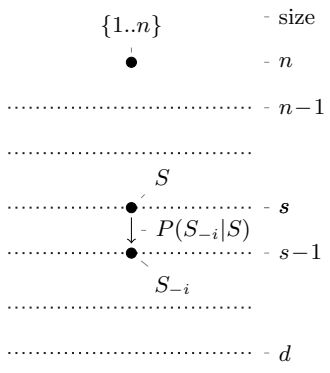

Figure 2.2:

Reverse iterative sampling.

We represent the sampling by a directed acyclic graph (DAG), with a single root node corresponding to

the full set $\{1..n\}$. Starting from the root, we proceed along the edges of the graph, iteratively removing elements from the set S (see Figure 2.2). Concretely, consider a DAG with levels $s = n, n-1, ..., d$. Level s contains $\binom{n}{s}$ nodes for sets $S \subseteq \{1..n\}$ of size s. Every node S at level $s > d$ has s directed edges to the nodes $S_{-i} = S \backslash \{i\}$ at the next lower level. These edges are labeled with a conditional probability vector $P(S_{-i}|S)$, where the event S occurs if the sampling process visits node S as it traces a (directed) path in the DAG from the root node $\{1..n\}$ to a node at level d. Such paths have $n - d$ edges. It is natural to assign probabilities to shorter paths as well going from any node to a node at a lower level. The probability of such paths is again the product of its edge probabilities. It also follows that the probability $P(S)$ of visiting node S (via a path from the root) is the sum of the probabilities of all paths from root to S. Finally, the probability $P(\{1..n\})$ of the root node is 1 and more generally, the total probability of all nodes at each layer is 1.

We associate a formula $\mathbf{F}(S)$ with each set node S in the DAG. The following key equality lets us compute expectations.

Lemma 2.1. *If for all $S \subseteq \{1..n\}$ of size greater than d we have*

$$\mathbf{F}(S) = \sum_{i \in S} P(S_{-i}|S)\mathbf{F}(S_{-i}),$$

then for any $s \in \{d..n\}$: $\mathbb{E}_{S:|S|=s}[\mathbf{F}(S)] = \sum_{S:|S|=s} P(S)\mathbf{F}(S) = \mathbf{F}(\{1..n\})$.

Proof. Suffices to show that expectations at successive layers s and $s-1$ are equal for

$s > d$:

$$\begin{aligned}\sum_{S:|S|=s} P(S)\,\mathbf{F}(S) &= \sum_{S:|S|=s} P(S)\sum_{i\in S} P(S_{-i}|S)\,\mathbf{F}(S_{-i}) \\ &= \sum_{S:|S|=s}\ \sum_{i\in S} P(S)P(S_{-i}|S)\mathbf{F}(S_{-i}) \\ &= \sum_{T:|T|=s-1} \underbrace{\sum_{j\notin T} P(T_{+j})P(T|T_{+j})}_{P(T)}\,\mathbf{F}(T).\end{aligned}$$

Note that the r.h.s. of the first line has one summand per edge leaving level s, and the r.h.s. of the last line has one summand per edge arriving at level $s-1$. Now the last equality holds because the edges leaving level s are exactly those arriving at level $s-1$, and the summand for each edge in both expressions is equivalent. ■

2.2.1 Volume sampling

Given a tall full rank matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a sample size $s \in \{d..n\}$, volume sampling chooses subset $S \subseteq \{1..n\}$ of size s with probability proportional to squared volume spanned by the columns of submatrix[2] $\mathbf{X}_S$ and this squared volume equals $\det(\mathbf{X}_S^\top \mathbf{X}_S)$. The following theorem uses the above DAG setup to compute the normalization constant for this distribution. Note that all subsets S of volume 0 will be ignored, since they are unreachable in the proposed sampling procedure.

Theorem 2.1. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$, where $d \le n$ and $\det(\mathbf{X}^\top\mathbf{X}) > 0$. For any set S of size $s > d$ for which $\det(\mathbf{X}_S^\top\mathbf{X}_S) > 0$, define the probability of the edge from S to S_{-i} for*

[2]For sample size $s = d$, the rows and columns of $\mathbf{X}_S$ have the same length and $\det(\mathbf{X}_S^\top\mathbf{X}_S)$ is also the squared volume spanned by the rows $\mathbf{X}_S$.

$i \in S$ *as:*

$$P(S_{-i}|S) \stackrel{def}{=} \frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}})}{(s-d)\det(\mathbf{X}_S^\top \mathbf{X}_S)} = \frac{1-\mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i}{s-d},$$

(reverse iterative volume sampling)

where $\mathbf{x}_i$ *is the* i*th row of* $\mathbf{X}$*. In this case* $P(S_{-i}|S)$ *is a proper probability distribution. If* $\det(\mathbf{X}_S^\top \mathbf{X}_S) = 0$*, then simply set* $P(S_{-i}|S)$ *to* $\frac{1}{s}$*. With these definitions,* $\sum_{S:|S|=s} P(S) = 1$ *for all* $s \in \{d..n\}$ *and the probability of all paths from the root to any subset* S *of size at least* d *is*

$$P(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{s-d}\det(\mathbf{X}^\top \mathbf{X})}. \qquad \textbf{(volume sampling)}$$

The rewrite of the ratio $\frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}})}{\det(\mathbf{X}_S^\top \mathbf{X}_S)}$ as $1 - \mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i$ is Sylvester's Theorem for determinants. Incidentally, this is the only property of determinants used in this section.

Note also that the theorem implies the following formula:

$$\sum_{S:|S|=s} \det(\mathbf{X}_S^\top \mathbf{X}_S) = \binom{n-d}{s-d} \det(\mathbf{X}^\top \mathbf{X}).$$

When $s = d$, then the binomial coefficient is 1 and the above becomes the Cauchy-Binet formula. The below proof thus results in a minimalist proof of this classical formula. It uses the reverse iterative sampling (Figure 2.2) and proceeds by showing the fact that all paths from the root to node S have the same probability. For the sake of completeness we also give a more direct inductive proof of the above formula in Section 2.2.2.

Proof. First, for any node S st $|S| = s > d$ and $\det(\mathbf{X}_S^\top \mathbf{X}_S) > 0$, the probabilities out

of S sum to 1:

$$\sum_{i\in S} P(S_{-i}|S) = \sum_{i\in S} \frac{1-\mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top)}{s-d} = \frac{s-\mathrm{tr}((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}_S^\top \mathbf{X}_S)}{s-d} = \frac{s-d}{s-d} = 1.$$

It remains to show the formula for the probability $P(S)$ of all paths visiting node S. If $\det(\mathbf{X}_S^\top \mathbf{X}_S) = 0$, then one edge on any path from the root to S has probability 0. This edge goes from a superset of S with positive volume to a superset of S that has volume 0. Since all paths have probability 0, $P(S) = 0$ in this case.

Now assume $\det(\mathbf{X}_S^\top \mathbf{X}_S) > 0$ and consider any path from the root $\{1..n\}$ to S. There are $(n-s)!$ such paths all going through sets with positive volume. The fractions of determinants in the probabilities along each path telescope and the additional factors accumulate to the same product. So the probability of all paths from the root to S is the same and the total probability into S is

$$\frac{(n-s)!}{(n-d)\dots(s-d+1)} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\det(\mathbf{X}^\top \mathbf{X})} = \frac{1}{\binom{n-d}{s-d}} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\det(\mathbf{X}^\top \mathbf{X})}. \qquad \blacksquare$$

An immediate consequence of the above sampling procedure is the following property of volume sampling, which states that this distribution is closed under subsampling. We also give a direct proof to highlight the combinatorics of volume sampling.

Corollary 2.1. *For any $\mathbf{X} \in \mathbb{R}^{n\times d}$ and $n \geq t > s \geq d$, the following hierarchical sampling procedure:*

$$T \overset{t}{\sim} \mathbf{X} \qquad \textit{(size } t \textit{ volume sampling from } \mathbf{X}\textit{)},$$

$$S \overset{s}{\sim} \mathbf{X}_T \qquad \textit{(size } s \textit{ volume sampling from } \mathbf{X}_T\textit{)}$$

returns a set S which is distributed according to size s volume sampling from $\mathbf{X}$.

Proof. We start with the Law of Total Probability and then use the probability formula for volume sampling from the above theorem. Here $P(T \cap S)$ means the probability of all paths going through node T at level t and then node S at level s. If $S \not\subseteq T$, then $P(T \cap S) = 0$.

$$\begin{aligned} P(S) &= \sum_{T:\, S\subseteq T} \overbrace{P(S \,|\, T) \qquad\qquad P(T)}^{P(T\cap S)} \\ &= \sum_{T:\, S\subseteq T} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{t-d}{s-d}\cancel{\det(\mathbf{X}_T^\top \mathbf{X}_T)}} \quad \frac{\cancel{\det(\mathbf{X}_T^\top \mathbf{X}_T)}}{\binom{n-d}{t-d}\det(\mathbf{X}^\top \mathbf{X})} \\ &= \binom{n-s}{t-s} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{t-d}{s-d}\binom{n-d}{t-d}\det(\mathbf{X}^\top \mathbf{X})} = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{s-d}\det(\mathbf{X}^\top \mathbf{X})}. \end{aligned}$$

Note that for all sets T containing S, the probability $P(T \cap S)$ is the same, and there are $\binom{n-s}{t-s}$ such sets. ∎

The main competitor of volume sampling is i.i.d. sampling of the rows of $\mathbf{X}$ w.r.t. the statistical leverage scores. For an input matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, the leverage score of the i-th row $\mathbf{x}_i^\top$ of $\mathbf{X}$ is defined as

$$l_i \overset{def}{=} \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i.$$

Recall that this quantity appeared in the definition of conditional probability $P(S_{-i}|S)$ in Theorem 2.1, where the leverage score was computed w.r.t. the submatrix $\mathbf{X}_S$. In fact, there is a more basic relationship between leverage scores and volume sampling: If set S is sampled according to size $s = d$ volume sampling, then the leverage score l_i of row i is the marginal probability $P(i \in S)$ of including the i-th row into S. A general formula for the marginals of size s volume sampling is given in the following proposition:

Proposition 2.1. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full rank matrix and $s \in \{d..n\}$. If $S \subseteq \{1..n\}$ is sampled according to size s volume sampling, then for any $i \in \{1..n\}$,*

$$P(i \in S) = \frac{s-d}{n-d} + \frac{n-s}{n-d}\overbrace{\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i}^{l_i}.$$

Proof. Instead of $P(i \in S)$ we will first compute $P(i \notin S)$:

$$\begin{aligned}
P(i \notin S) &= \sum_{S:|S|=s, i\notin S} \frac{\det(\mathbf{X}_S^\top\mathbf{X}_S)}{\binom{n-d}{s-d}\det(\mathbf{X}^\top\mathbf{X})} \\
&= \sum_{S:|S|=s, i\notin S} \frac{\sum_{T\subseteq S:|T|=d}\det(\mathbf{X}_T^\top\mathbf{X}_T)}{\binom{n-d}{s-d}\det(\mathbf{X}^\top\mathbf{X})} \\
&= \frac{\binom{n-d-1}{s-d}\overbrace{\sum_{T:|T|=d, i\notin T}\det((\mathbf{X}_{-i})_T^\top(\mathbf{X}_{-i})_T)}^{\det(\mathbf{X}_{-i}^\top\mathbf{X}_{-i})}}{\binom{n-d}{s-d}\det(\mathbf{X}^\top\mathbf{X})} \\
&= \frac{n-s}{n-d}\,(1 - \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i),
\end{aligned}$$

where we used Cauchy-Binet twice and the fact that every set $T : |T| = d, i \notin T$ appears in $\binom{n-d-1}{s-d}$ sets $S : |S| = s, i \notin S$. Now, the marginal probability follows from the fact that $P(i \in S) = 1 - P(i \notin S)$. ∎

2.2.2 Inductive proof of Cauchy-Binet

The most common form of the Cauchy-Binet equation deals with two real $n \times d$ matrices $\mathbf{A}, \mathbf{B}$: $\sum_{S\,:\,|S|=d}\det(\mathbf{A}_S^\top\mathbf{B}_S) = \det(\mathbf{A}^\top\mathbf{B})$. It is easy to generalize volume sampling and Theorem 2.1 to this "asymmetric" version. Here we give an alternate inductive proof.

Theorem 2.2. *For* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times d}$ *and* $n - 1 \geq s \geq d$ *:*

$$\det(\mathbf{A}^\top \mathbf{B}) = \frac{1}{\binom{n-d}{s-d}} \sum_{S\,:\,|S|=s} \det(\mathbf{A}_S^\top \mathbf{B}_S).$$

Proof. For $i \in \{1..n\}$, let $\mathbf{a}_i, \mathbf{b}_i$ denote the i-th row of $\mathbf{A}, \mathbf{B}$, respectively. S is a size s subset of $\{1..n\}$. We rewrite the range restriction $n - 1 \geq s \geq d$ for size s as $1 \leq n - s \leq n - d$ and induct on $n - s$. For the base case, $n - s = 1$ or $s = n - 1$, we need to show that

$$\det(\mathbf{A}^\top \mathbf{B}) = \frac{1}{n-d} \sum_{i=1}^{n} \det(\mathbf{A}_{-i}^\top \mathbf{B}_{-i}).$$

This clearly holds if $\det(\mathbf{A}^\top \mathbf{B}) = 0$. Otherwise, by Sylvester's Theorem

$$\sum_{i=1}^{n} \frac{\det(\overbrace{\mathbf{A}_{-i}^\top \mathbf{B}_{-i}}^{\mathbf{A}^\top \mathbf{B} - \mathbf{a}_i \mathbf{b}_i^\top})}{\det(\mathbf{A}^\top \mathbf{B})} = \sum_{i=1}^{n} (1 - \mathbf{a}_i^\top (\mathbf{A}^\top \mathbf{B})^{-1} \mathbf{b}_i) = n - \overbrace{\operatorname{tr}((\mathbf{A}^\top \mathbf{B})^{-1} \mathbf{A}^\top \mathbf{B})}^{d}.$$

Induction: Assume $2 \leq n - s \leq n - d$.

$$\begin{aligned}
\det(\mathbf{A}^\top \mathbf{B}) &= \frac{1}{n-d} \sum_{i=1}^{n} \det(\mathbf{A}_{-i}^\top \mathbf{B}_{-i}) \\
&= \frac{1}{n-d} \sum_{i=1}^{n} \sum_{S\,:\,|S|=s,\, i \notin S} \frac{1}{\binom{n-1-d}{s-d}} \det(\mathbf{A}_S^\top \mathbf{B}_S) \\
&= \underbrace{\frac{n-s}{n-d} \frac{1}{\binom{n-1-d}{s-d}}}_{\frac{1}{\binom{n-d}{s-d}}} \sum_{S\,:\,|S|=s} \det(\mathbf{A}_S^\top \mathbf{B}_S).
\end{aligned}$$

Note that for the induction step, S is a subset of size s from a set of size $n - 1$ and we have the range restriction $1 \leq n - 1 - s \leq n - 1 - d$. Clearly, $n - 1 - s$ is one smaller than $n - s$. For the last equality, notice that each set $S \subseteq \{1..n\} : |S| = s$ is counted $n - s$ times in the double sum. ∎

2.2.3 Expectation formulas for volume sampling

All expectations in the remainder of the chapter are w.r.t. volume sampling. We use the short-hand $\mathbb{E}[\mathbf{F}(S)]$ for expectation with volume sampling where the size of the sampled set is fixed to s. The expectation formulas for two choices of $\mathbf{F}(S)$ are proven in Theorems 2.3 and 2.4. By Lemma 2.1 it suffices to show $\mathbf{F}(S) = \sum_{i \in S} P(S_{-i}|S)\mathbf{F}(S_{-i})$ for volume sampling. We also present a related expectation formula (Theorem 2.5), which is proven later using different techniques.

Recall that $\mathbf{X}_S$ is the submatrix of rows indexed by $S \subseteq \{1..n\}$. We also use a version of $\mathbf{X}$ in which all but the rows of S are zeroed out. This matrix equals $\mathbf{I}_S\mathbf{X}$ where $\mathbf{I}_S$ is an n-dimensional diagonal matrix with $(\mathbf{I}_S)_{ii} = 1$ if $i \in S$ and 0 otherwise (see Figure 2.1).

Theorem 2.3. *Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a tall full rank matrix (i.e. $n \geq d$). For $s \in \{d..n\}$, let $S \subseteq \{1..n\}$ be a size s volume sampled set over $\mathbf{X}$. Then*

$$\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+] = \mathbf{X}^+.$$

For the special case of $s = d$, this fact was known in the linear algebra literature [BTT90, BI92]. It was shown there using elementary properties of the determinant such as Cramer's rule.[3] The proof methodology developed here based on reverse iterative volume sampling is very different. We believe that this fundamental formula lies at the core of why volume sampling is important in many applications. In this work, we focus on its application to linear regression. However, [AB13] discuss many problems where

[3] Using the composition property of volume sampling (Corollary 2.1), the $s > d$ case of the theorem can be reduced to the $s = d$ case. However, we give a different self-contained proof.

controlling the pseudoinverse of a submatrix is essential. For those applications, it is important to establish variance bounds for the above expectation and volume sampling once again offers very concrete guarantees. We obtain them by showing the following formula, which can be viewed as a second moment for this estimator.

Theorem 2.4. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full rank matrix and $s \in \{d..n\}$. If size s volume sampling over $\mathbf{X}$ has full support, then*

$$\mathbb{E}[\ \underbrace{(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}}_{(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}}\ ] = \frac{n-d+1}{s-d+1}\ \underbrace{(\mathbf{X}^\top\mathbf{X})^{-1}}_{\mathbf{X}^+\mathbf{X}^{+\top}}.$$

In the case when volume sampling does not have full support, then the matrix equality "$=$" above is replaced by the positive-definite inequality "$\preceq$".

The condition that size s volume sampling over $\mathbf{X}$ has full support is equivalent to $\det(\mathbf{X}_S^\top\mathbf{X}_S) > 0$ for all $S \subseteq \{1..n\}$ of size s. Note that if size s volume sampling has full support, then size $t > s$ also has full support. So full support for the smallest size d (often phrased as $\mathbf{X}$ being *in general position*) implies that volume sampling w.r.t. any size $s \geq d$ has full support.

The above theorem immediately gives an expectation formula for the Frobenius norm $\|(\mathbf{I}_S\mathbf{X})^+\|_F$ of the estimator:

$$\mathbb{E}\left[\|(\mathbf{I}_S\mathbf{X})^+\|_F^2\right] = \mathbb{E}[\mathrm{tr}((\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top})] = \frac{n-d+1}{s-d+1}\|\mathbf{X}^+\|_F^2. \tag{2.1}$$

This norm formula was shown by [AB13], with numerous applications. Theorem 2.4 can be viewed as a much stronger pre-trace version of the known norm formula. Also our proof techniques are quite different and much simpler. Note that if size s volume sampling for $\mathbf{X}$ does not have full support, then (2.1) becomes an inequality.

We now mention a second application of the above theorem in the context of linear regression for the case when the response vector $\mathbf{y}$ is modeled as a noisy linear transformation, i.e. $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\xi}$ for some $\mathbf{w} \in \mathbb{R}^d$ and an i.i.d. mean zero noise vector $\boldsymbol{\xi} \in \mathbb{R}^n$. In this case the matrix $(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}$ can be interpreted as the covariance matrix of least-squares estimator $\mathbf{w}^*(S)$ (for a fixed set S) and Theorem 2.4 gives an exact formula for the covariance matrix of $\mathbf{w}^*(S)$ under volume sampling. In Chapter 3, we give an extended version of this result which provides even stronger guarantees for regularized least-squares estimators under this model (Theorems 3.1 and 3.2).

Note that except for the above application, all results in this chapter hold for arbitrary response vectors $\mathbf{y}$. By combining Theorems 2.3 and 2.4, we can obtain a covariance-type formula[4] for the pseudoinverse matrix estimator:

$$\begin{aligned}
&\mathbb{E}[((\mathbf{I}_S\mathbf{X})^+ - \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+])\,((\mathbf{I}_S\mathbf{X})^+ - \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+])^\top] \\
&\qquad = \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}] - \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]\,\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]^\top \\
&\qquad = \frac{n-d+1}{s-d+1}\,\mathbf{X}^+\mathbf{X}^{+\top} - \mathbf{X}^+\mathbf{X}^{+\top} \\
&\qquad = \frac{n-s}{s-d+1}\,\mathbf{X}^+\mathbf{X}^{+\top}. \qquad (2.2)
\end{aligned}$$

We now give the background for a third matrix expectation formula for volume sampling. Pseudoinverses can be used to compute the projection matrix onto the span of columns of matrix $\mathbf{X}$, which is defined as follows:

$$\mathbf{P}_\mathbf{X} \overset{def}{=} \mathbf{X}\overbrace{(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top}^{\mathbf{X}^+}.$$

[4]This notion of "covariance" is used in random matrix theory, i.e. for a random matrix $\mathbf{M}$ we analyze $\mathbb{E}[(\mathbf{M} - \mathbb{E}[\mathbf{M}])(\mathbf{M} - \mathbb{E}[\mathbf{M}])^\top]$. See for example [Tro12].

Applying Theorem 2.3 leads us immediately to the following unbiased matrix estimator for the projection matrix:

$$\mathbb{E}[\mathbf{X}(\mathbf{I}_S\mathbf{X})^+] = \mathbf{X}\,\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+] = \mathbf{X}\mathbf{X}^+ = \mathbf{P}_\mathbf{X}.$$

Note that this matrix estimator $\mathbf{X}(\mathbf{I}_S\mathbf{X})^+$ is closely connected to linear regression: It can be used to transform the response vector $\mathbf{y}$ into the prediction vector $\widehat{\mathbf{y}}(S)$ of subsampled least squares solution $\mathbf{w}^*(S)$ as follows:

$$\widehat{\mathbf{y}}(S) = \mathbf{X}\underbrace{(\mathbf{I}_S\mathbf{X})^+\mathbf{y}}_{\mathbf{w}^*(S)}.$$

In this case, volume sampling once again provides a covariance-type matrix expectation formula.

Theorem 2.5. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be a full rank matrix. If matrix $\mathbf{X}$ is in general position and $S \subseteq \{1..n\}$ is sampled according to size d volume sampling, then*

$$\mathbb{E}[\underbrace{(\mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2}_{(\mathbf{I}_S\mathbf{X})^{+\top}\mathbf{X}^\top\mathbf{X}(\mathbf{I}_S\mathbf{X})^+}] - \mathbf{P}_\mathbf{X} = d\,(\mathbf{I} - \mathbf{P}_\mathbf{X}).$$

If $\mathbf{X}$ is not in general position, then the matrix equality "$=$" is replaced by the positive-definite inequality "$\preceq$".

Note that this third expectation formula is limited to sample size $s = d$. It is a direct consequence of Theorem 2.6 given in the next section which relates the expected loss of a subsampled least squares estimator to the loss of the optimum least squares estimator. Unlike the first two formulas given in Theorems 2.3 and 2.4, its proof does not rely on the methodology of Lemma 2.1, i.e., on showing that the expectations at all

levels of a certain DAG associated with the sampling process are the same. We defer the proof of this third expectation formula to Section 2.4.1. No extension of this third formula to sample size $s > d$ is known.

Proof of Theorem 2.3

We apply Lemma 2.1 with $\mathbf{F}(S) = (\mathbf{I}_S\mathbf{X})^+$. It suffices to show that $\mathbf{F}(S) = \sum_{i\in S} P(S_{-i}|S)\mathbf{F}(S_{-i})$ for $P(S_{-i}|S) = \frac{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i}{s-d}$, i.e.:

$$(\mathbf{I}_S\mathbf{X})^+ = \sum_{i\in S} \frac{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i}{s-d} \underbrace{(\mathbf{I}_{S_{-i}}\mathbf{X})^+}_{(\mathbf{X}_{S_{-i}}^\top\mathbf{X}_{S_{-i}})^{-1}(\mathbf{I}_{S_{-i}}\mathbf{X})^\top}.$$

We first apply Sherman-Morrison to $(\mathbf{X}_{S_{-i}}^\top\mathbf{X}_{S_{-i}})^{-1} = (\mathbf{X}_S^\top\mathbf{X}_S - \mathbf{x}_i\mathbf{x}_i^\top)^{-1}$ on the r.h.s. of the above:

$$\sum_i \frac{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i}{s-d} \left((\mathbf{X}_S^\top\mathbf{X}_S)^{-1} + \frac{(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}}{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i}\right)((\mathbf{I}_S\mathbf{X})^\top - \mathbf{x}_i\mathbf{e}_i^\top).$$

Next we expand the last two factors into 4 terms. The expectation of the first term, i.e. $(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}(\mathbf{I}_S\mathbf{X})^\top$, is equal to $(\mathbf{I}_S\mathbf{X})^+$ (which is the l.h.s.) and the expectations of the remaining three terms times $s-d$ sum to 0:

$$-\sum_{i\in S}(1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i)\,(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i\mathbf{e}_i^\top + \cancel{(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\sum_{i\in S}\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}}(\mathbf{I}_S\mathbf{X})^\top$$
$$-\sum_{i\in S}(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i\,(\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S)^{-1}\mathbf{x}_i)\mathbf{e}_i^\top = 0.$$

In Section 2.5, we give an intriguing alternate proof of Theorem 2.3 based on a derivative formula for determinants from [PP12] (which can be obtained via matrix differential calculus [MN99]) and the Cauchy-Binet formula.

Proof of Theorem 2.4

Choose $\mathbf{F}(S) = \frac{s-d+1}{n-d+1}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}$. By Lemma 2.1 it suffices to show $\mathbf{F}(S) = \sum_{i\in S} P(S_{-i}|S)\mathbf{F}(S_{-i})$ for volume sampling:

$$\frac{s-d+1}{\cancel{n-d+1}}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1} = \sum_{i\in S} \frac{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i}{\cancel{s-d}} \frac{\cancel{s-d}}{\cancel{n-d+1}}(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}})^{-1}.$$

To show this we apply Sherman-Morrison to $(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}})^{-1}$ on the r.h.s.:

$$\sum_{i\in S}(1-\mathbf{x}_i^\top(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i)\left((\mathbf{X}_S^\top \mathbf{X}_S)^{-1} + \frac{(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}}{1-\mathbf{x}_i^\top(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i}\right)$$

$$= (s-d)(\mathbf{X}_S^\top \mathbf{X}_S)^{-1} + \cancel{(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\sum_{i\in S}\mathbf{x}_i\mathbf{x}_i^\top}(\mathbf{X}_S^\top \mathbf{X}_S)^{-1} = (s-d+1)\,(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}.$$

If some denominators $1-\mathbf{x}_i^\top(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{x}_i$ are zero, then we only sum over i for which the denominators are positive. In this case the above matrix equality becomes a positive-definite inequality $\preceq$.

2.3 Linear regression with smallest number of responses

Our main motivation for studying volume sampling came from asking the following simple question. Suppose we want to solve a d-dimensional linear regression problem with an input matrix $\mathbf{X}$ of n rows in $\mathbb{R}^d$ and a response vector $\mathbf{y} \in \mathbb{R}^n$, i.e. find $\mathbf{w} \in \mathbb{R}^d$ that minimizes the least squares loss $\|\mathbf{X}\mathbf{w}-\mathbf{y}\|^2$ on all n rows. We use $L(\mathbf{w})$ to denote this loss. The optimal weight vector minimizes $L(\mathbf{w})$, i.e.

$$\mathbf{w}^* \overset{def}{=} \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} L(\mathbf{w}) = \mathbf{X}^+\mathbf{y}.$$

Computing it requires access to the input matrix $\mathbf{X}$ and the response vector $\mathbf{y}$. Assume we are given $\mathbf{X}$ but the access to response vector $\mathbf{y}$ is restricted. We are allowed to pick

a random subset $S \subseteq \{1..n\}$ of fixed size s for which the responses $\mathbf{y}_S$ for the submatrix $\mathbf{X}_S$ are revealed to us, and then must produce a weight vector $\mathbf{w}(\mathbf{X}, S, \mathbf{y}_S) \in \mathbb{R}^d$ from a subset of row indices S of the input matrix $\mathbf{X}$ and the corresponding responses $\mathbf{y}_S$. Our goal in this section is to find a distribution on the subsets S of size s and a *weight function* $\mathbf{w}(\mathbf{X}, S, \mathbf{y}_S)$ s.t.[5]

$$\forall\,(\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n\times d} \times \mathbb{R}^{n\times 1}: \quad \mathbb{E}\left[L(\mathbf{w}(\mathbf{X}, S, \mathbf{y}_S))\right] \leq (1+c)\, L(\mathbf{w}^*),$$

where c must be a fixed constant (that is independent of $\mathbf{X}$ and $\mathbf{y}$). Throughout the chapter we use the one argument shorthand $\mathbf{w}(S)$ for the weight function $\mathbf{w}(\mathbf{X}, S, \mathbf{y}_S)$. We assume that attaining response values is expensive and ask the question: What is the smallest number of responses (i.e. smallest size of S) for which such a multiplicative bound is possible? We will use volume sampling to show that attaining d response values is sufficient and show that less than d responses is not.

Before we state our main upper bound based on volume sampling, we make the following key observation: If for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$ there is a weight vector $\mathbf{w}(S)$ that has loss zero, then the algorithm has to predict with such a consistent weight vector. This is because in that case the responses $\mathbf{y}_S$ can be extended to a response vector $\mathbf{y}$ for all of $\mathbf{X}$ s.t. $L(\mathbf{w}^*) = 0$. Thus since we aim for a multiplicative loss bound, we force the algorithm to predict with the optimum solution $\mathbf{w}^*(S) \overset{def}{=} \mathbf{X}_S^+ \mathbf{y}_S$ whenever the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$ has loss 0. In particular, when $|S| = d$ and $\mathbf{X}_S$ has full rank,

[5]Since the learner is given $\mathbf{X}$, it is natural to define the optimal multiplicative constant specialized for each $\mathbf{X}$: $c_{\mathbf{X},s} = \min_c \min_{P(\cdot),\mathbf{w}(\cdot)} \max_{\mathbf{y}} \mathbb{E}_P\left[L(\mathbf{w}(\mathbf{X}, S, \mathbf{y}_S))\right] \leq (1+c)\, L(\mathbf{w}^*)$, where the domain for distribution $P(\cdot)$ and weight function $\mathbf{w}(\cdot)$ are sets of size s. Showing specialized bounds for $c_{\mathbf{X},s}$ is left for future research.

then there is a unique consistent solution $\mathbf{w}^*(S)$ for the subproblem and the learner must use the weight function $\mathbf{w}(S) = \mathbf{w}^*(S)$.

Theorem 2.6. *If the input matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ is in general position, then for any response vector $\mathbf{y} \in \mathbb{R}^n$, the expected square loss (on all n rows of $\mathbf{X}$) of the optimal solution $\mathbf{w}^*(S)$ for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$, with the d-element set S obtained from volume sampling, is given by*

$$\mathbb{E}[L(\mathbf{w}^*(S))] = (d+1)\, L(\mathbf{w}^*).$$

If $\mathbf{X}$ is not in general position, then the expected loss is upper-bounded by $(d+1)\, L(\mathbf{w}^*)$.

This bound is already non-obvious for dimension 1, when the multiplicative factor is 2 (as discussed in Section 1.1). Note, that if there is a bias term in dimension 1, then the factor becomes 3. In dimension d, it is instructive to look at the case when the square loss of the optimum solution is zero, i.e. there is a weight vector $\mathbf{w}^* \in \mathbb{R}^d$ s.t. $\mathbf{X}\mathbf{w}^* = \mathbf{y}$. In this case the response values of any d linearly independent rows of $\mathbf{X}$ determine the optimum solution and the multiplicative loss formula of the theorem clearly holds. The formula specifies how noise-free case generalizes gracefully to the noisy case in that for volume sampling, the expected square loss of the solution obtained from d row response pairs is always by a factor of

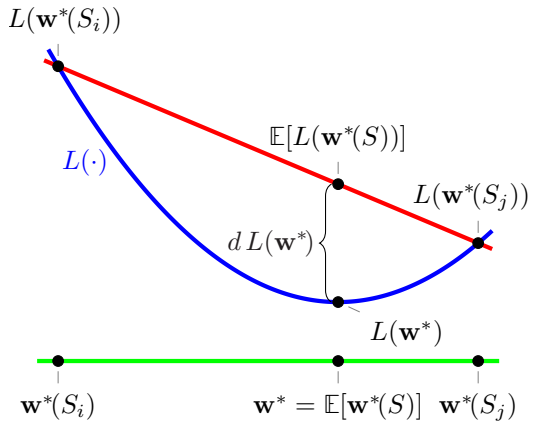

Figure 2.3: Unbiased estimator $\mathbf{w}^*(S)$ in expectation suffers loss $(d+1)\, L(\mathbf{w}^*)$.

at most $d+1$ larger than the square loss of the optimum solution. Moreover, since $\mathbb{E}[\mathbf{w}^*(S)] = \mathbf{w}^*$ and the loss function $L(\cdot)$ is convex, we have by Jensen's inequality that

$$\mathbb{E}\big[L(\mathbf{w}^*(S))\big] \geq L\big(\mathbb{E}[\mathbf{w}^*(S)]\big) = L(\mathbf{w}^*).$$

The above theorem now states that the gap $\mathbb{E}[L(\mathbf{w}^*(S))] - L(\mathbf{w}^*)$ in Jensen's inequality (which coincides with the "regret" of the estimator) equals $d\,L(\mathbf{w}^*)$, when the expectation is w.r.t. size d volume sampling and $\mathbf{X}$ is in general position (See Figure 2.3 for a schematic). As we will show in Section 2.6, this gap also equals the variance $\mathbb{E}[\|\mathbf{X}\mathbf{w}^*(S) - \mathbf{X}\mathbf{w}^*\|^2]$ of the predictions since the estimator is unbiased. In summary:

$$\underbrace{\mathbb{E}\big[L(\mathbf{w}^*(S))\big] - L(\overbrace{\mathbf{w}^*}^{\mathbb{E}[\mathbf{w}^*(S)]})}_{\text{regret}} = \underbrace{d\,L(\mathbf{w}^*)}_{\text{gap in Jensen's}} = \underbrace{\mathbb{E}\big[\|\mathbf{X}\mathbf{w}^*(S) - \mathbf{X}\mathbf{w}^*\|^2\big]}_{\text{variance}}.$$

For the remainder of this section we make a number of observations and present some lower bounds that highlight the above bound. Then, in Section 2.4 we prove the theorem and a matrix expectation formula that is implied by it.

2.3.1 When X is not in general position

The above theorem gives a surprising equality for the expected loss of a volume-sampled solution. However, this equality is only guaranteed to hold when matrix $\mathbf{X}$ is in general position. We give a minimal example problem where the matrix $\mathbf{X}$ is not in general position and the equality of Theorem 2.6 turns into a strict inequality. This shows that for the equality, the general position assumption is necessary. If we apply even an infinitesimal additive perturbation to the matrix $\mathbf{X}$ of the example problem, then the resulting matrix $\mathbf{X}_\epsilon$ is in general position and the equality holds. Note that even

though the optimum loss $L(\mathbf{w}^*)$ does not change significantly under such a perturbation, the expected sampling loss $\mathbb{E}[L(\mathbf{w}^*(S))]$ has to jump sufficiently to close the gap in the inequality. In our minimal example problem, $n = 3$ and $d = 2$, and

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

We have three 2-element subsets to sample from: $S_1 = \{1, 2\},\ S_2 = \{2, 3\},\ S_3 = \{1, 3\}$. Notice that the first two rows of $\mathbf{X}$ are identical, which means that the probability of sampling set S_1 is 0 in the volume sampling process. The other two subsets, S_2 and S_3, form identical submatrices $\mathbf{X}_{S_2} = \mathbf{X}_{S_3}$. Therefore they are equally probable. The optimal weight vectors for these sets are $\mathbf{w}^*(S_2) = (0, 0)^\top$ and $\mathbf{w}^*(S_3) = (0, 1)^\top$. Also $\mathbf{w}^* = (0, \frac{1}{2})^\top$ and the expected loss is bounded as:

$$\mathbb{E}[L(\mathbf{w}^*(S))] = \underbrace{\frac{1}{2}\overbrace{L(\mathbf{w}^*(S_2))}^{1} + \frac{1}{2}\overbrace{L(\mathbf{w}^*(S_3))}^{1}}_{1} \ < \ \underbrace{\overbrace{(d+1)}^{3}\overbrace{L(\mathbf{w}^*)}^{1/2}}_{3/2}.$$

Now consider a slightly perturbed input matrix

$$\mathbf{X}_\epsilon = \begin{pmatrix} 1 & 1+\epsilon \\ 1 & 1 \\ 1 & 0 \end{pmatrix},$$

where $\epsilon > 0$ is arbitrarily small (We keep the response vector $\mathbf{y}$ the same). Now, there is no $d \times d$ submatrix that is singular, so the upper bound from Theorem 2.6 must be tight. The reason is that even though subset S_1 still has very small probability, its loss is very large, so the expectation is significantly affected by this component, no matter

how small ϵ is. We see this directly in the calculations. Let $\mathbf{w}^*$ and $\mathbf{w}^*(S_i)$ be the corresponding solutions for the perturbed problem and its subproblems. The volumes of the subproblems and their losses are:

$$\begin{aligned}
&\det(\mathbf{X}_{S_1}^\top \mathbf{X}_{S_1}) = \epsilon^2 && L(\mathbf{w}^*(S_1)) = \epsilon^{-2} \\
&\det(\mathbf{X}_{S_2}^\top \mathbf{X}_{S_2}) = 1 && L(\mathbf{w}^*(S_2)) = 1 && L(\mathbf{w}^*) = \frac{1}{2(1+\epsilon+\epsilon^2)}. \\
&\det(\mathbf{X}_{S_3}^\top \mathbf{X}_{S_3}) = (1+\epsilon)^2 && L(\mathbf{w}^*(S_3)) = (1+\epsilon)^{-2}
\end{aligned}$$

Note that for each subproblem, the product of volume times loss is equal to 1. Now the expected loss can be easily computed, and we can see that the gap in the bound disappears (the denominator is the normalizing constant for volume sampling):

$$\mathbb{E}[L(\mathbf{w}^*(S))] = \frac{1+1+1}{\epsilon^2+1+(1+\epsilon)^2} = (d+1)\, L(\mathbf{w}^*).$$

2.3.2 Lower-bounds for selecting d responses

The factor $d+1$ in Theorem 2.6 cannot, in general, be improved when selecting only d responses:

Proposition 2.2. *For any d, there exists a least squares problem $(\mathbf{X}, \mathbf{y})$ with $d+1$ rows in $\mathbb{R}^d$ such that for every d-element index set $S \subseteq \{1\,..\,d+1\}$, we have*

$$L(\mathbf{w}^*(S)) = (d+1)\, L(\mathbf{w}^*).$$

Proof. Choose the input vectors $\mathbf{x}_i$ (and rows $\mathbf{x}_i^\top$) as the $d+1$ corners of any simplex in $\mathbb{R}^d$ centered at the origin and choose all $d+1$ responses as the same non-zero value α. For any α, the optimal solution $\mathbf{w}^*$ will be the all-zeros vector with loss

$$L(\mathbf{w}^*) = (d+1)\, \alpha^2.$$

On the other hand, taking any size d subset of indices $S \subseteq \{1\,..\,d+1\}$, the subproblem solution $\mathbf{w}^*(S)$ will only produce loss on the left out input vector $\mathbf{x}_i$, indexed with $i \notin S$. To obtain the prediction on x_i, we use a simple geometric argument. Observe that since the simplex is centered, we can write the origin of $\mathbb{R}^d$ in terms of the corners of the simplex as

$$\mathbf{0} = \sum_k \mathbf{x}_k = \mathbf{x}_i + d\,\bar{\mathbf{x}}_{-i}, \qquad \text{where } \bar{\mathbf{x}}_{-i} \stackrel{def}{=} \frac{1}{d}\sum_{k\neq i} \mathbf{x}_k.$$

Thus, the left out input vector $\mathbf{x}_i$ equals $-d\,\bar{\mathbf{x}}_{-i}$. The prediction of $\mathbf{w}^*(S)$ on this vector is

$$\widehat{y}_i = \mathbf{x}_i^\top \mathbf{w}^*(S) = -d\left(\frac{1}{d}\sum_{k\neq i} \mathbf{x}_k^\top\right)\mathbf{w}^*(S) = -\sum_{k\neq i} \mathbf{x}_k^\top \mathbf{w}^*(S) = -d\alpha.$$

It follows that the loss of $\mathbf{w}^*(S)$ equals

$$L(\mathbf{w}^*(S)) = (\widehat{y}_i - y_i)^2 = (-d\alpha - \alpha)^2 = (d+1)^2\alpha^2 = (d+1)\,L(\mathbf{w}^*). \qquad \blacksquare$$

Moreover, it is easy to show that no *deterministic* algorithm for selecting d rows (without knowing the responses) can guarantee a multiplicative loss bound with a factor less than n/d [BDM13]. For the sake of completeness, we show this here for $d = 1$:

Proposition 2.3. *For any $n \times 1$ input matrix $\mathbf{X}$ of all 1's and any deterministic algorithm that chooses some singleton set $S = \{i\}$, there is a response vector $\mathbf{y}$ for which the loss of the subproblem and the optimal loss are related as follows:*

$$L(\mathbf{w}^*(S)) = n\,L(\mathbf{w}^*).$$

Proof. If the response vector $\mathbf{y}$ is the vector of n 1's except for a single 0 at index i, then we have

$$\underbrace{L(\overbrace{\mathbf{w}^*(\{i\})}^{0})}_{n-1} = n\ \underbrace{L(\overbrace{\mathbf{w}^*}^{\frac{n-1}{n}})}_{\frac{n-1}{n}}.$$

■

Note that for the 1-dimensional example used in the proof, volume sampling would pick the set S uniformly. For this distribution, the multiplicative factor drops from n downto 2, that is $\mathbb{E}[L(\mathbf{w}^*(S))] = \frac{1}{n}(n-1) + \frac{n-1}{n}1 = 2\ L(\mathbf{w}^*)$.

2.3.3 The importance of joint sampling

Three properties of volume sampling play a crucial role in achieving a multiplicative loss bound:

1. *Randomness*: No deterministic algorithm guarantees such a bound (see Proposition 2.3).

2. *The chosen submatrices must have full rank*: Choosing any rank deficient submatrix with positive probability, does not allow for a multiplicative bound (see Propositions 2.4 and 2.5).

3. *Jointness*: No i.i.d. sampling procedure can achieve a multiplicative loss bound with $O(d)$ responses (see Corollary 2.2).

By jointly selecting subset S, volume sampling ensures that the corresponding input vectors $\mathbf{x}_i$ are well spread out in the input space $\mathbb{R}^d$. In particular, volume

sampling does not put any probability mass on sets S such that the rank of submatrix $\mathbf{X}_S$ is less than d. Intuitively, selecting rank deficient row subsets should not be effective, since such a choice leads to an under-determined least squares problem. We make this simple statement more precise by showing that any randomized algorithm, that with positive probability selects a rank deficient row subset, cannot achieve a multiplicative loss bound. Intuitively if the algorithm picks a rank deficient subset then it is not clear how it should select the weight vector $\mathbf{w}(S)$ given input matrix $\mathbf{X}$, subset S and responses $\mathbf{y}_S$. We reasoned before that $\mathbf{w}(S)$ must have loss 0 on the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. However if $\text{rank}(\mathbf{X}_S) < d$, then the choice of the weight vector $\mathbf{w}(S)$ with loss 0 is not unique and this causes positive loss for some response vector $\mathbf{y}$.

Proposition 2.4. *If for any input matrix $\mathbf{X}$, the algorithm samples a rank deficient subset S of rows with positive probability, then the expected loss of the algorithm cannot be bounded by a constant times the optimum loss for all response vectors $\mathbf{y}$.*

Note that this means in particular that if $\mathbf{X}$ has rank d, then sampling $d-1$ size subsets with positive probability does not allow for a constant factor approximation.

Proof. Let S be a rank deficient subset chosen with probability $P(S) > 0$. Since in our setup the bound has to hold for all response vectors $\mathbf{y}$ we can imagine an adversary choosing a worst-case $\mathbf{y}$. This adversary gives all rows of $\mathbf{X}_S$ the response value zero. Let $\mathbf{w}(S)$ be the plane produced by the algorithm when choosing S and receiving the responses 0 for $\mathbf{X}_S$. Let $i \in \{1..n\}$ s.t. $\mathbf{x}_i^\top \notin \text{row-span}(\mathbf{X}_S)$ and let $\mathbf{w}^*$ be any weight vector that gives response value 0 to all rows of $\mathbf{X}_S$ and response value $\mathbf{x}_i^\top \mathbf{w}(S) + Y$ to

$\mathbf{x}_i$. The adversary chooses $\mathbf{y}$ as $\mathbf{X}\mathbf{w}^*$, i.e. it gives all points $\mathbf{x}_j$ not indexed by S and different from $\mathbf{x}_i$ the response values $\mathbf{x}_j^\top \mathbf{w}^*$ as well. Now $\mathbf{w}^*$ has total loss 0 but $\mathbf{w}(S)$ has loss Y^2 on $\mathbf{x}_i$ and the algorithm's expected total loss is $\geq P(S)\, Y^2$. ■

We now strengthen the above proposition in that whenever the sample S is rank deficient then the loss of the optimum is zero while the loss of the algorithm is positive. However note that this proposition is weaker than the above in that it only holds for specific input matrices.

Proposition 2.5. *Let $d \leq n$ and let $\mathbf{X}$ be any input matrix of rank d consisting of n standard basis row vectors in $\mathbb{R}^d$. Then for any randomized learning algorithm that with probability p selects a subset S s.t. $\operatorname{rank}(\mathbf{X}_S) < d$ and any weight function $\mathbf{w}(\cdot)$, there is a response vector $\mathbf{y}$, satisfying:*

$$L(\mathbf{w}^*) = 0, \quad \textit{and} \quad L(\mathbf{w}(S)) > 0 \quad \textit{with probability at least } p.$$

Proof. Let $Q = \{1, 2, \ldots, 2^n\}$. The adversarial response vector $\mathbf{y}$ is constructed by carefully selecting one of the weight vectors $\mathbf{w}^* \in Q^d$, and setting the response vector $\mathbf{y}$ to $\mathbf{X}\mathbf{w}^*$. This ensures that $L(\mathbf{w}^*) = 0$ and since $\mathbf{X}$ consists of standard basis row vectors, the components of $\mathbf{y}$ lie in Q as well. Note that if the learner does not discover $\mathbf{w}^*$ exactly, it will incur positive loss. Let $\mathcal{H}$ be the set of all rank deficient sets in $\mathbf{X}$, i.e. those that lack at least one of the standard basis vectors:

$$\mathcal{H} = \{S \subseteq \{1..n\} \ : \ \operatorname{rank}(\mathbf{X}_S) < d\}.$$

Suppose that given matrix $\mathbf{X}$, the learner uses weight function $\mathbf{w}(S, \mathbf{y}_S)$. (Note that for the sake of concreteness we stopped using the single argument shorthand for the weight function during this proof.) We will count the number of possible inputs to this function, when S is a rank deficient index set of the rows of $\mathbf{X}$ and the response vector $\mathbf{y}_S$ is consistent with some $\mathbf{w}^* \in Q^d$. For any fixed rank deficient set S, let t be the number of distinct basis vectors appearing in $\mathbf{X}_S$. Clearly $t \leq d-1$. Fix a subset $T \subseteq S$ of size t s.t. $\mathbf{X}_T$ contains all t basis vectors of $\mathbf{X}_S$ exactly once (Thus the basis vectors in $\mathbf{X}_{S\setminus T}$ are all duplicates). Since $\mathbf{y} \in Q^n$, the components of $\mathbf{y}_S$ also lie in Q and $\mathbf{y}_S$ is determined by the responses of $\mathbf{y}_T$. Clearly there are at most $|Q|^{d-1}$ choices for $\mathbf{y}_T$. It follows that the number of possible input pairs $(S, \mathbf{y}_S)$ for function $\mathbf{w}(\cdot, \cdot)$ under the above restrictions can be bounded as

$$\left|\left\{(S, \mathbf{y}_S) \,:\, [S \in \mathcal{H}] \text{ and } [\mathbf{y}_S = \mathbf{X}_S\mathbf{w}^* \text{ for } \mathbf{w}^* \in Q^d]\right\}\right| \leq \underbrace{|\mathcal{H}|}_{<2^n} \underbrace{\max_{S\in\mathcal{H}} |\{\mathbf{X}_S\mathbf{w}^* \,:\, \mathbf{w}^* \in Q^d\}|}_{\leq |Q|^{d-1}}$$

$$< 2^n |Q|^{d-1} = |Q^d|.$$

So for every weight function $\mathbf{w}(\cdot, \cdot)$, there exists $\mathbf{w}^* \in Q^d$ that is missed by $\{\mathbf{w}(S, \mathbf{y}_S) : S \in \mathcal{H}\}$. Selecting $\mathbf{y} = \mathbf{X}\mathbf{w}^*$ for the adversarial response vector, we guarantee that the learner picks the wrong solution for every rank deficient set S and therefore receives positive loss with probability at least p. ■

Using Proposition 2.5, we show that any i.i.d. row sampling distribution (like for example leverage score sampling) requires $\Omega(d \log d)$ samples to get any multiplicative loss bound, either with high probability or in expectation.

Corollary 2.2. *Let $d \leq n$ and let $\mathbf{X}$ be any input matrix of rank d consisting of n standard basis row vectors in $\mathbb{R}^d$. Then for any randomized learning algorithm which selects a random multiset $S \subseteq \{1..n\}$ of size $|S| \leq (d-1)\ln(d)$ via i.i.d. sampling from any distribution and uses any weight function $\mathbf{w}(S)$, there is a response vector $\mathbf{y}$ satisfying:*

$$L(\mathbf{w}^*) = 0, \quad \text{and} \quad L(\mathbf{w}(S)) > 0 \quad \text{with probability at least } 1/2.$$

Proof. Any i.i.d. sample of size at most $(d-1)\ln(d)$ with probability at least $1/2$ does not contain all of the unique standard basis vectors (Coupon Collector Problem[6]). Thus, with probability at least $1/2$ submatrix $\mathbf{X}_S$ has rank less than d. Now, for any such algorithm we can use Proposition 2.5 to select a consistent response vector $\mathbf{y}$ such that with probability at least $1/2$ we have $L(\mathbf{w}(S)) > 0$. ■

Note that the corollary requires $\mathbf{X}$ to be of a restricted form that contains a lot of duplicate rows. It is open whether this corollary still holds when $\mathbf{X}$ is an arbitrary full rank matrix.

2.4 Loss expectation formula (proof of Theorem 2.6)

First, we discuss several key connections between linear regression and volume, which are used in the proof. Note that the loss $L(\mathbf{w}^*)$ suffered by the optimum weight vector can be written as $\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$, the squared Euclidean distance between prediction

[6]This was proven for uniform sampling in Theorem 1.24 of [AD11]. It can be shown that uniform sampling is the best case for Coupon Collector Problem [Hol01], so the bound holds for any i.i.d. sampling.

vector $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}^*$ and the response vector $\mathbf{y}$. Since $\widehat{\mathbf{y}}$ is minimizing the distance from $\mathbf{y}$ to the subspace of $\mathbb{R}^n$ spanning the feature vectors $\{\mathbf{f}_1, \ldots, \mathbf{f}_d\}$ (columns of $\mathbf{X}$), it has to be the *projection* of $\mathbf{y}$ onto that subspace (see Figure 2.4). We denote this projection as $\mathbf{P}_\mathbf{X}\,\mathbf{y}$, as defined in Section 2.2.3. Note that $\mathbf{P}_\mathbf{X}$ is a linear mapping from $\mathbb{R}^n$ onto the column span of the matrix $\mathbf{X}$ such that

$$\text{for } \mathbf{u} \in \text{span}(\mathbf{X}) \quad \mathbf{u} = \mathbf{P}_\mathbf{X}\,\mathbf{y} \quad \Leftrightarrow \quad \mathbf{P}_\mathbf{X}\,(\mathbf{u} - \mathbf{y}) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}^\top(\mathbf{u} - \mathbf{y}) = \mathbf{0}. \tag{2.3}$$

We next give a second geometric interpretation of the length $\|\widehat{\mathbf{y}} - \mathbf{y}\|^2$. Let $\mathcal{P}$ be the parallelepiped formed by the d column/feature vectors of the input matrix $\mathbf{X}$. Furthermore, consider the extended input matrix produced by adding the response vector $\mathbf{y}$ to $\mathbf{X}$ as an extra column:

$$\widetilde{\mathbf{X}} \stackrel{def}{=} (\mathbf{X}, \mathbf{y}) \in \mathbb{R}^{n \times (d+1)}. \tag{2.4}$$

Using the "base × height" formula we can relate the volume of $\mathcal{P}$ to the volume of $\widetilde{\mathcal{P}}$, the parallelepiped formed by the $d+1$ columns of $\widetilde{\mathbf{X}}$. Observe that $\widetilde{\mathcal{P}}$ has $\mathcal{P}$ as one of its faces, with the response vector $\mathbf{y}$ representing the edge that protrudes from that face. Hence the volume of $\widetilde{\mathcal{P}}$ is the product of the volume of $\mathcal{P}$ and the distance between $\mathbf{y}$ and span($\mathbf{X}$).

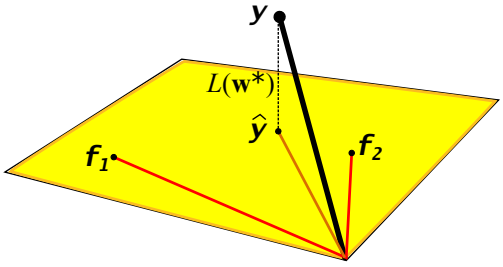

Figure 2.4: Prediction vector $\widehat{\mathbf{y}}$ is a projection of $\mathbf{y}$ onto the span of features $\mathbf{f}_i$.

This distance equals $\|\widehat{\mathbf{y}} - \mathbf{y}\|$, since as discussed above, $\widehat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto

span($\mathbf{X}$). Thus we have

$$\det(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}) = \det(\mathbf{X}^\top \mathbf{X})\, L(\mathbf{w}^*). \tag{2.5}$$

Next, we present a proposition whose corollary is key to proving Theorem 2.6. Suppose that we select one test row from the input matrix and use the remaining $n-1$ row response pairs as the training set. The proposition relates the loss of the obtained solution on the test row to the total leave-one-out loss an all rows.

Proposition 2.6. *For any index $i \in \{1..n\}$, let $\mathbf{w}^*(-i)$ be the solution to the reduced linear regression problem $(\mathbf{X}_{-i}, \mathbf{y}_{-i})$. Then*

$$L(\mathbf{w}^*(-i)) - L(\mathbf{w}^*) = \overbrace{\mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i}^{\frac{\det(\mathbf{X}^\top \mathbf{X}) - \det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})}{\det(\mathbf{X}^\top \mathbf{X})}} \ell_i(\mathbf{w}^*(-i)),$$

where $\ell_i(\mathbf{w}) \stackrel{def}{=} (\mathbf{x}_i^\top \mathbf{w} - y_i)^2$ is the square loss of $\mathbf{w}$ on the ith point.

An algebraic proof of this proposition essentially appears in the proof of Theorem 11.7 in [CBL06]. For the sake of completeness we give a new geometric proof of this proposition in Section 2.4.2 using basic properties of volume, thus stressing the connection to volume sampling.

Note that if matrix $\mathbf{X}$ has exactly $n = d+1$ rows and the training matrix $\mathbf{X}_{-i}$ is full rank, then $\mathbf{w}^*(-i)$ has loss zero on all training rows. In this case we obtain a simpler relationship than the proposition.

Corollary 2.3. *If $\mathbf{X}$ has $d+1$ rows and $\operatorname{rank}(\mathbf{X}_{-i}) = d$, then defining $\widetilde{\mathbf{X}}$ as in (2.4), we have*

$$\det(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}}) = \det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})\, \ell_i(\mathbf{w}^*(-i)).$$

Proof. By Proposition 2.6 and the fact that $L(\mathbf{w}^*(-i)) = \ell_i(\mathbf{w}^*(-i))$, we have

$$\det(\mathbf{X}^\top \mathbf{X})\; L(\mathbf{w}^*) = \det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})\; \ell_i(\mathbf{w}^*(-i)).$$

The corollary now follows from the "base $\times$ height" formula for volume. ■

Returning to the proof of Theorem 2.6, our goal is to find the expected loss $\mathbb{E}[L(\mathbf{w}^*(S))]$, where S is a size d volume sampled set. First, we rewrite the expectation as follows:

$$\begin{aligned}\mathbb{E}[L(\mathbf{w}^*(S))] &= \sum_{S,|S|=d} P(S) L(\mathbf{w}^*(S)) = \sum_{S,|S|=d} P(S) \sum_{j=1}^{n} \ell_j(\mathbf{w}^*(S)) \\ &= \sum_{S,|S|=d} \sum_{j \notin S} P(S)\; \ell_j(\mathbf{w}^*(S)) = \sum_{T,|T|=d+1} \sum_{j \in T} P(T_{-j})\; \ell_j(\mathbf{w}^*(T_{-j})). \qquad (2.6)\end{aligned}$$

We now use Corollary 2.3 on the matrix $\mathbf{X}_T$ and test row $\mathbf{x}_j^\top$ (assuming $\operatorname{rank}(\mathbf{X}_{T_{-j}}) = d$):

$$P(T_{-j})\; \ell_j(\mathbf{w}^*(T_{-j})) = \frac{\det(\mathbf{X}_{T_{-j}}^\top \mathbf{X}_{T_{-j}})}{\det(\mathbf{X}^\top \mathbf{X})}\; \ell_j(\mathbf{w}^*(T_{-j})) = \frac{\det(\widetilde{\mathbf{X}}_T^\top \widetilde{\mathbf{X}}_T)}{\det(\mathbf{X}^\top \mathbf{X})}. \qquad (2.7)$$

Since the summand does not depend on the index $j \in T$, the inner summation in (2.6) becomes a multiplication by $d+1$. This lets us write the expected loss as:

$$\mathbb{E}[L(\mathbf{w}^*(S))] = \frac{d+1}{\det(\mathbf{X}^\top \mathbf{X})} \sum_{T,|T|=d+1} \det(\widetilde{\mathbf{X}}_T^\top \widetilde{\mathbf{X}}_T) \overset{(1)}{=} (d+1) \frac{\det(\widetilde{\mathbf{X}}^\top \widetilde{\mathbf{X}})}{\det(\mathbf{X}^\top \mathbf{X})} \overset{(2)}{=} (d+1)\, L(\mathbf{w}^*), \qquad (2.8)$$

where (1) follows from the Cauchy-Binet formula and (2) is an application of the "base $\times$ height" formula. If $\mathbf{X}$ is not in general position, then for some summands in (2.7), $\operatorname{rank}(\mathbf{X}_{T_{-j}}) < d$ and $P(T_{-j}) = 0$. Thus the left-hand side of (2.7) is 0, while the right-hand side is non-negative, so (2.8) becomes an inequality, completing the proof of Theorem 2.6.

2.4.1 Lifting expectations to matrix form (proof of Theorem 2.5)

We show the matrix expectation formula of Theorem 2.5 as a corollary to the loss expectation formula of Theorem 2.6. The key observation is that the loss formula holds for arbitrary response vector $\mathbf{y}$, which allows us to "lift" it to the matrix form. Note, that the loss of least squares estimator can be written in terms of the projection matrix $\mathbf{P}_{\mathbf{X}}$:

$$L(\mathbf{w}^*) = \|\mathbf{y} - \widehat{\mathbf{y}}\|^2 = \|(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\mathbf{y}\|^2 = \mathbf{y}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{X}})^2\mathbf{y} \overset{(*)}{=} \mathbf{y}^\top(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\,\mathbf{y},$$

where in $(*)$ we used the following property of a projection matrix: $\mathbf{P}_{\mathbf{X}}^2 = \mathbf{P}_{\mathbf{X}}$. Writing the loss expectation of the subsampled estimator in the same form, we obtain:

$$\begin{aligned}\mathbb{E}[L(\mathbf{w}^*(S))] &= \mathbb{E}[\|\mathbf{y} - \widehat{\mathbf{y}}(S)\|^2] = \mathbb{E}[\|(\mathbf{I} - \mathbf{X}(\mathbf{I}_S\mathbf{X})^+)\mathbf{y}\|^2] \\ &= \mathbb{E}[\mathbf{y}^\top(\mathbf{I} - \mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2\,\mathbf{y}] = \mathbf{y}^\top\mathbb{E}[(\mathbf{I} - \mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2]\,\mathbf{y}.\end{aligned}$$

Crucially, we are able to extract the response vector $\mathbf{y}$ out of the expectation formula, which allows us to write the formula from Theorem 2.6 as follows:

$$\mathbf{y}^\top\ \mathbb{E}[(\mathbf{I} - \mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2]\ \mathbf{y} = \mathbf{y}^\top(d+1)(\mathbf{I} - \mathbf{P}_{\mathbf{X}})\ \mathbf{y}, \quad \forall\,\mathbf{y} \in \mathbb{R}^n.$$

We now use the following elementary fact: If for two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, we have $\mathbf{y}^\top\mathbf{A}\mathbf{y} = \mathbf{y}^\top\mathbf{B}\mathbf{y}$, $\forall\mathbf{y} \in \mathbb{R}^n$, then $\mathbf{A} = \mathbf{B}$. This gives the matrix expectation formula:

$$\mathbb{E}[(\mathbf{I} - \mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2] = (d+1)(\mathbf{I} - \mathbf{P}_{\mathbf{X}}).$$

Expanding square on the l.h.s. of the above and applying Theorem 2.3, we obtain the covariance-type equivalent form stated in Theorem 2.5:

$$\mathbf{I} - 2\overbrace{\mathbb{E}[\mathbf{X}(\mathbf{I}_S\mathbf{X})^+]}^{\mathbf{P_X}} + \mathbb{E}[(\mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2] = (d+1)(\mathbf{I} - \mathbf{P_X})$$

$$\iff \qquad \mathbb{E}[(\mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2] - \mathbf{P_X} = d\,(\mathbf{I} - \mathbf{P_X}).$$

2.4.2 Leave-one-out loss formula (proof of Proposition 2.6)

The main idea behind this proof is to construct variants of the input matrix $\mathbf{X}$ and relate their volumes. We use the following standard properties of the determinant:

Proposition 2.7. *For any matrix* $\mathbf{M}$, $\det(\mathbf{M}^\top\mathbf{M}) = \det(\widetilde{\mathbf{M}}^\top\widetilde{\mathbf{M}})$ *where* $\widetilde{\mathbf{M}}$ *is produced from* $\mathbf{M}$ *through the following operations:*

1. $\widetilde{\mathbf{M}}$ *equals* $\mathbf{M}$ *except that column* $\mathbf{m}_j$ *is replaced by* $\mathbf{m}_j + \alpha\mathbf{m}_i$*, where* $\mathbf{m}_i$ *is another column of* $\mathbf{M}$*;*

2. $\widetilde{\mathbf{M}}$ *equals* $\mathbf{M}$ *except that two rows are swapped.*

By Part 2 above, we can assume w.l.o.g. that $i = n$, i.e. that the test row in Proposition 2.6 is the last row of $\mathbf{X}$. Recall, that the columns of $\mathbf{X}$ are the feature vectors, denoted by $\mathbf{f}_1, \ldots, \mathbf{f}_d$. Moreover, the optimal prediction vector on the full dataset, $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}^*$, is a projection of $\mathbf{y}$ onto the subspace spanned by the features/columns of $\mathbf{X}$, denoted as $\widehat{\mathbf{y}} = \mathbf{P_X}\,\mathbf{y}$. Let us define a vector $\overline{\mathbf{y}}$ as

$$\overline{\mathbf{y}}^\top \stackrel{def}{=} (\text{——}\widehat{\mathbf{y}}_{-n}^\top\text{——}, y_n), \tag{2.9}$$

where $\widehat{\mathbf{y}}_{-n} \stackrel{def}{=} \mathbf{X}_{-n}\mathbf{w}^*(-n)$ is the optimal prediction vector for the training problem $(\mathbf{X}_{-n}, \mathbf{y}_{-n})$. Note, that if $\text{rank}(\mathbf{X}_{-n}) < d$, then $\mathbf{w}^*(-n)$ may not be unique, but we can pick any weight vector as long as it minimizes the loss on the training set $\{1..n-1\}$. Next, we show the following claim:

Claim 2.7. *The best achievable loss for the problem* $(\mathbf{X}, \mathbf{y})$ *can be decomposed as follows:*

$$L(\mathbf{w}^*) = L(\mathbf{w}^*(-n)) - \ell_n(\mathbf{w}^*(-n)) + \|\overline{\mathbf{y}} - \widehat{\mathbf{y}}\|^2. \tag{2.10}$$

Proof. First, we will show that $\overline{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the subspace spanned by all features and the unit vector $\mathbf{e}_n \in \mathbb{R}^n$ (where n corresponds to the test row). That is, we want to show that $\overline{\mathbf{y}} = \mathbf{P}_{(\mathbf{X},\mathbf{e}_n)}\,\mathbf{y}$. Denote $\widetilde{\mathbf{y}}$ as that projection. Observe that $\widetilde{y}_n = y_n$, because if this was not true, we could construct a vector $\widetilde{\mathbf{y}} + (y_n - \widetilde{y}_n)\mathbf{e}_n$ that is closer to $\mathbf{y}$ than $\widetilde{\mathbf{y}}$ and lies in $\text{span}(\mathbf{X}, \mathbf{e}_n)$. Thus, the projection does not incur any loss along the n-th dimension and can be reduced to the remaining $n-1$ dimensions, which corresponds to solving the training problem $(\mathbf{X}_{-n}, \mathbf{y}_{-n})$. Using the definition of $\overline{\mathbf{y}}$ in (2.9), this shows that $\widetilde{\mathbf{y}} = \mathbf{P}_{(\mathbf{X},\mathbf{e}_n)}\,\mathbf{y}$ equals $\overline{\mathbf{y}}$.

Next, we will show that $\widehat{\mathbf{y}}$ is the projection of $\overline{\mathbf{y}}$ onto $\text{span}(\mathbf{X})$, i.e. that $\mathbf{P}_{\mathbf{X}}\,\overline{\mathbf{y}} = \widehat{\mathbf{y}}$. By the linearity of projection, we have

$$\begin{aligned}
\mathbf{P}_{\mathbf{X}}\,\overline{\mathbf{y}} &= \mathbf{P}_{\mathbf{X}}(\overline{\mathbf{y}} - \mathbf{y} + \mathbf{y}) \\
&= \mathbf{P}_{\mathbf{X}}(\overline{\mathbf{y}} - \mathbf{y}) + \mathbf{P}_{\mathbf{X}}\,\mathbf{y} \\
&= \mathbf{P}_{\mathbf{X}}(\overline{\mathbf{y}} - \mathbf{y}) + \widehat{\mathbf{y}}.
\end{aligned}$$

We already showed that $\overline{\mathbf{y}} = \mathbf{P}_{(\mathbf{X},\mathbf{e}_n)}\,\mathbf{y}$. Therefore, the vector $\overline{\mathbf{y}} - \mathbf{y}$ is orthogonal to the column vectors of $\mathbf{X}$, and thus $\mathbf{P}_{\mathbf{X}}(\overline{\mathbf{y}} - \mathbf{y}) = 0$. This shows that $\mathbf{P}_{\mathbf{X}}\,\overline{\mathbf{y}} = \widehat{\mathbf{y}}$.

Finally, note that since $\overline{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto $\text{span}(\mathbf{X}, \mathbf{e}_n)$ and $\widehat{\mathbf{y}} \in \text{span}(\mathbf{X}, \mathbf{e}_n)$, vector $\overline{\mathbf{y}} - \mathbf{y}$ is orthogonal to vector $\overline{\mathbf{y}} - \widehat{\mathbf{y}}$ and by the Pythagorean Theorem we have

$$\|\widehat{\mathbf{y}} - \mathbf{y}\|^2 = \|\overline{\mathbf{y}} - \mathbf{y}\|^2 + \|\overline{\mathbf{y}} - \widehat{\mathbf{y}}\|^2.$$

Using the definition of $\overline{\mathbf{y}}$ in (2.9), we have

$$\|\overline{\mathbf{y}} - \mathbf{y}\|^2 = \|\widehat{\mathbf{y}}_{-n} - \mathbf{y}_{-n}\|^2 = L(\mathbf{w}^*(-n)) - \ell_n(\mathbf{w}^*(-n)),$$

concluding the proof of the claim. ■

Continuing with the proof of Proposition 2.6, we now construct a matrix $\overline{\mathbf{X}}$ by adding vector $\overline{\mathbf{y}}$ as an extra column to matrix $\mathbf{X}$:

$$\overline{\mathbf{X}} \stackrel{def}{=} (\mathbf{X}\,,\,\overline{\mathbf{y}}) = \left(\begin{array}{c|c} \mathbf{X}_{-n} & \widehat{\mathbf{y}}_{-n} \\ \hline \mathbf{x}_n^\top & y_n \end{array}\right). \tag{2.11}$$

Applying "base × height" and Claim 2.7, we compute the volume spanned by $\overline{\mathbf{X}}$:

$$\det(\overline{\mathbf{X}}^\top \overline{\mathbf{X}}) = \det(\mathbf{X}^\top \mathbf{X})\ \|\overline{\mathbf{y}} - \widehat{\mathbf{y}}\|^2 = \det(\mathbf{X}^\top \mathbf{X})\ (L(\mathbf{w}^*) - L(\mathbf{w}^*(-n)) + \ell_n(\mathbf{w}^*(-n))). \tag{2.12}$$

Next, we use the fact that volume is preserved under elementary column operations (Part 1 of Proposition 2.7). Note, that prediction vector $\widehat{\mathbf{y}}_{-n}$ is a linear combination of the columns of $\mathbf{X}_{-n}$, with the coefficients given by $\mathbf{w}^*(-n)$. Therefore, looking at the block structure of $\overline{\mathbf{X}}$ (see (2.11)), we observe that performing column operations on

the last column of $\overline{\mathbf{X}}$ with coefficients given by negative $\mathbf{w}^*(-n)$, we can zero out that column except for its last element:

$$\overline{\mathbf{y}} - \mathbf{X}\,\mathbf{w}^*(-n) = r\,\mathbf{e}_n,$$

where $r \stackrel{def}{=} y_n - \mathbf{x}_n^\top \mathbf{w}^*(-n)$ (see transformation (a) in (2.13)). Now, we consider two cases, depending on whether or not r equals zero. If $r \neq 0$, then we further transform the matrix by a second transformation (b), which zeros out the last row (the test row) using column operations. The entire sequence of operations, resulting in a matrix we call $\overline{\mathbf{X}}_0$, is shown below:

$$\overline{\mathbf{X}} = \left(\begin{array}{c|c} \mathbf{X}_{-n} & \widehat{\mathbf{y}}_{-n} \\ \hline \mathbf{x}_n^\top & y_n \end{array}\right) \stackrel{\text{(a)}}{\to} \left(\begin{array}{c|c} \mathbf{X}_{-n} & 0 \\ \hline \mathbf{x}_n^\top & r \end{array}\right) \stackrel{\text{(b)}}{\to} \left(\begin{array}{c|c} \mathbf{X}_{-n} & 0 \\ \hline 0 & r \end{array}\right) = \overline{\mathbf{X}}_0 \qquad (2.13)$$

Note, that due to the block-diagonal structure of $\overline{\mathbf{X}}_0$, its volume can be easily described by the "base $\times$ height" formula:

$$\det(\overline{\mathbf{X}}_0^\top \overline{\mathbf{X}}_0) = \det(\mathbf{X}_{-n}^\top \mathbf{X}_{-n})\; r^2 = \det(\mathbf{X}_{-n}^\top \mathbf{X}_{-n})\; \ell_n(\mathbf{w}^*(-n)). \qquad (2.14)$$

Since $\det(\overline{\mathbf{X}}^\top \overline{\mathbf{X}}) = \det(\overline{\mathbf{X}}_0^\top \overline{\mathbf{X}}_0)$, we can combine (2.12) and (2.14) to obtain the desired result.

Finally, if $r = 0$ we cannot perform transformation (b). However, in this case matrix $\overline{\mathbf{X}}$ has volume 0, and moreover, $\ell_n(\mathbf{w}^*(-n)) = r^2 = 0$, so once again we have

$$\det(\overline{\mathbf{X}}^\top \overline{\mathbf{X}}) = 0 = \det(\mathbf{X}_{-n}^\top \mathbf{X}_{-n})\; \ell_n(\mathbf{w}^*(-n)),$$

which concludes the proof of Proposition 2.6.

2.5 Matrix differentials as a tool for volume sampling

In this section, we briefly discuss matrix differential calculus and its intriguing application to volume sampling. A thorough treatment of matrix differential calculus can be found in [MN99]. Given a function $f : \mathbb{R}^{n\times d} \to \mathbb{R}$ and a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ we are interested in finding a matrix $\mathbf{A} \in \mathbb{R}^{n\times d}$, such that $\mathbf{A}_{ij} = \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}_{ij}}$. Following standard rules of differential calculus, we can compute this derivative by computing the "differential" of function $f(\mathbf{X})$, denoted $\mathrm{d}f(\mathbf{X})$, which is the linear approximation of f, i.e.

$$f(\mathbf{X} + \mathrm{d}\,\mathbf{X}) = f(\mathbf{X}) + \mathrm{d}f(\mathbf{X}) + (\text{higher order terms}).$$

For the formal definition we refer to [MN99]. The matrix derivative of f can be found by transforming the differential into a trace form:

$$\mathrm{d}f(\mathbf{X}) = \mathrm{tr}(\mathbf{A}^\top \mathrm{d}\mathbf{X}) \quad \Longleftrightarrow \quad \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} = \mathbf{A}. \tag{2.15}$$

In our analysis, we will need the following standard transformations allowed in computing a matrix differential (here, $\mathbf{X}$ and $\mathbf{Y}$ are used loosely as matrix functions, while $\mathbf{B}$ is a matrix constant):

(a) $$\mathrm{d}\,\mathbf{X}^\top = (\mathrm{d}\mathbf{X})^\top,$$

(b) $$\mathrm{d}\,\mathbf{B}\mathbf{X} = \mathbf{B}\,\mathrm{d}\mathbf{X},$$

(c) $$\mathrm{d}\,\mathbf{X}\mathbf{Y} = (\mathrm{d}\mathbf{X})\,\mathbf{Y} + \mathbf{X}\,(\mathrm{d}\mathbf{Y}),$$

(d) $$\mathrm{d}\det(\mathbf{X}) = \det(\mathbf{X})\,\mathrm{tr}(\,\mathbf{X}^{-1}\,\mathrm{d}\mathbf{X}\,).$$

As a key example for us, we derive the differential of $\det(\mathbf{X}_S^\top \mathbf{X}_S)$ for any fixed set S using the above rules and basic properties of the trace:

$$\begin{aligned}
\mathrm{d}\,\det(\mathbf{X}_S^\top \mathbf{X}_S) &\overset{\text{(d)}}{=} \det(\mathbf{X}_S^\top \mathbf{X}_S)\,\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\,\mathrm{d}(\overbrace{\mathbf{X}_S^\top \mathbf{X}_S}^{\mathbf{X}^\top(\mathbf{I}_S\mathbf{X})})\big) \\
&\overset{\text{(c)}}{=} \det(\mathbf{X}_S^\top \mathbf{X}_S)\,\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}((\mathrm{d}\,\mathbf{X}^\top)\mathbf{I}_S\mathbf{X} + \mathbf{X}^\top(\mathrm{d}\,\mathbf{I}_S\mathbf{X}))\big) \\
&\overset{\text{(b)}}{=} \det(\mathbf{X}_S^\top \mathbf{X}_S)\,\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}(\mathbf{X}^\top\mathbf{I}_S(\mathrm{d}\,\mathbf{X}^\top)^\top + \mathbf{X}^\top\mathbf{I}_S(\mathrm{d}\mathbf{X}))\big) \\
&\overset{\text{(a)}}{=} 2\det(\mathbf{X}_S^\top \mathbf{X}_S)\,\mathrm{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\mathbf{X}^\top\mathbf{I}_S\,\mathrm{d}\mathbf{X}\big) = 2\det(\mathbf{X}_S^\top \mathbf{X}_S)\,\mathrm{tr}\big((\mathbf{I}_S\mathbf{X})^+\,\mathrm{d}\mathbf{X}\big).
\end{aligned}$$

Thus, by the rule given in (2.15), we showed that $\frac{\partial \det(\mathbf{X}_S^\top \mathbf{X}_S)}{\partial \mathbf{X}} = 2\det(\mathbf{X}_S^\top \mathbf{X}_S)(\mathbf{I}_S\mathbf{X})^{+\top}$. This fact can be used to prove the pseudoinverse expectation formula of Theorem 2.3. The proof begins with generalized Cauchy-Binet for size s volume sampling:

$$\sum_{S,|S|=s} \det(\mathbf{X}_S^\top \mathbf{X}_S) = \binom{n-d}{s-d} \det(\mathbf{X}^\top \mathbf{X}).$$

Now, we take a derivative w.r.t. $\mathbf{X}$ on both sides

$$\begin{aligned}
&\sum_{S,|S|=s} 2\det(\mathbf{X}_S^\top \mathbf{X}_S)\ \ (\mathbf{I}_S\mathbf{X})^{+\top} = \binom{n-d}{s-d}\ \ 2\det(\mathbf{X}^\top \mathbf{X})\ \ \mathbf{X}^{+\top} \\
\iff \quad &\underbrace{\sum_{S,|S|=s} \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{s-d}\det(\mathbf{X}^\top \mathbf{X})}\ \ (\mathbf{I}_S\mathbf{X})^{+\top}}_{\mathbb{E}[(\mathbf{I}_S\mathbf{X})^{+\top}]} = \mathbf{X}^{+\top}.
\end{aligned}$$

2.6 Conclusion of the chapter and conjectures

We analyze linear regression when the learner is given the entire input matrix $\mathbf{X}$ which contains the points in $\mathbb{R}^d$ as rows. The response vector $\mathbf{y}$ contains one real response per row and is hidden from the learner. However the learner can request the

responses $\mathbf{y}_S$ for a small index set S of the points. The learner then produces a weight vector $\mathbf{w}(S)$ from the input matrix $\mathbf{X}$ and the requested responses $\mathbf{y}_S$. Our goal is to find a way to sample S and construct a weight function $\mathbf{w}(S)$ s.t. $\mathbb{E}[L(\mathbf{w}(S))] \leq (1+c)\ L(\mathbf{w}^*)$, where the multiplicative factor $1+c$ is bounded for all input matrices $\mathbf{X}$ and response vectors $\mathbf{y}$. Recall that $L(\cdot)$ denotes the square loss on all rows and $\mathbf{w}^*$ is the optimal solution based on all responses.

We show in this chapter that the smallest size of S for which this goal can be achieved is d (There is no sampling procedure for sets of size less than d and weight function $\mathbf{w}(S)$ for which this factor is finite). We also prove that when sets S of size d are drawn proportional to the squared volume of $\mathbf{X}_S$ (i.e. $\det(\mathbf{X}_S^\top \mathbf{X}_S)$), then $\mathbb{E}[L(\mathbf{w}^*(S))] \leq (d+1)L(\mathbf{w}^*)$, where the factor $d+1$ is optimal for some $\mathbf{X}$ and $\mathbf{y}$. Here $\mathbf{w}^*(S)$ denotes the linear least squares solution for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. The lower-bounds show that similar results are not possible for certain classes of algorithms: For any deterministic algorithm selecting a set S of size d the multiplicative factor can be at least n (the number of rows of the input matrix $\mathbf{X}$); also, any i.i.d. sampling procedure (such as leverage score sampling) requires $\Omega(d \log d)$ responses to achieve a finite factor.

We study volume sampling in more detail and develop a method for proving matrix expectation formulas. This method leads to exact formulas for $\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+]$ and $\mathbb{E}[(\mathbf{I}_S\mathbf{X})^+(\mathbf{I}_S\mathbf{X})^{+\top}]$ when the subset S of s row indices is sampled by volume sampling. The formulas hold for any fixed size $s \in \{d..n\}$. These new expectation formulas imply that the solution $\mathbf{w}^*(S)$ for a volume sampled subproblem of a linear regression problem

is unbiased:

$$\mathbb{E}[\mathbf{w}^*(S)] = \mathbb{E}[(\mathbf{I}_S\mathbf{X})^+\mathbf{y}] = \mathbf{X}^+\mathbf{y} = \mathbf{w}^*.$$

We also give an additional second order formula for $\mathbb{E}[(\mathbf{X}(\mathbf{I}_S\mathbf{X})^+)^2]$. However, this formula relies on our inequality $\mathbb{E}[L(\mathbf{w}^*(S))] \leq (d+1)L(\mathbf{w}^*)$ that only holds for volume sampling of size $s = d$. Generalizing this formula to sample size s larger than d is a challenging open problem.

A natural more general goal is to get arbitrarily close to the optimum loss. That is, for any ϵ, what is the smallest sample size $|S| = s$ for which there is a sampling distribution over subsets S and a weight function $\mathbf{w}(S)$ built from $\mathbf{X}$ and $\mathbf{y}_S$, such that $\mathbb{E}[L(\mathbf{w}(S))] \leq (1+\epsilon)\, L(\mathbf{w}^*)$. A related bound for i.i.d. leverage score sampling states that a sample size of $O(d \log d + \frac{d}{\epsilon})$ suffices to achieve a $1+\epsilon$ factor *with high probability* (this fact follows from standard techniques [Woo14] presented here in Section 4.4.1), however this does not imply multiplicative bounds *in expectation*.

We conjecture that some form of volume sampling can be used to achieve the $1+\epsilon$ factor with sample size $O(\frac{d}{\epsilon})$, in expectation. How close can we get with the techniques presented in this chapter? We showed that size d volume sampling achieves a factor of $1+d$, but we do not know how to generalize this proof to sample size larger than d. However, one unique property of the volume-sampled estimator $\mathbf{w}^*(S)$ that can be useful here is that it is an *unbiased estimator* of $\mathbf{w}^*$. As we shall see now, this basic property has many benefits. For any unbiased estimator (i.e. $\mathbb{E}[\mathbf{w}(S)] = \mathbf{w}^*$) and optimal prediction vector $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{w}^*$, consider the following rudimentary version of a

bias-variance decomposition:

$$\mathbb{E}\,\underbrace{\|\mathbf{X}\,\mathbf{w}(S)-\mathbf{y}\|^2}_{L(\mathbf{w}(S))} = \mathbb{E}\,\|\mathbf{X}\,\mathbf{w}(S)-\widehat{\mathbf{y}}+\widehat{\mathbf{y}}-\mathbf{y}\|^2 = \mathbb{E}\,\|\mathbf{X}\,\mathbf{w}(S)-\widehat{\mathbf{y}}\|^2 + \underbrace{\|\widehat{\mathbf{y}}-\mathbf{y}\|^2}_{L(\mathbf{w}^*)}.$$

The unbiasedness of the estimator assures that the cross term $(\mathbf{X}\overbrace{\mathbb{E}[\mathbf{w}(S)]}^{\mathbf{w}^*}-\widehat{\mathbf{y}})^\top(\widehat{\mathbf{y}}-\mathbf{y})$ is 0. Therefore a $1+c$ factor loss bound is equivalent to a c factor variance bound, i.e.

$$\overbrace{\mathbb{E}[L(\mathbf{w}(S))] \le (1+c)\,L(\mathbf{w}^*)}^{\text{loss bound}} \quad\Longleftrightarrow\quad \overbrace{\mathbb{E}\,\|\mathbf{X}\,\mathbf{w}(S)-\widehat{\mathbf{y}}\|^2 \le c\,L(\mathbf{w}^*)}^{\text{variance bound}}. \tag{2.16}$$

To reduce the variance of any unbiased estimator $\mathbf{w}(S)$ (i.e. $\mathbb{E}[\mathbf{w}(S)]=\mathbf{w}^*$) with sample size s, we can draw k independent samples $S_1,\ldots,S_k$ of size s each and predict with the average estimator $\frac{1}{k}\sum_{j=1}^k \mathbf{w}(S_j)$. If the loss bound from (2.16) holds for $\mathbf{w}(S)$, then the average estimator satisfies

$$\mathbb{E}\left[L\left(\frac{1}{k}\sum\nolimits_j \mathbf{w}(S_j)\right)\right] \le \left(1+\frac{c}{k}\right)\,L(\mathbf{w}^*).$$

Setting $k=c/\epsilon$, we need $s\,c/\epsilon$ responses to get a $1+\epsilon$ approximation. We showed that size d volume sampling achieves $c=d$. Thus with our current proof techniques, we need d^2/ϵ examples to get a $1+\epsilon$ factor approximation.

In conclusion the basic open problem is the following: Is there a size $O(d/\epsilon)$ *unbiased estimator* that achieves a $1+\epsilon$ factor approximation? By the above averaging method this is equivalent to the following question: Is there a size $O(d)$ unbiased estimator that achieves a constant factor? This is because once we have an unbiased estimator that achieves a constant factor, then by averaging $1/\epsilon$ copies, we get the $1+O(\epsilon)$ factor. Ideally the special unbiased estimators resulting from volume sampling can achieve this

feat. We conclude with our favorite open problem: Does size $O(d)$ volume sampling achieve a constant factor approximation? The following two chapters make progress towards addressing this open problem: in Chapter 3 we show that when the response vector is linear plus noise of mean zero, then the desired sample size is achievable (although the notion of loss in that setting is replaced with *mean squared prediction error*); in Chapter 4 we return to the worst-case response vector setting, and show a surprising lower bound indicating that volume sampling, as defined in this chapter, does not offer $1+\epsilon$ approximations for $\epsilon < \frac{1}{2}$ with small sample sizes. We then propose leveraged volume sampling, which produces an unbiased estimator satisfying the $1+\epsilon$ loss bound with high probability (but not in expectation) for sample size $O(d \log d + \frac{d}{\epsilon})$, matching that of i.i.d. leverage score sampling.

Chapter 3

Regularized volume sampling

3.1 Introduction

A simple approach for implementing volume sampling introduced in the previous chapter is to start with the full set of column indices $S = \{1..n\}$ and then (in reverse order) select an index i in each iteration to be eliminated from set S with probability proportional to the change in matrix volume caused by removing the ith column:

$$\text{Sample } i \sim P(i \,|\, S) = \frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}})}{(|S| - d)\det(\mathbf{X}_S^\top \mathbf{X}_S)}, \tag{3.1}$$

$$\text{Update } S \leftarrow S - i. \qquad \textbf{(reverse iterative volume sampling)}$$

As shown in Theorem 2.1, this procedure samples a set S of fixed size according to the distribution

$$P(S) \propto \det(\mathbf{X}_S^\top \mathbf{X}_S). \tag{3.2}$$

Note that when $|S| < d$, then all matrices $\mathbf{X}_S^\top \mathbf{X}_S$ are singular, and so the distribution becomes undefined. Motivated by this limitation, we propose a regularized variant,

called λ-regularized volume sampling:

$$\text{Sample } i \sim P(i \mid S) \propto \frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})}{\det(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})}, \tag{3.3}$$

$$\text{Update } S \leftarrow S - i. \qquad (\lambda\textbf{-regularized volume sampling})$$

The normalization factor of this conditional probability (i.e. the sum of (3.3) over $i \in S$) can be computed using Sylvester's theorem:

$$\begin{aligned}\sum_{i \in S} \frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})}{\det(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})} &= \sum_{i \in S} \big(1 - \mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{x}_i\big) \\ &= |S| - \operatorname{tr}\big(\mathbf{X}_S (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{X}_S^\top\big) \\ &= |S| - d + \lambda \operatorname{tr}\big((\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}\big). \end{aligned} \tag{3.4}$$

Note that in the special case of no regularization (i.e. $\lambda = 0$) the last trace vanishes and (3.4) is equal to $|S| - d$, so we recover standard volume sampling. However, when $\lambda > 0$, then the last term is non-zero and depends on the entire matrix $\mathbf{X}_S$. This makes regularized volume sampling more complicated and certain equalities proven in the previous chapter for $\lambda = 0$ no longer hold. In particular, the analogous closed form of the sampling probability $P(S)$ given in (3.2) is not recovered because the paths from node $\{1..n\}$ to node S in the directed acyclic graph described in Chapter 2 (see Figure 2.2) do not all have the same probability.

Nevertheless, we are able to show that the proposed λ-regularized distribution exhibits a fundamental connection to ridge regression. Given a matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, we consider the task of fitting a linear model to a vector of responses $\mathbf{y} = \mathbf{X}\widetilde{\mathbf{w}} + \boldsymbol{\xi}$, where $\widetilde{\mathbf{w}} \in \mathbb{R}^d$ and the noise $\boldsymbol{\xi} \in \mathbb{R}^n$ is a mean zero random vector with covariance matrix

$\mathrm{Var}[\boldsymbol{\xi}] \preceq \sigma^2\mathbf{I}$ for some $\sigma > 0$. A classical solution to this task is the ridge estimator:

$$\mathbf{w}^*_\lambda = \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda\|\mathbf{w}\|^2 = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}.$$

In this setting, we are interested in producing an estimator based on a small subset of responses indexed by set S. We will show that if S is sampled according to λ-regularized volume sampling with $\lambda \leq \frac{\sigma^2}{\|\widetilde{\mathbf{w}}\|^2}$ then the ridge estimator for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$

$$\mathbf{w}^*_\lambda(S) = (\mathbf{X}_S^\top\mathbf{X}_S + \lambda\mathbf{I})^{-1}\mathbf{X}_S^\top\mathbf{y}_S$$

has strong generalization properties with respect to the full problem $(\mathbf{X}, \mathbf{y})$. In particular, we prove that if the subset S has size s, then the mean squared prediction error (MSPE) of estimator $\mathbf{w}^*_\lambda(S)$ over the entire dataset $\mathbf{X}$ is bounded as follows:

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2\Big] \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1},$$
$$\text{where} \quad d_\lambda = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top)$$

is the statistical dimension. If λ_i are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$, then $d_\lambda = \sum_{i=1}^d \frac{\lambda_i}{\lambda_i+\lambda}$. Note that d_λ is decreasing with λ and $d_0 = d$. If the spectrum of the matrix $\mathbf{X}^\top\mathbf{X}$ decreases quickly then d_λ does so as well with increasing λ. When λ is properly tuned then d_λ is the effective degrees of freedom of $\mathbf{X}$. Our new lower bounds show that the above upper bound for regularized volume sampling is essentially optimal with respect to the choice of a subsampling procedure.

Recall that volume sampling can be viewed as a non-i.i.d. extension of leverage score sampling [DMIMW12], a widely used method where columns are sampled

independently according to their leverage scores. We show that any i.i.d. subsampling with respect to any fixed distribution such as leverage score sampling can require $\Omega(d_\lambda \ln(d_\lambda))$ labels to achieve any generalization for ridge regression, compared to $O(d_\lambda)$ for regularized volume sampling. We reinforce this claim experimentally in Section 3.5.

The main obstacle against using volume sampling in practice has been high computational cost. In particular, the only previously known polynomial time algorithm for exact volume sampling was $O(n^4 s)$ [LJS17], whereas exact leverage score sampling[1] is $O(nd^2)$. For many modern datasets, the number of examples n is much larger than d, which makes existing algorithms for volume sampling infeasible. In this chapter, we give an easy-to-implement volume sampling algorithm that runs in time $O(nd^2)$. Thus we give the first volume sampling procedure which is essentially linear in n and matches the time complexity of exact leverage score sampling. Finally our procedure also achieves regularized volume sampling for any $\lambda > 0$ with the same running time.

Outline of the chapter. In the following section we show a matrix expectation inequality for λ-regularized volume sampling, which we then use for the statistical analysis of volume sampled ridge estimators in Section 3.3. Next, in Section 3.4 we present two efficient algorithms for regularized volume sampling. Finally, we evaluate the runtime of our algorithms on several standard linear regression datasets, and compare the prediction performance of the subsampled ridge estimator under volume sampling versus leverage score sampling (Section 3.5). We conclude with a brief summary of this chapter

[1] Approximate leverage score sampling methods achieve even better runtime of $\widetilde{O}(nd + d^3)$.

in Section 3.6.

3.2 A matrix expectation inequality

In the previous chapter we showed an important matrix expectation formula for standard volume sampling (Theorem 2.4) which states that if matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ is in general position and set S is sampled according to size $s \geq d$ volume sampling, then

$$\mathbb{E}\big[(\mathbf{X}_S^\top \mathbf{X}_S)^{-1}\big] = \frac{n-d+1}{s-d+1}(\mathbf{X}^\top\mathbf{X})^{-1}.$$

If $\mathbf{X}$ is not in general position, the above equality "=" is replaced with a positive semi-definite inequality "$\preceq$". We showed this using a proof technique based on reverse iterative sampling, which can also be applied to regularized volume sampling, resulting in the following extension of the above formula:

Theorem 3.1. *For any $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\lambda \geq 0$, let S be sampled according to λ-regularized size s volume sampling from $\mathbf{X}$. Then,*

$$\mathbb{E}\big[(\mathbf{X}_S^\top \mathbf{X}_S + \lambda\mathbf{I})^{-1}\big] \preceq \frac{n-d_\lambda+1}{s-d_\lambda+1}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}$$

for any $s \geq d_\lambda = \mathrm{tr}(\mathbf{X}(\mathbf{X}^\top\mathbf{X}+\lambda\mathbf{I})^{-1}\mathbf{X}^\top)$.

Remark. *In many settings, $d_\lambda \ll d$, thus unlike Theorem 2.4, the above result offers meaningful bounds for sampling sets S of size smaller than d. Also, note that the above inequality does not turn into an equality when $\mathbf{X}$ is in general position.*

Proof. To obtain Theorem 3.1, we use essentially the same methodology as described in Lemma 2.1, except in the regularized case equality is replaced with inequality. Recall

that using Sylvester's theorem we can compute the unnormalized conditional probability from (3.3):

$$h_i = \frac{\det(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})}{\det(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})} = 1 - \mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{x}_i.$$

From now on, we will use $\mathbf{Z}_\lambda(S) = \mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I}$ as a shorthand in the proofs. Next, letting $M = \sum_{i \in S} h_i$, we compute unnormalized expectation by applying the Sherman-Morrison formula:

$$\begin{aligned}
M\, \mathbb{E}\big[(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})^{-1} \,|\, S\big] &= \sum_{i \in S} h_i \Bigg(\overbrace{\mathbf{Z}_\lambda(S)^{-1} + \frac{\mathbf{Z}_\lambda(S)^{-1} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}}{1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1} \mathbf{x}_i}}^{\mathbf{Z}_\lambda(S_{-i})^{-1}} \Bigg) \\
&= M\, \mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1} \Big(\sum_{i \in S} \mathbf{x}_i \mathbf{x}_i^\top \Big) \mathbf{Z}_\lambda(S)^{-1} \\
&= M\, \mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1} (\mathbf{Z}_\lambda(S) - \lambda \mathbf{I}) \mathbf{Z}_\lambda(S)^{-1} \\
&= M\, \mathbf{Z}_\lambda(S)^{-1} + \mathbf{Z}_\lambda(S)^{-1} - \lambda \mathbf{Z}_\lambda(S)^{-2} \;\preceq\; (M+1)\, \mathbf{Z}_\lambda(S)^{-1}.
\end{aligned}$$

Finally, the normalization factor M (which we already computed in (3.4)) can be lower-bounded using the λ-statistical dimension d_λ of matrix $\mathbf{X}$:

$$M = \sum_{i \in S} (1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1} \mathbf{x}_i) = s - d + \lambda \operatorname{tr}(\mathbf{Z}_\lambda(S)^{-1}) \geq s - \Big(\underbrace{d - \lambda \operatorname{tr}(\mathbf{Z}_\lambda(\{1..n\})^{-1})}_{d_\lambda} \Big).$$

Putting the bounds together, we obtain that:

$$\mathbb{E}\big[(\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})^{-1} \,|\, S\big] \preceq \frac{s - d_\lambda + 1}{s - d_\lambda} (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}.$$

To prove Theorem 3.1 it remains to chain the conditional expectations along the sequence of subsets obtained by λ-regularized volume sampling:

$$\mathbb{E}\big[\mathbf{Z}_\lambda(S)^{-1}\big] \preceq \left(\prod_{t=s+1}^{n} \frac{t - d_\lambda + 1}{t - d_\lambda} \right) \mathbf{Z}_\lambda(\{1..n\})^{-1} = \frac{n - d_\lambda + 1}{s - d_\lambda + 1} (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}. \qquad \blacksquare$$

3.3 Ridge regression with noisy responses

We apply the above result to obtain statistical guarantees for subsampling with regularized estimators.

Theorem 3.2. *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *and* $\widetilde{\mathbf{w}} \in \mathbb{R}^d$*, and suppose that* $\mathbf{y} = \mathbf{X}\widetilde{\mathbf{w}} + \boldsymbol{\xi}$*, where* $\boldsymbol{\xi}$ *is a mean zero vector with* $\mathrm{Var}[\boldsymbol{\xi}] \preceq \sigma^2\,\mathbf{I}$*. Let* S *be sampled according to* λ*-regularized size* $s \geq d_\lambda$ *volume sampling from* $\mathbf{X}$ *and* $\mathbf{w}^*_\lambda(S)$ *be the* λ*-ridge estimator of* $\widetilde{\mathbf{w}}$ *computed from subproblem* $(\mathbf{X}_S, \mathbf{y}_S)$*. Then, if* $\lambda \leq \frac{\sigma^2}{\|\widetilde{\mathbf{w}}\|^2}$*, we have*

$$\textit{(mean squared prediction error)} \quad \mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2\Big] \leq \frac{\sigma^2 d_\lambda}{s - d_\lambda + 1},$$

$$\textit{(mean squared error)} \qquad \mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\big[\|\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}}\|^2\big] \leq \frac{\sigma^2 n\,\mathrm{tr}((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1})}{s - d_\lambda + 1}.$$

Next, we present two lower-bounds for MSPE of a subsampled ridge estimator which show that the statistical guarantees achieved by regularized volume sampling are nearly optimal for $s \gg d_\lambda$ and better than standard approaches for $s = O(d_\lambda)$. In particular, we show that non-i.i.d. nature of volume sampling is essential if we want to achieve good generalization when the number of responses is close to d_λ. Namely, for certain data matrices, any subsampling procedure selecting examples in an i.i.d. fashion (e.g., leverage score sampling), requires more than $d_\lambda \ln(d_\lambda)$ responses to achieve MSPE below σ^2, whereas volume sampling obtains that bound for any matrix with $2d_\lambda$ responses.

Theorem 3.3. *For any* $p \geq 1$ *and* $\sigma \geq 0$*, there is* $d \geq p$ *such that for any sufficiently*

large n divisible by d there exists a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$ such that

$$d_\lambda(\mathbf{X}) \geq p \quad \text{for any} \quad 0 \leq \lambda \leq \sigma^2,$$

and for each of the following two statements there is a vector $\widetilde{\mathbf{w}} \in \mathbb{R}^d$ for which the corresponding regression problem $\mathbf{y} = \mathbf{X}\widetilde{\mathbf{w}} + \boldsymbol{\xi}$ with $\mathrm{Var}[\boldsymbol{\xi}] = \sigma^2\mathbf{I}$ satisfies that statement:

1. *For any subset $S \subseteq \{1..n\}$ of size s,*

$$\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2\Big] \geq \frac{\sigma^2 d_\lambda}{s + d_\lambda};$$

2. *For multiset $S \subseteq \{1..n\}$ of size $s \leq (d_\lambda - 1)\ln(d_\lambda)$, sampled i.i.d. from any distribution,*

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2\Big] \geq \sigma^2.$$

3.3.1 Upper bounds (proof of Theorem 3.2)

Standard analysis for the ridge regression estimator follows by performing bias-variance decomposition of the error, and then selecting λ so that bias can be appropriately bounded. We will recall this calculation for a fixed subproblem $(\mathbf{X}_S, \mathbf{y}_S)$. First, we compute the bias of the ridge estimator for a fixed set S (recall the shorthand $\mathbf{Z}_\lambda(S) = \mathbf{X}_S^\top\mathbf{X}_S + \lambda\mathbf{I}$):

$$\begin{aligned}
\mathrm{Bias}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)] &= \mathbb{E}[\mathbf{w}^*_\lambda(S)] - \widetilde{\mathbf{w}} = \mathbb{E}_{\boldsymbol{\xi}}\left[\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S^\top\mathbf{y}_S\right] - \widetilde{\mathbf{w}} \\
&= \mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S^\top\left(\mathbf{X}_S\widetilde{\mathbf{w}} + \cancel{\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\xi}_S]}\right) - \widetilde{\mathbf{w}} \\
&= (\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S^\top\mathbf{X}_S - \mathbf{I})\widetilde{\mathbf{w}} = -\lambda\,\mathbf{Z}_\lambda(S)^{-1}\widetilde{\mathbf{w}}.
\end{aligned}$$

Similarly, the covariance matrix of $\mathbf{w}^*_\lambda(S)$ is given by:

$$\begin{aligned}\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)] &= \mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S^\top \mathrm{Var}_{\boldsymbol{\xi}}[\boldsymbol{\xi}_S]\mathbf{X}_S\mathbf{Z}_\lambda(S)^{-1} \\ &\preceq \sigma^2\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}_S^\top\mathbf{X}_S\mathbf{Z}_\lambda(S)^{-1} = \sigma^2(\mathbf{Z}_\lambda(S)^{-1} - \lambda\,\mathbf{Z}_\lambda(S)^{-2}).\end{aligned}$$

Mean squared error of the ridge estimator for a fixed subset S can now be bounded by:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\big[\|\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}}\|^2\big] &= \mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)]) + \|\mathrm{Bias}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)]\|^2 \\ &\le \sigma^2\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1} - \lambda\mathbf{Z}_\lambda(S)^{-2}) + \lambda^2\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-2}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top) \\ &\le \sigma^2\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}) + \lambda\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-2})(\lambda\|\widetilde{\mathbf{w}}\|^2 - \sigma^2) \qquad (3.5) \\ &\le \sigma^2\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}), \qquad (3.6)\end{aligned}$$

where in (3.5) we applied Cauchy-Schwartz inequality for matrix trace, and in (3.6) we used the assumption that $\lambda \le \frac{\sigma^2}{\|\widetilde{\mathbf{w}}\|^2}$. Thus, taking expectation over the sampling of set S, we get

$$\begin{aligned}\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\big[\|\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}}\|^2\big] &\le \sigma^2\mathbb{E}_S\big[\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1})\big] \\ \text{(Theorem 3.1)} \quad &\le \sigma^2\frac{n - d_\lambda + 1}{s - d_\lambda + 1}\mathrm{tr}(\mathbf{Z}_\lambda(\{1..n\})^{-1}) \qquad (3.7) \\ &\le \frac{\sigma^2\, n\,\mathrm{tr}((\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1})}{s - d_\lambda + 1}.\end{aligned}$$

Next, we bound the mean squared prediction error. As before, we start with the standard bias-variance decomposition for fixed set S:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\big[\|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2\big] &= \mathrm{tr}(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}\mathbf{w}^*_\lambda(S)]) + \|\mathbf{X}(\mathbb{E}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)] - \widetilde{\mathbf{w}})\|^2 \\ &\le \sigma^2\mathrm{tr}(\mathbf{X}(\mathbf{Z}_\lambda(S)^{-1} - \lambda\,\mathbf{Z}_\lambda(S)^{-2})\mathbf{X}^\top) + \lambda^2\mathrm{tr}(\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{Z}_\lambda(S)^{-1}\widetilde{\mathbf{w}}\widetilde{\mathbf{w}}^\top) \\ &\le \sigma^2\mathrm{tr}(\mathbf{X}\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}^\top) + \lambda\,\mathrm{tr}(\mathbf{X}\mathbf{Z}_\lambda(S)^{-2}\mathbf{X}^\top)(\lambda\|\widetilde{\mathbf{w}}\|^2 - \sigma^2) \\ &\le \sigma^2\mathrm{tr}(\mathbf{X}\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}^\top).\end{aligned}$$

Once again, taking expectation over subset S, we have

$$\mathbb{E}_S\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S)-\widetilde{\mathbf{w}})\|^2\Big] \le \frac{\sigma^2}{n}\mathbb{E}_S\big[\mathrm{tr}(\mathbf{X}\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}^\top)\big] = \frac{\sigma^2}{n}\mathrm{tr}(\mathbf{X}\,\mathbb{E}_S[\mathbf{Z}_\lambda(S)^{-1}]\,\mathbf{X}^\top)$$
$$\text{(Theorem 3.1)} \quad \le \frac{\sigma^2}{n}\frac{n-d_\lambda+1}{s-d_\lambda+1}\mathrm{tr}(\mathbf{X}\mathbf{Z}_\lambda(\{1..n\})^{-1}\mathbf{X}^\top) \le \frac{\sigma^2 d_\lambda}{s-d_\lambda+1}. \quad (3.8)$$

The key part of proving both bounds is the application of Theorem 3.1. For MSE, we only used the trace version of the inequality (see (3.7)), however to obtain the bound on MSPE we used the more general positive semi-definite inequality in (3.8).

3.3.2 Lower bounds (proof of Theorem 3.3)

Let $d = \lceil p\rceil + 1$ and $n \ge \lceil\sigma^2\rceil d(d-1)$ be divisible by d. We define

$$\mathbf{X} \overset{def}{=} [\mathbf{I},...,\mathbf{I}]^\top \in \mathbb{R}^{n\times d}, \qquad \widetilde{\mathbf{w}}^\top \overset{def}{=} [a\sigma,...,a\sigma] \in \mathbb{R}^d$$

for some $a > 0$. For any $\lambda \le \sigma^2$, the λ-statistical dimension of $\mathbf{X}$ is

$$d_\lambda = \mathrm{tr}(\mathbf{X}\,\mathbf{Z}_\lambda(\{1..n\})^{-1}\mathbf{X}^\top) \ge \frac{\lceil\sigma^2\rceil d(d-1)}{\lceil\sigma^2\rceil(d-1)+\lambda} \ge \frac{d(d-1)}{d-1+1} \ge p.$$

Let $S \subseteq \{1..n\}$ be any set of size s, and for $i \in \{1..d\}$ let $s_i \overset{def}{=} |\{i \in S : \mathbf{x}_i = \mathbf{e}_i\}|$. The prediction variance of estimator $\mathbf{w}^*_\lambda(S)$ is equal to

$$\mathrm{tr}\big(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}\mathbf{w}^*_\lambda(S)]\big) = \sigma^2\mathrm{tr}(\mathbf{X}(\mathbf{Z}_\lambda(S)^{-1} - \lambda\mathbf{Z}_\lambda(S)^{-2})\mathbf{X}^\top)$$
$$= \frac{\sigma^2 n}{d}\sum_{i=1}^d\left(\frac{1}{s_i+\lambda} - \frac{\lambda}{(s_i+\lambda)^2}\right) = \frac{\sigma^2 n}{d}\sum_{i=1}^d\frac{s_i}{(s_i+\lambda)^2}.$$

The prediction bias of estimator $\mathbf{w}^*_\lambda(S)$ is equal to

$$\|\mathbf{X}(\mathbb{E}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)] - \widetilde{\mathbf{w}})\|^2 = \lambda^2\widetilde{\mathbf{w}}^\top\mathbf{Z}_\lambda(S)^{-1}\mathbf{X}^\top\mathbf{X}\mathbf{Z}_\lambda(S)^{-1}\widetilde{\mathbf{w}}$$
$$= \frac{\lambda^2 a^2\sigma^2 n}{d}\,\mathrm{tr}\big(\mathbf{Z}_\lambda(S)^{-2}\big) = \frac{\lambda^2 a^2\sigma^2 n}{d}\sum_{i=1}^d\frac{1}{(s_i+\lambda)^2}.$$

Thus, MSPE of estimator $\mathbf{w}^*_\lambda(S)$ is given by:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S)-\widetilde{\mathbf{w}})\|^2\Big] &= \frac{1}{n}\mathrm{tr}\big(\mathrm{Var}_{\boldsymbol{\xi}}[\mathbf{X}\mathbf{w}^*_\lambda(S)]\big)+\frac{1}{n}\|\mathbf{X}(\mathbb{E}_{\boldsymbol{\xi}}[\mathbf{w}^*_\lambda(S)]-\widetilde{\mathbf{w}})\|^2\\ &= \frac{\sigma^2}{d}\sum_{i=1}^{d}\left(\frac{s_i}{(s_i+\lambda)^2}+\frac{a^2\lambda^2}{(s_i+\lambda)^2}\right)=\frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{s_i+a^2\lambda^2}{(s_i+\lambda)^2}.\end{aligned}$$

Next, we find the λ that minimizes this expression. Taking the derivative with respect to λ we get:

$$\frac{\partial}{\partial\lambda}\left(\frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{s_i+a^2\lambda^2}{(s_i+\lambda)^2}\right)=\frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{2s_i(\lambda-a^{-2})}{(s_i+\lambda)^3}.$$

Thus, since at least one s_i has to be greater than 0, for any set S the derivative is negative for $\lambda < a^{-2}$ and positive for $\lambda > a^{-2}$, and the unique minimum of MSPE is achieved at $\lambda = a^{-2}$, regardless of which subset S is chosen. So, as we are seeking a lower bound, we can focus on the case of $\lambda = a^{-2}$.

Proof of Part 1. Let $a = 1$. As shown above, we can assume that $\lambda = 1$. In this case the formula simplifies to:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}\Big[\frac{1}{n}\|\mathbf{X}(\mathbf{w}^*_\lambda(S)-\widetilde{\mathbf{w}})\|^2\Big] &= \frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{s_i+1}{(s_i+1)^2}=\frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{1}{s_i+1}\\ &\overset{(*)}{\geq}\frac{\sigma^2}{\frac{s}{d}+1}=\frac{\sigma^2 d}{s+d}\geq\frac{\sigma^2 d_\lambda}{s+d_\lambda},\end{aligned}$$

where $(*)$ follows by applying Jensen's inequality to convex function $\phi(x)=\frac{1}{x+1}$.

Proof of Part 2. Let $a=\sqrt{2d}$. As shown above, we can assume that $\lambda = 1/(2d)$. Suppose that multiset S is sampled i.i.d. from some distribution over set $\{1..n\}$. Similarly as in Corollary 2.2, we exploit the Coupon Collector's problem, i.e. that if

$|S| \leq (d-1)\ln(d)$, then with probability at least $1/2$ there is $i \in \{1..d\}$ such that $s_i = 0$ (ie, one of the unit vectors $\mathbf{e}_i$ was never selected). Thus, MSPE can be lower-bounded as follows:

$$\mathbb{E}_S \mathbb{E}_{\boldsymbol{\xi}} \Big[\frac{1}{n} \|\mathbf{X}(\mathbf{w}^*_\lambda(S) - \widetilde{\mathbf{w}})\|^2 \Big] \geq \frac{1}{2} \frac{\sigma^2}{d} \frac{s_i + a^2\lambda^2}{(s_i+\lambda)^2} = \frac{\sigma^2}{2d} \frac{2d\lambda^2}{\lambda^2} = \sigma^2.$$

3.4 Efficient algorithms for regularized volume sampling

In this section we propose algorithms for efficiently performing volume sampling. This addresses the question posed by [AB13], asking for a polynomial-time algorithm for the case when the size of set S is $s > d$. [DR10] gave an algorithm for the case when $s = d$, which was later slightly improved by [GS12], running in time $O(nd^3)$. Recently, [LJS17] offered an algorithm for arbitrary s, which has complexity $O(n^4 s)$. We propose two new methods, which use our reverse iterative sampling technique to achieve faster running times for volume sampling of any size s. Both algorithms apply to the more general setting of λ-regularized volume sampling, and produce standard volume sampling as a special case for $\lambda = 0$ and $s \geq d$. The first algorithm has a deterministic runtime of $O(n-s+d)nd)$, whereas the second one is an accelerated version which with high probability finishes in time $O(nd^2)$. Thus, we obtain a direct improvement over [LJS17] by a factor of at least n^2, and in the special case of $s = d$, by a factor of d over the algorithm of [GS12].

Our algorithms implement reverse iterative sampling from Theorem 2.1. We start with the full index set $S = \{1..n\}$. In one step of the algorithm, we remove one

row from an index set S. After removing q rows, we are left with the index set of size $n-q$ that is distributed according to volume sampling for row set size $n-q$, and we proceed until our set S has the desired size s. The primary cost of the procedure is updating the conditional distribution $P(S_{-i}|S)$ at every step. It is convenient to store it using the unnormalized weights defined in (3.3) which, via Sylvester's theorem, can be computed as $h_i = 1 - \mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{x}_i$. Doing this naively, we would first compute $(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}$ which takes $O(nd^2)$ time[2]. After that for each i, we would multiply this matrix by $\mathbf{x}_i$ in time $O(d^2)$ to get the h_i's. The overall runtime of this naive method becomes:

$$\overbrace{\#\text{ of steps}}^{n-s} \times \Big(\overbrace{\text{compute } (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}}^{O(nd^2)} + \overbrace{\#\text{ of weights}}^{\leq n} \times \overbrace{\text{compute } h_i}^{O(d^2)} \Big) \;=\; O((n-s)nd^2).$$

We improve on this by observing that both the matrix $(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}$ and the weights h_i can be efficiently computed from the one obtained in the previous step by using the Sherman-Morrison formula. This lets us update the matrix inverse $(\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}$ in $O(d^2)$ time instead of $O(nd^2)$. We propose two strategies for dealing with the cost of maintaining the unnormalized probabilities:

1. Maintain all h_i's at every step, performing a cheap update step for every one of them;

2. Use rejection sampling, which avoids computing all h_i's, but makes each one more expensive.

[2]We are primarily interested in the case where $n \geq d$ and we state our time bounds under that assumption. However, when $\lambda > 0$, our techniques can be easily adapted to the case of $n < d$.

Algorithm 3.1: RegVol($\mathbf{X}, s, \lambda$)

$\mathbf{Z} \leftarrow (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$

$\forall_{i \in \{1..n\}} \quad h_i \leftarrow 1 - \mathbf{x}_i^\top \mathbf{Z} \mathbf{x}_i$

$S \leftarrow \{1..n\}$

while $|S| > s$

  Sample $i \propto h_i$ out of S

  $S \leftarrow S - i$

  $\mathbf{v} \leftarrow \mathbf{Z}\mathbf{x}_i / \sqrt{h_i}$

  $\forall_{j \in S} \quad h_j \leftarrow h_j - (\mathbf{x}_j^\top \mathbf{v})^2$

  $\mathbf{Z} \leftarrow \mathbf{Z} + \mathbf{v}\mathbf{v}^\top$

end

return S

Algorithm 3.2: FastRegVol($\mathbf{X}, s, \lambda$)

$\mathbf{Z} \leftarrow (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$

$S \leftarrow \{1..n\}$

while $|S| > \max\{s, 2d\}$

  repeat

    Sample i uniformly out of S

    $h_i \leftarrow 1 - \mathbf{x}_i^\top \mathbf{Z} \mathbf{x}_i$

    Sample $A \sim \text{Bernoulli}(h_i)$

  until $A = 1$

  $S \leftarrow S - i$

  $\mathbf{Z} \leftarrow \mathbf{Z} + h_i^{-1} \mathbf{Z}\mathbf{x}_i \mathbf{x}_i^\top \mathbf{Z}$

end

if $s < 2d$, $S \leftarrow \text{RegVol}(\mathbf{X}_S, s, \lambda)$

end

return S

As we can see, there is a trade-off between those strategies. In the following lemma, we will show that updating the value of h_i, given its value in the previous step only costs $O(d)$ time as opposed to $O(d^2)$. However, the number of h_i's that need to be computed for rejection sampling (explained shortly) can be far smaller.

Lemma 3.1. *For any matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$, set $S \subseteq \{1..n\}$ and two distinct indices $i, j \in S$, we have*

$$1 - \mathbf{x}_j^\top (\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})^{-1} \mathbf{x}_j = h_j - (\mathbf{x}_j^\top \mathbf{v})^2,$$

where $h_j = \mathbf{x}_j^\top (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{x}_j$ and $\mathbf{v} = \frac{1}{\sqrt{h_i}} (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1} \mathbf{x}_i$.

Proof. Letting $\mathbf{Z}_\lambda(S) = \mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I}$, we have

$$\begin{aligned} h_j - (\mathbf{x}_j^\top \mathbf{v})^2 &= 1 - \mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_j - \frac{(\mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i)^2}{1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i} \\ &= 1 - \mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_j - \frac{\mathbf{x}_j^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_j}{1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i} \\ &= 1 - \mathbf{x}_j^\top \left(\mathbf{Z}_\lambda(S)^{-1} + \frac{\mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i\mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}}{1 - \mathbf{x}_i^\top \mathbf{Z}_\lambda(S)^{-1}\mathbf{x}_i} \right) \mathbf{x}_j \\ &\overset{(*)}{=} 1 - \mathbf{x}_j^\top (\mathbf{X}_{S_{-i}}^\top \mathbf{X}_{S_{-i}} + \lambda \mathbf{I})^{-1}\mathbf{x}_j, \end{aligned}$$

where $(*)$ follows from the Sherman-Morrison formula. ∎

Thus the overall time complexity of reverse iterative sampling when using the first strategy goes down by a factor of d compared to the naive version (except for an initialization cost which stays at $O(nd^2)$).

Theorem 3.4. *Algorithm RegVol produces an index set S of rows distributed according to λ-regularized size s volume sampling over $\mathbf{X}$ in time $O((n-s+d)nd)$.*

Proof. Using Lemma 3.1 for h_i and the Sherman-Morrison formula for $\mathbf{Z}$, the following invariants hold at the beginning of the **while** loop:

$$h_i = 1 - \mathbf{x}_i^\top (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}\mathbf{x}_i \qquad \text{and} \qquad \mathbf{Z} = (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})^{-1}.$$

Runtime: Computing the initial $\mathbf{Z} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1}$ takes $O(nd^2)$, as does computing the initial values of h_j's. Inside the **while** loop, updating h_j's takes $O(|S|d) = O(nd)$ and updating $\mathbf{Z}$ takes $O(d^2)$. The overall runtime becomes $O(nd^2 + (n-s)nd) = O((n-s+d)nd)$, which completes the proof. ∎

Next we present algorithm FastRegVol, which is based on the rejection sampling strategy. Our key observation is that updating the full conditional distribution $P(S_{-i}|S)$ is wasteful, since the distribution changes very slowly throughout the procedure. Moreover, the unnormalized weights h_i, which are computed in the process are all bounded by 1. Thus, to sample from the correct distribution at any given iteration, we can employ rejection sampling as follows:

1. Sample i uniformly from set S,
2. Compute h_i,
3. Accept with probability h_i,
4. Otherwise, draw another sample.

Note that this rejection sampling can be employed locally, within each iteration of the algorithm. Thus, one rejection does not revert us back to the beginning of the algorithm. Moreover, if the probability of acceptance is high, then this strategy requires computing only a small number of weights per iteration of the algorithm, as opposed to updating all of them. This turns out to be the case for a majority of the steps of the algorithm, except at the very end (for $s \leq 2d$), were the conditional probabilities start changing more drastically. At that point, it becomes more efficient to use the first algorithm, RegVol.

Theorem 3.5. *For any $\lambda, \delta, s \geq 0$, algorithm FastRegVol samples according to λ-*

regularized size s volume sampling, and with probability at least $1-\delta$ runs in time

$$O\bigg(\Big(n+\log\big(n/d\big)\log\big(1/\delta\big)\Big)d^2\bigg).$$

Proof. We analyze the efficiency of rejection sampling in FastRegVol. Let R_t be a random variable corresponding to the number of trials needed in the **repeat** loop from line 4 in FastRegVol at the point when $|S| = t$. Note that conditioning on the algorithm's history, R_t is distributed according to geometric distribution $\mathrm{Ge}(q_t)$ with success probability:

$$q_t = \frac{1}{t}\sum_{i\in S}\big(1-\mathbf{x}_i^\top(\mathbf{X}_S^\top\mathbf{X}_S+\lambda\mathbf{I})^{-1}\mathbf{x}_i\big) \geq \frac{t-d}{t} \geq \frac{1}{2}.$$

Thus, even though variables R_t are not themselves independent, they can be upper-bounded by a sequence of independent variables $\widehat{R}_t \sim \mathrm{Ge}(\frac{t-d}{t})$. The expectation of the total number of trials in FastRegVol, $\bar{R} = \sum_t R_t$, can thus be bounded as follows:

$$\mathbb{E}[\bar{R}] \leq \sum_{t=2d}^{n}\mathbb{E}[\widehat{R}_t] = \sum_{t=2d}^{n}\frac{t}{t-d} \leq 2n.$$

Next, we will obtain a similar bound with high probability instead of in expectation. Here, we will have to use the fact that the variables $\widehat{R}_t$ are independent, which means that we can upper-bound their sum with high probability using standard concentration bounds for geometric distribution. For example, using Corollary 2.2 from [Jan18] one can immediately show that with probability at least $1-\delta$ we have $\bar{R} = O(n\ln\delta^{-1})$. However, more careful analysis shows an even better dependence on δ.

Lemma 3.2. *Let $\widehat{R}_t \sim \mathrm{Ge}(\frac{t-d}{t})$ be independent random variables. Then w.p. at least*

$1-\delta$

$$\sum_{t=2d}^{n} \widehat{R}_t = O\Big(n + \log\big(n/d\big)\log\big(1/\delta\big)\Big).$$

Each trial of rejection sampling requires computing one weight h_i in time $O(d^2)$. The overall time complexity of FastRegVol also includes computation and updating of matrix $\mathbf{Z}$ (in time $O(nd^2)$), rejection sampling which takes $O\left(\left(n + \log\left(\frac{n}{d}\right)\log\left(\frac{1}{\delta}\right)\right)d^2\right)$ time, and (if $s < 2d$) the RegVol portion, taking $O(d^3)$. ■

Proof of Lemma 3.2

As observed by [Jan18], tail-bounds for the sum of geometric random variables depend on the minimum acceptance probability among those variables. Note that for the vast majority of $\widehat{R}_t$'s the acceptance probability is very close to 1, so intuitively we should be able to take advantage of this to improve our tail bounds. To that end, we partition the variables into groups of roughly similar acceptance probability and then separately bound the sum of variables in each group. Let $J = \log(\frac{n}{d})$ (w.l.o.g. assume that J is an integer). For $1 \le j \le J$, let $I_j = \{d2^j, d2^j + 1, .., d2^{j+1}\}$ represent the j-th partition. We use the following notation for each partition:

$$\bar{R}_j \stackrel{def}{=} \sum_{t \in I_j} R_t, \qquad \mu_j \stackrel{def}{=} \mathbb{E}[\bar{R}_j], \qquad r_j \stackrel{def}{=} \min_{t \in I_j} \frac{t-d}{t}, \qquad \gamma_j \stackrel{def}{=} \frac{\log(\delta^{-1})}{d2^{j-2}} + 3.$$

Now, we apply Theorem 2.3 of [Jan18] to $\bar{R}_j$, obtaining

$$P(\bar{R}_j \ge \gamma_j\mu_j) \le \gamma_j^{-1}(1-r_j)^{(\gamma_j - 1 - \ln\gamma_j)\mu_j} \stackrel{(1)}{\le} (1-r_j)^{\gamma_j\mu_j/4} \stackrel{(2)}{\le} 2^{-j\gamma_j d2^{j-2}},$$

Dataset	$n \times d$	RegVol	FastRegVol	LSS
cadata	21k×8	33.5s	0.9s	0.1s
MSD	464k×90	>24hr	39s	12s
cpusmall	8k×12	1.7s	0.4s	0.07s
abalone	4k×8	0.5s	0.2s	0.03s

Table 3.1: A list of used regression datasets, with runtime comparison between RegVol and FastRegVol. We also provide the runtime for obtaining exact leverage score samples (LSS).

where (1) follows since $\gamma_j \geq 3$, and (2) holds because $\mu_j \geq d2^j$ and $r_j \geq 1 - 2^{-j}$. Moreover, for the chosen γ_j we have

$$j\gamma_j d2^{j-2} = j\log(\delta^{-1}) + 3jd2^{j-2} \geq \log(\delta^{-1}) + j = \log(2^j\delta^{-1}).$$

Let A denote the event that $\bar{R}_j \leq \gamma_j\mu_j$ for all $j \leq J$. Applying union bound, we get

$$P(A) \geq 1 - \sum_{j=1}^{J} P(\bar{R}_j \geq \gamma_j\mu_j) \geq 1 - \sum_{j=1}^{J} 2^{-\log(2^j\delta^{-1})} = 1 - \sum_{j=1}^{J} \frac{\delta}{2^j} \geq 1 - \delta.$$

If A holds, then we obtain the desired bound:

$$\begin{aligned}\sum_{t=2d}^{n} \widehat{R}_t &\leq \sum_{j=1}^{J} \gamma_j\mu_j \leq \sum_{j=1}^{J} \left(\frac{\log(\delta^{-1})}{d2^{j-2}} + 3\right) d2^{j+1} = 8J\log(\delta^{-1}) + 6\sum_{j=1}^{J} d2^j \\ &= O\Big(\log\big(n/d\big)\log\big(1/\delta\big) + n\Big).\end{aligned}$$

3.5 Experiments

In this section we experimentally evaluate the proposed volume sampling algorithms in terms of runtime and in the task of subsampling for linear regression. The list of implemented algorithms is:

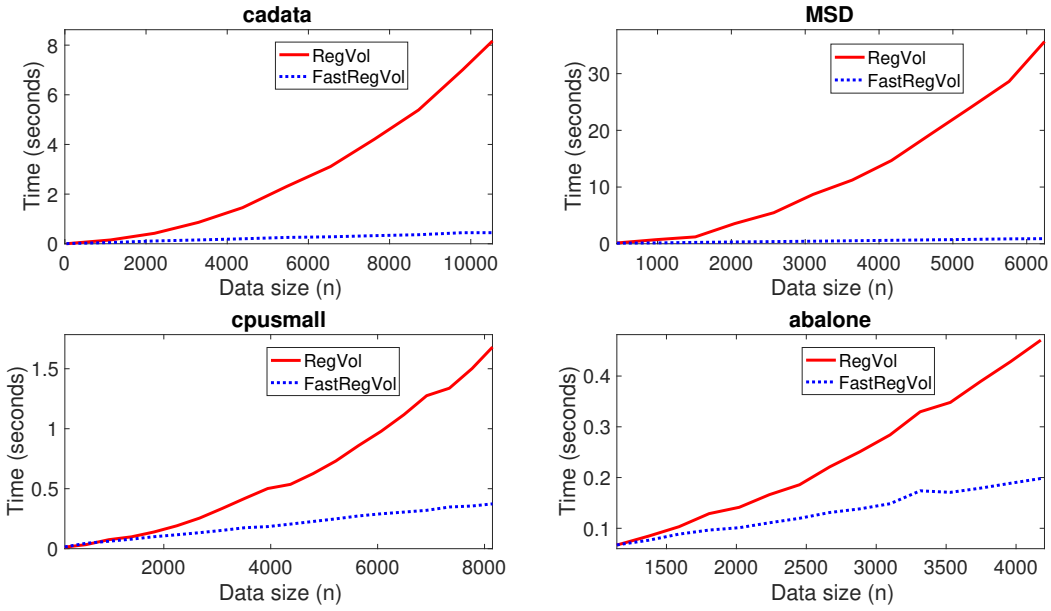

Figure 3.1: Comparison of runtime between FastRegVol and RegVol on four libsvm regression datasets [CL11], with the methods ran on data subsets of varying size (n).

1. Regularized volume sampling (algorithms FastRegVol and RegVol),

2. Leverage score sampling[3] (LSS) – a popular i.i.d. sampling technique [Mah11], where examples are selected w.p. $P(i) = (\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i)/d$.

3.5.1 Runtime comparison between the algorithms

The experiments were performed on several benchmark linear regression datasets from the libsvm repository [CL11]. Table 3.1 lists those datasets along with running times for sampling dimension many columns with each method. Dataset *MSD* was too big for RegVol to finish in reasonable time, however FastRegVol finished in less than 40

[3]Regularized variants of leverage scores have also been considered in context of kernel ridge regression [AM15]. However, in our experiments regularizing leverage scores did not provide any improvements.

seconds. In Figure 3.1 we plot the runtime against varying values of n (using portions of the datasets), to compare how FastRegVol and RegVol scale with respect to the data size. We observe that FastRegVol exhibits linear dependence on n, thus it is much better suited for running on large datasets.

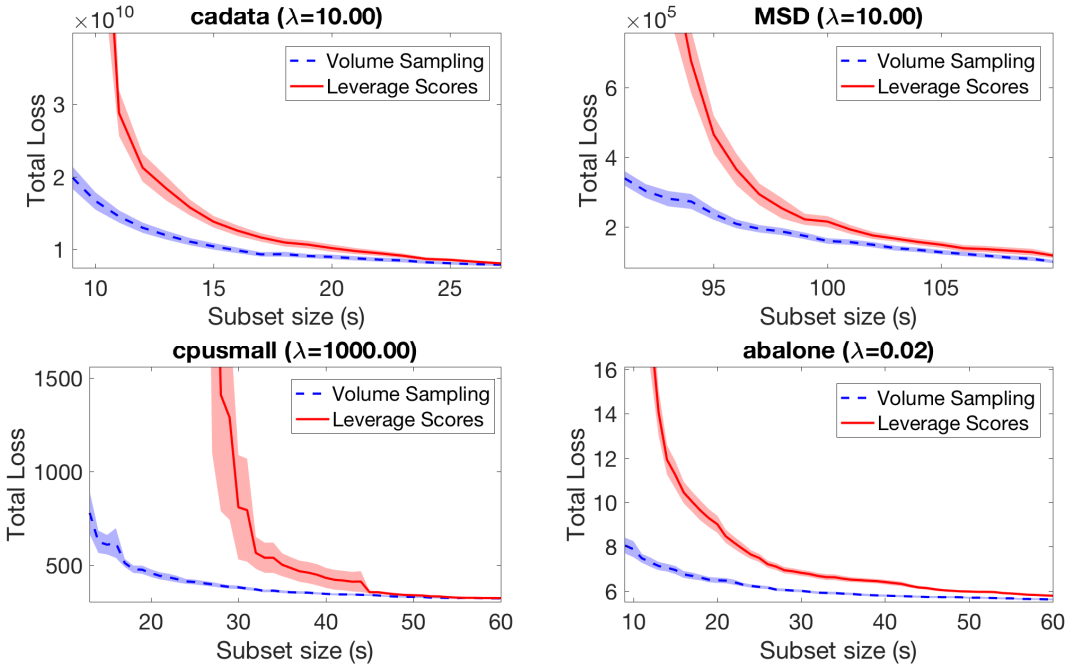

Figure 3.2: Comparison of loss of the subsampled ridge estimator when using regularized volume sampling vs using leverage score sampling on four datasets.

3.5.2 Subset selection for ridge regression

We applied volume sampling to the task of subset selection for linear regression, by evaluating the subsampled ridge estimator $\mathbf{w}^*_\lambda(S)$ using the total loss over the full

dataset, i.e.

$$\text{Total loss:} \quad \frac{1}{n}\|\mathbf{X}\mathbf{w}^*_\lambda(S) - \mathbf{y}\|^2, \quad \text{where} \quad \mathbf{w}^*_\lambda(S) = (\mathbf{X}_S^\top \mathbf{X}_S + \lambda \mathbf{I})\mathbf{X}_S^\top \mathbf{y}_S.$$

We evaluated the estimators for a range of subset sizes and values of λ, when the subsets are sampled according to λ-regularized volume sampling[4] and leverage score sampling. The results were averaged over 20 runs of each experiment. For clarity, Figure 3.2 shows the results only with one value of λ for each dataset, chosen so that the subsampled ridge estimator performed best (on average over all samples of preselected size s). Note that for leverage scores we did the appropriate rescaling of the instances before solving for $\mathbf{w}^*_\lambda(S)$ for the sampled subproblems (see [Mah11] for details). Volume sampling does not require any rescaling. The results on all datasets show that when only a small number of responses s is obtainable, then regularized volume sampling offers better estimators than leverage score sampling (as predicted by Theorems 3.2 and 3.3). The lower-bound from Theorem 3.3 part 2 can be observed for dataset *cpusmall*, where $d = 12$ and $d \log d \approx 30$.

3.6 Conclusion of the chapter

We proposed a sampling procedure called regularized volume sampling, which offers near-optimal statistical guarantees for subsampled ridge estimators. We also gave a new algorithm for volume sampling which is essentially as efficient as i.i.d. leverage score sampling.

[4] Our experiments suggest that using the same λ for sampling and for computing the ridge estimator works best.

Chapter 4

Leveraged volume sampling

4.1 Introduction

In the previous chapters we established that volume sampling is closely connected to linear least squares, and provided theoretical and experimental evidence that it is an effective tool for subset selection. Recall that in our setting the input points in $\mathbb{R}^d$ are provided, but the associated response for each point is withheld unless explicitly requested. In this chapter, we return to the worst-case response model of Chapter 2 (as opposed to the noisy response model of Chapter 3). The goal is to sample the responses for just a small subset of inputs, and then produce a weight vector whose total square loss on all n points is at most $1+\epsilon$ times that of the optimum, i.e., find $\widehat{\mathbf{w}}$ such that $L(\widehat{\mathbf{w}}) \leq (1+\epsilon)L(\mathbf{w}^*)$ (where $L(\cdot)$ is the square loss and $\mathbf{w}^*$ is the optimum). Unlike in Chapter 2, where we focused on bounds *in expectation*, in this chapter we will primarily focus on establishing the $1+\epsilon$ bound *with high probability*, i.e., w.p. at least $1-\delta$,

where the sample size depends both on ϵ and on failure probability δ (as well as on dimension d). Given the evidence of previous two chapters, it is surprising that using volume sampling in the context of linear regression with worst-case responses may in some cases lead to severely suboptimal performance, as we show in this chapter. We construct an example in which, even after sampling up to half of the responses, the loss of the weight vector from volume sampling is with a significant probability larger than the minimum loss by at least a fixed factor >1. Indeed, this poor behavior arises because for any sample size $>d$, the marginal probabilities from volume sampling are a mixture of uniform probabilities and leverage score probabilities, and uniform sampling is well-known to be suboptimal when the leverage scores are highly non-uniform.

A possible recourse is to abandon volume sampling in favor of leverage score sampling [DMM06, Woo14]. However, as discussed in the previous chapters, all i.i.d. sampling methods, including leverage score sampling, suffer from a coupon collector problem that prevents their effective use at small sample sizes. Moreover, the resulting weight vector is a *biased* estimator of the least squares solution based on all responses. This bias is a nuisance when averaging multiple solutions (e.g., as produced in distributed settings). In contrast, volume sampling offers multiplicative loss bounds even with sample sizes as small as d and it is the *only* known non-trivial method that gives unbiased weight vectors (see Chapter 2).

We develop a new solution, called *leveraged volume sampling*, that retains the aforementioned benefits of volume sampling while avoiding its flaws. Specifically, we propose a variant of volume sampling based on rescaling the input points to "correct"

the resulting marginals. On the algorithmic side, this leads to a new "determinantal rejection sampling" procedure which offers significant computational advantages over existing volume sampling algorithms, while at the same time being strikingly simple to implement. We prove that this new sampling scheme retains the benefits of volume sampling (like unbiasedness) but avoids the bad behavior demonstrated in our lower bound example. Along the way, we prove a new generalization of the Cauchy-Binet formula, which is needed for the rejection sampling denominator. Finally, we develop a new method for proving matrix tail bounds for leveraged volume sampling. Our analysis shows that the unbiased least-squares estimator constructed this way achieves a $1+\epsilon$ approximation factor from a sample of size $O(d \log d + d/\epsilon)$.

Experiments. Figure 4.1 presents experimental evidence on a benchmark dataset (*cpusmall_scale* from the libsvm collection [CL11]) that the potential bad behavior of volume sampling proven in our lower bound does occur in practice. Section 4.6 shows more datasets and a detailed discussion of the experiments. In summary, leveraged volume sampling avoids the bad behavior of standard volume sampling, and performs considerably better than leverage score sampling, especially for small sample sizes s.

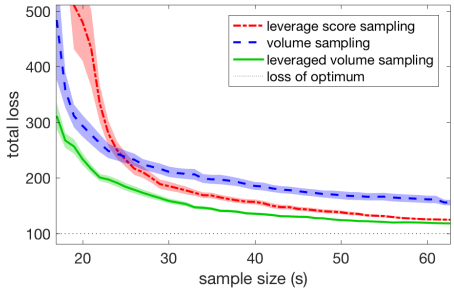

Figure 4.1: Plots of the total loss for the sampling methods (averaged over 100 runs) versus sample size (shading is standard error) for a libsvm dataset *cpusmall_scale* [CL11].

Outline of the chapter. In the next section, we present our lower bound for standard volume sampling. A new variant of rescaled volume sampling is introduced in Section 4.3. We develop techniques for proving matrix expectation formulas for this variant which show that for any rescaling the weight vector produced for the subproblem is unbiased. In this section we also show how leverage scores emerge as the natural choice of rescaling.

We prove multiplicative loss bounds for leveraged volume sampling in Section 4.4, by establishing two important conditions which are hard to prove for joint sampling procedures. Next, we present a surprisingly simple and efficient algorithm for leveraged volume sampling based on determinantal rejection sampling (Section 4.5): Other than the preprocessing step of computing leverage scores, the runtime does not depend on n (a major improvement over other volume sampling algorithms). Experimental evaluation of leveraged volume sampling in comparison to standard volume sampling and i.i.d. leverage score sampling is performed in Section 4.6. We conclude in Section 4.7 with an open problem.

4.2 Lower bound for standard volume sampling

Recall from Chapter 2 that given $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a size $s \geq d$, standard volume sampling jointly chooses a set S of s indices in $[n] \stackrel{def}{=} \{1..n\}$ with probability

$$P(S) = \frac{\det(\mathbf{X}_S^\top \mathbf{X}_S)}{\binom{n-d}{s-d} \det(\mathbf{X}^\top \mathbf{X})},$$

where $\mathbf{X}_S$ is the submatrix of the rows from $\mathbf{X}$ indexed by the set S. In the context of linear least squares, the learner then obtains the responses y_i, for $i \in S$, and uses the optimum solution $\mathbf{w}^*(S) = (\mathbf{X}_S)^+ \mathbf{y}_S$ for the subproblem $(\mathbf{X}_S, \mathbf{y}_S)$ as its weight vector. The goal is to obtain a multiplicative loss bound, i.e., that for some $\epsilon > 0$,

$$L(\mathbf{w}^*(S)) \leq (1+\epsilon)\, L(\mathbf{w}^*), \quad \text{where } \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \overbrace{\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2}^{L(\mathbf{w})}.$$

We show that standard volume sampling cannot guarantee $1+\epsilon$ multiplicative loss bounds on some instances, unless over half of the rows are chosen to be in the subsample.

Theorem 4.1. *Let $(\mathbf{X}, \mathbf{y})$ be an $n \times d$ least squares problem, such that*

$$\mathbf{X} = \begin{pmatrix} \mathbf{I}_{d\times d} \\ \hline \gamma\, \mathbf{I}_{d\times d} \\ \hline \vdots \\ \hline \gamma\, \mathbf{I}_{d\times d} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} \mathbf{1}_d \\ \hline \mathbf{0}_d \\ \hline \vdots \\ \hline \mathbf{0}_d \end{pmatrix}, \qquad \textit{where} \quad \gamma > 0.$$

Let $\mathbf{w}^*(S) = (\mathbf{X}_S)^+ \mathbf{y}_S$ be obtained from size s volume sampling for $(\mathbf{X}, \mathbf{y})$. Then,

$$\lim_{\gamma \to 0} \frac{\mathbb{E}[L(\mathbf{w}^*(S))]}{L(\mathbf{w}^*)} \;\geq\; 1 + \frac{n-s}{n-d}, \tag{4.1}$$

and there is a $\gamma > 0$ such that for any $s \leq \frac{n}{2}$,

$$P\left(L(\mathbf{w}^*(S)) \geq \left(1 + \frac{1}{2}\right) L(\mathbf{w}^*) \right) > \frac{1}{4}. \tag{4.2}$$

Proof. First, let us calculate $L(\mathbf{w}^*)$. Observe that

$$(\mathbf{X}^\top \mathbf{X})^{-1} = \overbrace{\left(1 + \frac{n-d}{d}\gamma^2\right)^{-1}}^{c}\, \mathbf{I},$$

$$\text{and} \quad \mathbf{w}^* = c\, \mathbf{X}^\top \mathbf{y} = c\, \mathbf{1}_d.$$

The loss $L(\mathbf{w})$ of any $\mathbf{w} \in \mathbb{R}^d$ can be decomposed as $L(\mathbf{w}) = \sum_{i=1}^d L_i(\mathbf{w})$, where $L_i(\mathbf{w})$ is the total loss incurred on all input vectors $\mathbf{e}_i$ or $\gamma\mathbf{e}_i$:

$$L_i(\mathbf{w}^*) = (1-c)^2 + \overbrace{\frac{n-d}{d}\gamma^2}^{\frac{1}{c}-1} c^2 = 1-c,$$

For $i \in [d]$, the i-th leverage score of $\mathbf{X}$ is equal $l_i = \mathbf{x}_i^\top (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i = c$, so we obtain that

$$L(\mathbf{w}^*) = d\,(1-c) = \sum_{i=1}^d (1-l_i). \tag{4.3}$$

Next, we compute $L(\mathbf{w}^*(S))$. Suppose that $S \subseteq \{1..n\}$ is produced by size s standard volume sampling. Note that if for some $1 \le i \le d$ we have $i \notin S$, then $(\mathbf{w}^*(S))_i = 0$ and therefore $L_i(\mathbf{w}^*(S)) = 1$. Moreover, denoting $b_i \overset{def}{=} \mathbf{1}_{[i\in S]}$,

$$(\mathbf{X}_S^\top\mathbf{X}_S)^{-1} \succeq (\mathbf{X}^\top\mathbf{X})^{-1} = c\,\mathbf{I}, \quad \text{and} \quad \mathbf{X}_S^\top\mathbf{y}_S = (b_1,\ldots,b_d)^\top,$$

so if $i \in S$, then $(\mathbf{w}^*(S))_i \ge c$ and

$$L_i(\mathbf{w}^*(S)) \ge \frac{n-d}{d}\,\gamma^2\,c^2 = \Big(\frac{1}{c}-1\Big)c^2 = c\,L_i(\mathbf{w}^*).$$

Putting the cases of $i \in S$ and $i \notin S$ together, we get

$$\begin{aligned} L_i(\mathbf{w}^*(S)) &\ge c\,L_i(\mathbf{w}^*) + (1 - c\,L_i(\mathbf{w}^*))\,(1-b_i) \\ &\ge c\,L_i(\mathbf{w}^*) + c^2(1-b_i). \end{aligned}$$

The marginal probability of the i-th row under volume sampling (see Proposition 2.1) is

$$P(i \in S) = \theta\; l_i + (1-\theta)\;1 = 1 - \theta\;(1-l_i), \quad \text{where } \theta = \frac{n-s}{n-d}. \tag{4.4}$$

Applying this formula, we note that

$$\mathbb{E}[1 - b_i] = 1 - P(i \in S) = \frac{n-s}{n-d}(1-c) = \frac{n-s}{n-d} L_i(\mathbf{w}^*).$$

Taking expectation over $L_i(\mathbf{w}^*(S))$ and summing the components over $i \in [d]$, we get

$$\mathbb{E}[L(\mathbf{w}^*(S))] \geq L(\mathbf{w}^*)\Big(c + c^2 \frac{n-s}{n-d}\Big).$$

Note that as $\gamma \to 0$, we have $c \to 1$, thus showing (4.1). It remains to show (4.2). We bound the probability that all of the first d input vectors were selected by volume sampling, using (4.3) in the process:

$$P\big([d] \subseteq S\big) \overset{(*)}{\leq} \prod_{i=1}^{d} P(i \in S) = \prod_{i=1}^{d} \Big(1 - \frac{n-s}{n-d}(1-l_i)\Big) \leq \exp\Big(-\frac{n-s}{n-d} \overbrace{L(\mathbf{w}^*)}^{\sum_{i=1}^{d}(1-l_i)}\Big),$$

where $(*)$ follows from negative associativity of volume sampling (see [LJS17]). If for some $i \in [d]$ we have $i \notin S$, then $L(\mathbf{w}^*(S)) \geq 1$. So for γ such that $L(\mathbf{w}^*) = \frac{2}{3}$ and any $s \leq \frac{n}{2}$:

$$P\bigg(L(\mathbf{w}^*(S)) \geq \Big(1 + \frac{1}{2}\Big)\overbrace{L(\mathbf{w}^*)}^{2/3}\bigg) \geq 1 - \exp\Big(-\frac{n-s}{n-d} \cdot \frac{2}{3}\Big) \geq 1 - \exp\Big(-\frac{1}{2} \cdot \frac{2}{3}\Big) > \frac{1}{4}. \blacksquare$$

Note that this lower bound only makes use of the negative associativity of volume sampling and the form of the marginals. However the tail bounds we prove in Section 4.4 rely on more subtle properties of volume sampling. We begin by creating a variant of volume sampling with rescaled marginals.

4.3 Rescaled volume sampling

Given any size $s \geq d$, our goal is to jointly sample s row indices $\pi_1, \ldots, \pi_s$ with replacement (instead of a *subset* S of $[n]$ of size s, we get a *sequence* $\pi \in [n]^s$).

The second difference to standard volume sampling is that we rescale the i-th row (and response) by $\frac{1}{\sqrt{q_i}}$, where $q = (q_1, ..., q_n)$ is any discrete distribution over the set of row indices $[n]$, such that $\sum_{i=1}^n q_i = 1$ and $q_i > 0$ for all $i \in [n]$. We now define q-rescaled size s volume sampling as a joint sampling distribution over $\pi \in [n]^s$, s.t.

$$q\text{-rescaled size } s \text{ volume sampling:} \qquad P(\pi) \;\sim\; \det\Big(\sum_{i=1}^{s} \frac{1}{q_{\pi_i}} \mathbf{x}_{\pi_i} \mathbf{x}_{\pi_i}^\top\Big) \;\prod_{i=1}^{s} q_{\pi_i}. \tag{4.5}$$

Using the following rescaling matrix $\mathbf{Q}_\pi \overset{def}{=} \sum_{i=1}^{|\pi|} \frac{1}{q_{\pi_i}} \mathbf{e}_{\pi_i} \mathbf{e}_{\pi_i}^\top \in \mathbb{R}^{n \times n}$, we rewrite the determinant as $\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})$. As in standard volume sampling, the normalization factor in rescaled volume sampling can be given in a closed form through a novel extension of the Cauchy-Binet formula.

Proposition 4.1. *For any $\mathbf{X} \in \mathbb{R}^{n \times d}$, $s \geq d$ and $q_1, \ldots, q_n > 0$, such that $\sum_{i=1}^n q_i = 1$, we have*

$$\sum_{\pi \in [n]^s} \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \prod_{i=1}^{s} q_{\pi_i} = s(s-1)...(s-d+1)\, \det(\mathbf{X}^\top \mathbf{X}).$$

Proof. In this proof, we illustrate a technique which is also useful for showing Theorem 4.3, as well as Proposition 4.3. The key idea is to first apply the Cauchy-Binet formula to the determinant term specified by a fixed sequence $\pi \in [n]^s$, and then apply it again at the end. Starting with a single term, we have

$$\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) = \sum_{S \in \binom{[s]}{d}} \det(\mathbf{X}^\top \mathbf{Q}_{\pi_S} \mathbf{X}) \overset{(*)}{=} \sum_{S \in \binom{[s]}{d}} \det(\mathbf{X}_{\pi_S})^2 \prod_{i \in S} \frac{1}{q_{\pi_i}}.$$

where $\binom{[s]}{d} \overset{def}{=} \{S \subseteq \{1, \ldots, s\} : \; |S| = d\}$ and π_S denotes a subsequence of π indexed by the elements of set S. Note that $(*)$ uses the fact that $\mathbf{X}_{\pi_S}$ is a square matrix. Next,

we compute the sum, using the above identity:

$$\begin{aligned}
\sum_{\pi\in[n]^s} \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \prod_{i=1}^{s} q_{\pi_i} &= \sum_{\pi\in[n]^s} \sum_{S\in\binom{[s]}{d}} \det(\mathbf{X}_{\pi_S})^2 \prod_{i\in[s]\setminus S} q_{\pi_i} \\
&\overset{(1)}{=} \binom{s}{d} \sum_{\bar{\pi}\in[n]^d} \det(\mathbf{X}_{\bar{\pi}})^2 \sum_{\tilde{\pi}\in[n]^{s-d}} \prod_{i=1}^{s-d} q_{\tilde{\pi}_i} \\
&= \binom{s}{d} \sum_{\bar{\pi}\in[n]^d} \det(\mathbf{X}_{\bar{\pi}})^2 \Big(\sum_{i=1}^{n} q_i\Big)^{s-d} \\
&\overset{(2)}{=} \binom{s}{d} d! \sum_{S\in\binom{[n]}{d}} \det(\mathbf{X}_S)^2 = s(s-1)...(s-d+1)\det(\mathbf{X}^\top\mathbf{X}),
\end{aligned}$$

Note that in (1) we separate π into two parts (subset S and its complement, $[s]\setminus S$) and sum over them separately. The binomial coefficient $\binom{s}{d}$ counts the number of ways that S can be "placed into" the sequence π. In (2) we observe that q_i's sum to 1, and that whenever $\bar{\pi}$ has repetitions, determinant $\det(\mathbf{X}_{\bar{\pi}})$ is zero, so we can switch to summing over sets. Finally, (3) again uses the standard size d Cauchy-Binet formula, now for the entire matrix $\mathbf{X}$. ■

The following proposition states that rescaled volume sampling is closed under subsampling with standard volume sampling, demonstrating a direct connection between the two distributions. This mirrors a corresponding composition property of standard volume sampling (see Corollary 2.1).

Proposition 4.2. *Consider the following sampling procedure, for $t > s$:*

$$\begin{aligned}
\pi &\overset{t}{\sim} \quad \mathbf{X} && \textit{(q-rescaled size } t \textit{ volume sampling)}, \\
S &\overset{s}{\sim} \quad \begin{pmatrix} \frac{1}{\sqrt{q_{\pi_1}}}\mathbf{x}_{\pi_1}^\top \\ \cdots \\ \frac{1}{\sqrt{q_{\pi_t}}}\mathbf{x}_{\pi_t}^\top \end{pmatrix} = \big(\mathbf{Q}_{[1..n]}^{1/2}\mathbf{X}\big)_\pi && \textit{(standard size } s \textit{ volume sampling)}.
\end{aligned}$$

Then π_S is distributed according to q-rescaled size s volume sampling from $\mathbf{X}$.

Proof. First step of the reverse iterative sampling procedure of standard volume sampling described in Section 2.2.1 involves removing one row from the given matrix with probability proportional to the square volume of that submatrix:

$$\forall_{i\in S} \qquad P(i \,|\, \pi_S) = \frac{\det(\mathbf{X}^\top \mathbf{Q}_{\pi_{S_{-i}}} \mathbf{X})}{(|S|-d)\det(\mathbf{X}^\top \mathbf{Q}_{\pi}\mathbf{X})}.$$

Suppose that $s = t-1$ and let $\tilde{\pi} = \pi_S \in [n]^{t-1}$ denote the sequence obtained after performing one step of the row-removal procedure. Then,

$$\begin{aligned} P(\tilde{\pi}) &= \sum_{i=1}^{n} t \quad \overbrace{P(i \,|\, [\tilde{\pi}, i])}^{\text{removing one row}} \qquad \overbrace{P([\tilde{\pi}, i])}^{\text{rescaled sampling}} \\ &= \sum_{i=1}^{n} t \, \frac{\det(\mathbf{X}^\top \mathbf{Q}_{\tilde{\pi}}\mathbf{X})}{(t-d)\det(\mathbf{X}^\top \mathbf{Q}_{[\tilde{\pi},i]}\mathbf{X})} \, \frac{\det(\mathbf{X}^\top \mathbf{Q}_{[\tilde{\pi},i]}\mathbf{X})\,(\prod_{j=1}^{t-1} q_{\tilde{\pi}_j})\, q_i}{\frac{t!}{(t-d)!}\det(\mathbf{X}^\top\mathbf{X})} \\ &= \frac{\det(\mathbf{X}^\top \mathbf{Q}_{\tilde{\pi}}\mathbf{X})(\prod_{j=1}^{t-1} q_{\tilde{\pi}_j})}{\frac{t-d}{t}\frac{t!}{(t-d)!}\det(\mathbf{X}^\top\mathbf{X})} \sum_{i=1}^{n} q_i = \frac{\det(\mathbf{X}^\top \mathbf{Q}_{\tilde{\pi}}\mathbf{X})\,(\prod_{j=1}^{s} q_{\tilde{\pi}_j})}{\frac{s!}{(s-d)!}\det(\mathbf{X}^\top\mathbf{X})}, \end{aligned}$$

where the factor t next to the sum counts the number of ways to place index i into the sequence $\tilde{\pi}$. Thus, by induction, for any $s < t$ the algorithm correctly samples from q-rescaled volume sampling. ■

4.3.1 Expectations for rescaled volume sampling

In this section, we show that the key properties of standard volume sampling, such as the unbiasedness of least squares estimators, are also exhibited by any q-rescaled volume sampling. First, we give a construction of an estimator most appropriate for this setting. Given a matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, vector $\mathbf{y} \in \mathbb{R}^n$ and a sequence $\pi \in [n]^s$, we are

interested in a least-squares problem $(\mathbf{Q}_\pi^{1/2}\mathbf{X}, \mathbf{Q}_\pi^{1/2}\mathbf{y})$, which selects instances indexed by π, and rescales each of them by the corresponding $1/\sqrt{q_i}$. This leads to a natural subsampled least squares estimator

$$\mathbf{w}^*(\pi) = \operatorname*{argmin}_{\mathbf{w}} \sum_{i=1}^{s} \frac{1}{q_{\pi_i}} \left(\mathbf{x}_{\pi_i}^\top \mathbf{w} - y_{\pi_i}\right)^2 = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^+ \mathbf{Q}_\pi^{1/2}\mathbf{y}.$$

The key property of standard volume sampling is that the subsampled least-squares estimator is unbiased. Surprisingly this property is retained for any q-rescaled volume sampling. As we shall see this will give us great leeway for choosing q to optimize our algorithms.

Theorem 4.2. *Given a full rank $\mathbf{X} \in \mathbb{R}^{n\times d}$ and a response vector $\mathbf{y} \in \mathbb{R}^n$, for any q as above, if π is sampled according to (4.5), then*

$$\mathbb{E}[\mathbf{w}^*(\pi)] = \mathbf{w}^*, \quad \textit{where} \quad \mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2.$$

Proof. We demonstrate an interesting application of the composition property (Proposition 4.2). Suppose that π is sampled according to q-rescaled size s volume sampling, and then set S is sampled according to standard size d volume sampling from $(\mathbf{Q}_{[1..n]}^{1/2}\mathbf{X})_\pi$. Note that $\mathbf{w}^*(\pi_S)$ is the exact solution of a system of d linear equations:

$$\frac{1}{\sqrt{q_{\pi_i}}}\mathbf{x}_{\pi_i}^\top \mathbf{w} = \frac{1}{\sqrt{q_{\pi_i}}} y_{\pi_i}, \qquad \text{for} \quad i \in S.$$

Thus, the rescaling of each equation by $\frac{1}{\sqrt{q_{\pi_i}}}$ cancels out, and we can simply write $\mathbf{w}^*(\pi_S) = (\mathbf{X}_{\pi_S})^+ \mathbf{y}_{\pi_S}$. Note that this is not the case for sets larger than d whenever the optimum solution incurs positive loss. Now, applying the unbiasedness formula

(Theorem 2.3) for standard volume sampling followed by the law of total expectation, we have:

$$\mathbb{E}[\mathbf{w}^*(\pi)] = \mathbb{E}\big[\,\mathbb{E}[\mathbf{w}^*(\pi_S)\,|\pi]\,\big] = \mathbb{E}[\mathbf{w}^*(\pi_S)].$$

Note that Proposition 4.2 states that π_S is distributed according to q-rescaled size d volume sampling, which is in fact the same as standard size d volume sampling, because the rescaling comes out of the determinant and then cancels out:

$$P(\pi_S) \sim \det(\mathbf{X}^\top \mathbf{Q}_{\pi_S}\mathbf{X}) \prod_{i\in S} q_{\pi_i} = \det(\mathbf{X}_{\pi_S}^\top \mathbf{X}_{\pi_S}) \left(\prod_{i\in S} \frac{1}{q_{\pi_i}}\right) \prod_{i\in S} q_{\pi_i} = \det(\mathbf{X}_{\pi_S}^\top \mathbf{X}_{\pi_S}).$$

Thus, we can simply apply Theorem 2.3 again, showing that $\mathbb{E}[\mathbf{w}^*(\pi_S)] = \mathbf{w}^*$. ■

The matrix variance formula for standard volume sampling from Theorem 2.4 has a natural extension to any rescaled volume sampling, turning here into an inequality.

Theorem 4.3. *Given a full rank $\mathbf{X} \in \mathbb{R}^{n\times d}$ and any q as above, if π is sampled according to (4.5), then*

$$\mathbb{E}\big[(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\big] \preceq \frac{1}{s-d+1}(\mathbf{X}^\top \mathbf{X})^{-1}.$$

Proof. We will prove that for any vector $\mathbf{v} \in \mathbb{R}^d$,

$$\mathbb{E}\big[\mathbf{v}^\top (\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\mathbf{v}\big] \leq \frac{\mathbf{v}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{v}}{s-d+1},$$

which immediately implies the corresponding matrix inequality. First, we use Sylvester's formula, which holds whenever a matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is full rank:

$$\det(\mathbf{A} + \mathbf{v}\mathbf{v}^\top) = \det(\mathbf{A})\big(1 + \mathbf{v}^\top \mathbf{A}^{-1}\mathbf{v}\big).$$

Note that whenever the matrix is not full rank, its determinant is 0 (in which case we avoid computing the matrix inverse), so we have for any $\pi \in [n]^s$:

$$\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})\ \mathbf{v}^\top (\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\mathbf{v} \leq \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X} + \mathbf{v}\mathbf{v}^\top) - \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})$$
$$\overset{(*)}{=} \sum_{S \in \binom{[s]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top \mathbf{X}_{\pi_S} + \mathbf{v}\mathbf{v}^\top) \prod_{i \in S} \frac{1}{q_{\pi_i}},$$

where $(*)$ follows from applying the Cauchy-Binet formula to both of the determinants, and canceling out common terms. Next, we proceed similarly as in the proof of Proposition 4.1, letting $Z = d!\binom{s}{d}\det(\mathbf{X}^\top \mathbf{X})$ and summing over all $\pi \in [n]^s$:

$$\begin{aligned}
Z\ \mathbb{E}\big[\mathbf{v}^\top (\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\mathbf{v}\big] &= \sum_{\pi \in [n]^s} \mathbf{v}^\top (\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X})^{-1}\mathbf{v} \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \prod_{i=1}^{s} q_{\pi_i} \\
&\leq \sum_{\pi \in [n]^s} \sum_{S \in \binom{[s]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top \mathbf{X}_{\pi_S} + \mathbf{v}\mathbf{v}^\top) \prod_{i \in [s]\setminus S} q_{\pi_i} \\
&= \binom{s}{d-1} \sum_{\bar{\pi} \in [n]^{d-1}} \det(\mathbf{X}_{\bar{\pi}}^\top \mathbf{X}_{\bar{\pi}} + \mathbf{v}\mathbf{v}^\top) \sum_{\tilde{\pi} \in [n]^{s-d+1}} \prod_{i=1}^{s-d+1} q_{\pi_i} \\
&= \binom{s}{d-1}(d-1)! \sum_{S \in \binom{[n]}{d-1}} \det(\mathbf{X}_S^\top \mathbf{X}_S + \mathbf{v}\mathbf{v}^\top) \\
&= \frac{d!\binom{s}{d}}{s-d+1}\big(\det(\mathbf{X}^\top \mathbf{X} + \mathbf{v}\mathbf{v}^\top) - \det(\mathbf{X}^\top \mathbf{X})\big) \\
&= Z\,\frac{\mathbf{v}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{v}}{s-d+1}.
\end{aligned}$$

■

4.3.2 Leveraged volume sampling: a natural rescaling

Rescaled volume sampling can be viewed as selecting a sequence π of s rank-1 covariates from the covariance matrix $\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top$. If $\pi_1, \ldots, \pi_s$ are sampled i.i.d. from q, i.e. $P(\pi) = \prod_{i=1}^s q_{\pi_i}$, then matrix $\frac{1}{s}\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}$ is an unbiased estimator

of the covariance matrix because $\mathbb{E}[q_{\pi_i}^{-1}\mathbf{x}_{\pi_i}\mathbf{x}_{\pi_i}^\top] = \mathbf{X}^\top\mathbf{X}$. In rescaled volume sampling (4.5), $P(\pi) \sim \big(\prod_{i=1}^s q_{\pi_i}\big)\frac{\det(\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{X}^\top\mathbf{X})}$, and the latter volume ratio introduces a bias to that estimator. However, we show that this bias vanishes when q is exactly proportional to the leverage scores.

Proposition 4.3. *For any q and $\mathbf{X}$ as before, if $\pi \in [n]^s$ is sampled according to (4.5), then*

$$\mathbb{E}[\mathbf{Q}_\pi] = (s-d)\,\mathbf{I} + \operatorname{diag}\Big(\frac{l_1}{q_1},\ldots,\frac{l_n}{q_n}\Big), \quad \textit{where} \quad l_i = \mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_i.$$

In particular, $\mathbb{E}[\frac{1}{s}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}] = \mathbf{X}^\top\mathbb{E}[\frac{1}{s}\mathbf{Q}_\pi]\mathbf{X} = \mathbf{X}^\top\mathbf{X}$ if and only if $q_i = \frac{l_i}{d} > 0$ for all $i \in [n]$.

Proof. First, we compute the marginal probability of a fixed element of sequence π containing a particular index $i \in [n]$ under q-rescaled volume sampling:

$$\begin{aligned} Z\ P(\pi_s{=}i) &= \sum_{\pi\in[n]^{s-1}} \det(\mathbf{X}^\top\mathbf{Q}_{[\pi,i]}\mathbf{X})\ q_i \prod_{t=1}^{s-1} q_{\pi_t} \\ &= \underbrace{q_i \sum_{\pi\in[n]^{s-1}} \sum_{S\in\binom{[s-1]}{d}} \det(\mathbf{X}_{\pi_S})^2 \prod_{t\in[s-1]\setminus S} q_{\pi_t}}_{T_1} + \underbrace{\sum_{\pi\in[n]^{s-1}} \sum_{S\in\binom{[s-1]}{d-1}} \det(\mathbf{X}_{\pi_S}^\top\mathbf{X}_{\pi_S} + \mathbf{x}_i\mathbf{x}_i^\top) \prod_{t\in[s-1]\setminus S} q_{\pi_t}}_{T_2}, \end{aligned}$$

where the first term can be computed by following the derivation in the proof of Proposition 4.1, obtaining $T_1 = q_i\frac{s-d}{s}\,Z$, and the second term is derived as in the proof of Theorem 4.3, obtaining $T_2 = \frac{l_i}{s}\,Z$. Putting this together, we get

$$P(\pi_s{=}i) = \frac{1}{s}\big((s-d)\,q_i + l_i\big).$$

Note that by symmetry this applies to any element of the sequence. We can now easily compute the desired expectation:

$$\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big] = \frac{1}{q_i}\sum_{t=1}^{s} P(\pi_t = i) = (s-d) + \frac{l_i}{q_i}. \qquad \blacksquare$$

The special rescaling defined by setting $q_i = \frac{l_i}{d}$, which we call *leveraged volume sampling*, has other remarkable properties, including algorithms and tail bounds, as shown in the following sections. As we shall see, these properties often hold when distribution q is merely approximately proportional to the leverage scores.

4.4 Multiplicative tail bounds for linear regression

An analysis of leverage score sampling, essentially following [Woo14, Section 2] which in turn draws from [Sar06], highlights two basic sufficient conditions on the (random) subsampling matrix $\mathbf{Q}_\pi$ that lead to multiplicative tail bounds for $L(\mathbf{w}^*(\pi))$.

To derive these conditions, it is convenient to shift to an orthogonalization of the linear regression task $(\mathbf{X}, \mathbf{y})$ by replacing matrix $\mathbf{X}$ with a matrix $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2} \in \mathbb{R}^{n\times d}$. It is easy to check that the columns of $\mathbf{U}$ have unit length and are orthogonal, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. Now, $\mathbf{v}^* = \mathbf{U}^\top\mathbf{y}$ is the least-squares solution for the orthogonal problem $(\mathbf{U}, \mathbf{y})$ and prediction vector $\mathbf{U}\mathbf{v}^* = \mathbf{U}\mathbf{U}^\top\mathbf{y}$ for $(\mathbf{U}, \mathbf{y})$ is the same as the prediction vector $\mathbf{X}\mathbf{w}^* = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ for the original problem $(\mathbf{X}, \mathbf{y})$. The same property holds for the subsampled estimators, i.e., $\mathbf{U}\mathbf{v}^*(\pi) = \mathbf{X}\mathbf{w}^*(\pi)$, where $\mathbf{v}^*(\pi) = (\mathbf{Q}_\pi^{1/2}\mathbf{U})^+\mathbf{Q}_\pi^{1/2}\mathbf{y}$. Volume sampling probabilities are also preserved under this transformation, so w.l.o.g.

we can work with the orthogonal problem. Now $L(\mathbf{v}^*(\pi))$ can be rewritten as

$$L(\mathbf{v}^*(\pi)) = \|\mathbf{U}\mathbf{v}^*(\pi) - \mathbf{y}\|^2 \stackrel{(1)}{=} \|\mathbf{U}\mathbf{v}^* - \mathbf{y}\|^2 + \|\mathbf{U}(\mathbf{v}^*(\pi) - \mathbf{v}^*)\|^2$$
$$\stackrel{(2)}{=} L(\mathbf{v}^*) + \|\mathbf{v}^*(\pi) - \mathbf{v}^*\|^2, \tag{4.6}$$

where (1) follows via Pythagorean theorem from the fact that $\mathbf{U}(\mathbf{v}^*(\pi) - \mathbf{v}^*)$ lies in the column span of $\mathbf{U}$ and the residual vector $\mathbf{r} = \mathbf{U}\mathbf{v}^* - \mathbf{y}$ is orthogonal to all columns of $\mathbf{U}$, and (2) follows from $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. By the definition of $\mathbf{v}^*(\pi)$, we can write $\|\mathbf{v}^*(\pi) - \mathbf{v}^*\|^2$ as follows:

$$\|\mathbf{v}^*(\pi) - \mathbf{v}^*\| = \|(\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U})^{-1} \ \ \mathbf{U}^\top\mathbf{Q}_\pi(\mathbf{y} - \mathbf{U}\mathbf{v}^*)\| \leq \|(\underset{d\times d}{\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}})^{-1}\| \, \|\underset{d\times 1}{\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}}\|, \tag{4.7}$$

where $\|\mathbf{A}\|$ denotes the matrix 2-norm (i.e., the largest singular value) of $\mathbf{A}$; when $\mathbf{A}$ is a vector, then $\|\mathbf{A}\|$ is its Euclidean norm. Thus, we break our task down to showing two key conditions:

1. *Matrix multiplication:* an upper bound on the Euclidean norm $\|\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}\|$,

2. *Subspace embedding:* an upper bound on the matrix 2-norm $\|(\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U})^{-1}\|$.

4.4.1 Tail bounds for i.i.d. leverage scores

As an example of applying this type of analysis for i.i.d. sampling, we show how to establish the above two conditions for leverage score sampling. Thus, for this section only, we will assume that sequence $\pi \in [n]^s$ is sampled according to i.i.d. leverage scores:

$$P(\pi) = \prod_{i=1}^{s} q_{\pi_i}, \qquad \text{where} \quad q_i = \frac{l_i}{d}.$$

Note that the estimator $\mathbf{w}^*(\pi)$ produced this way is no longer unbiased, which is the motivation behind developing leveraged volume sampling.

Orthogonalization described at the beginning of the section yields a simple form of the ith leverage score: $l_i = \|\mathbf{u}_i\|^2$. Using this fact, a matrix multiplication guarantee can be easily shown for leverage score sampling. Since $\mathbf{z}_i \overset{def}{=} r_{\pi_i}(d/\|\mathbf{u}_{\pi_i}\|^2)\mathbf{u}_{\pi_i}$ for $i \in [s]$ are i.i.d. random vectors with $\mathbb{E}[\mathbf{z}_i] = \mathbf{U}^\top \mathbf{r} = \mathbf{0}$ and $\mathbb{E}[\|\mathbf{z}_i\|^2] = d\|\mathbf{r}\|^2$,

$$\mathbb{E}\Big[\big\|\mathbf{U}^\top \mathbf{Q}_\pi\, \mathbf{r}\big\|^2\Big] = \mathbb{E}\left[\left\|\sum_{i=1}^{s} r_{\pi_i} \cdot \frac{d}{s\|\mathbf{u}_{\pi_i}\|^2}\, \mathbf{u}_{\pi_i}\right\|^2\right] = \mathbb{E}\left[\left\|\frac{1}{s}\sum_{i=1}^{s} \mathbf{z}_i\right\|^2\right] = \frac{\mathbb{E}[\|\mathbf{z}_1\|^2]}{s} = \frac{d}{s}\cdot\|\mathbf{r}\|^2.$$

Applying Markov's inequality, we conclude that sample size $s \geq \frac{8d}{\epsilon\delta}$ is sufficient to show that $\|\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{r}\|^2 \leq \epsilon\|\mathbf{r}\|^2$ with probability at least $1-\delta$.

We show the subspace embedding condition by decomposing matrix $\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U}$ into a sum of independent random matrices:

$$\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U} = \sum_{i=1}^{s} \mathbf{Z}_i, \quad \mathbf{Z}_i \overset{def}{=} \frac{d}{s\,\|\mathbf{u}_{\pi_i}\|^2}\mathbf{u}_{\pi_i}\mathbf{u}_{\pi_i}^\top,$$

where the 2-norm of each matrix is bounded by $\|\mathbf{Z}_i\| \leq \frac{d}{s}$. By a standard matrix Chernoff bound (see Corollary 5.2 and Remark 5.3 of [Tro12]):

$$P\left(\lambda_{\min}\Big(\sum_{t=1}^{s}\mathbf{Z}_t\Big) \leq \frac{1}{2}\right) \leq d\exp\left(-\frac{s}{8d}\right).$$

This lower bounds the smallest eigenvalue of $\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U}$ by $1/2$, and thus upper bounds the 2-norm of the inverse by 2, with sample size $s = O(d\ln(\frac{d}{\delta}))$.

Putting everything together with inequalities (4.6) and (4.7), we have shown that size $O(d\ln(\frac{d}{\delta}) + \frac{d}{\epsilon\delta})$ i.i.d. leverage score sampling yields a weight vector with loss at most $(1+\epsilon)L(\mathbf{w}^*)$ with probability at least $1-\delta$.

4.4.2 Tail bounds for leveraged volume sampling

We now show the guarantees of matrix multiplication and subspace embedding for leveraged volume sampling. Due to the jointness of this distribution, the task is considerably more challenging than for i.i.d. sampling, and requires different mathematical machinery.

We start with a theorem that implies strong guarantees for approximate matrix multiplication with leveraged volume sampling. Unlike with i.i.d. sampling, this result requires controlling the pairwise dependence between indices selected under rescaled volume sampling. Its proof is an interesting application of a classical Hadamard matrix product inequality from [AAHRJ87] (Proof in Section 4.4.3).

Theorem 4.4. *Let* $\mathbf{U} \in \mathbb{R}^{n\times d}$ *be a matrix s.t.* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. *If sequence* $\pi \in [n]^s$ *is selected using leveraged volume sampling of size* $s \geq \frac{2d}{\epsilon}$, *then for any* $\mathbf{r} \in \mathbb{R}^n$,

$$\mathbb{E}\left[\left\|\frac{1}{s}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{r} - \mathbf{U}^\top\mathbf{r}\right\|^2\right] \leq \epsilon\,\|\mathbf{r}\|^2.$$

Next, we turn to the subspace embedding condition. The following result is remarkable because standard matrix tail bounds used to prove this condition for leverage score sampling are not applicable to volume sampling. In fact, obtaining matrix Chernoff bounds for negatively associated joint distributions like volume sampling is an active area of research, as discussed in [HO14]. We address this challenge by defining a coupling procedure for volume sampling and uniform sampling without replacement, which leads to a curious reduction argument described in Section 4.4.4.

Theorem 4.5. *Let* $\mathbf{U} \in \mathbb{R}^{n\times d}$ *be a matrix s.t.* $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. *There is an absolute constant* C, *s.t. if sequence* $\pi \in [n]^s$ *is selected using leveraged volume sampling of size* $s \geq C\, d \ln(\frac{d}{\delta})$, *then*

$$P\left(\lambda_{\min}\left(\frac{1}{s}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\right) \leq \frac{1}{8}\right) \leq \delta.$$

Theorems 4.4 and 4.5 imply that the unbiased estimator $\mathbf{w}^*(\pi)$ produced from leveraged volume sampling achieves multiplicative tail bounds with sample size $s = O(d\log d + d/\epsilon)$.

Corollary 4.1. *Let* $\mathbf{X} \in \mathbb{R}^{n\times d}$ *be a full rank matrix. There is an absolute constant* C, *s.t. if sequence* $\pi \in [n]^s$ *is selected using leveraged volume sampling of size* $s \geq C\left(d\ln(\frac{d}{\delta}) + \frac{d}{\epsilon\delta}\right)$, *then for estimator*

$$\mathbf{w}^*(\pi) = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{Q}_\pi^{1/2}(\mathbf{X}\mathbf{w} - \mathbf{y})\|^2,$$

we have $L(\mathbf{w}^*(\pi)) \leq (1+\epsilon)\, L(\mathbf{w}^*)$ *with probability at least* $1-\delta$.

Proof. Let $\mathbf{U} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1/2}$. Combining Theorem 4.4 with Markov's inequality, we have that for large enough C, w.h.p.

$$\|\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}\|^2 \leq \epsilon\,\frac{s^2}{8^2}\|\mathbf{r}\|^2,$$

where $\mathbf{r} = \mathbf{y} - \mathbf{U}\mathbf{v}^*$. Finally following (4.6) and (4.7) above, we have that w.h.p.

$$\begin{aligned} L(\mathbf{w}^*(\pi)) &\leq L(\mathbf{w}^*) + \|(\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U})^{-1}\|^2\,\|\mathbf{U}^\top\mathbf{Q}_\pi\,\mathbf{r}\|^2 \\ &\leq L(\mathbf{w}^*) + \frac{8^2}{s^2}\,\epsilon\frac{s^2}{8^2}\,\|\mathbf{r}\|^2 \\ &= (1+\epsilon)\, L(\mathbf{w}^*). \end{aligned}$$

■

4.4.3 Matrix multiplication (proof of Theorem 4.4)

We rewrite the expected square norm as:

$$\mathbb{E}\left[\left\|\frac{1}{s}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{r}-\mathbf{U}^\top\mathbf{r}\right\|^2\right]=\mathbb{E}\left[\left\|\mathbf{U}^\top\Big(\frac{1}{s}\mathbf{Q}_\pi-\mathbf{I}\Big)\mathbf{r}\right\|^2\right]=\mathbb{E}\left[\mathbf{r}^\top\Big(\frac{1}{s}\mathbf{Q}_\pi-\mathbf{I}\Big)\mathbf{U}\mathbf{U}^\top\Big(\frac{1}{s}\mathbf{Q}_\pi-\mathbf{I}\Big)\mathbf{r}\right]$$

$$=\mathbf{r}^\top\,\mathbb{E}\left[\Big(\frac{1}{s}\mathbf{Q}_\pi-\mathbf{I}\Big)\mathbf{U}\mathbf{U}^\top\Big(\frac{1}{s}\mathbf{Q}_\pi-\mathbf{I}\Big)\right]\mathbf{r}$$

$$\leq\lambda_{\max}\Big(\underbrace{\big(\mathbb{E}[(z_i-1)(z_j-1)]\,\mathbf{u}_i^\top\mathbf{u}_j\big)_{ij}}_{\mathbf{M}}\Big)\,\|\mathbf{r}\|^2,\ \text{where } z_i=\frac{1}{s}(\mathbf{Q}_\pi)_{ii}.$$

It remains to bound $\lambda_{\max}(\mathbf{M})$. By Proposition 4.3, for leveraged volume sampling $\mathbb{E}[(\mathbf{Q}_\pi)_{ii}]=s$, so

$$\mathbb{E}[(z_i-1)(z_j-1)]=\frac{1}{s^2}\Big(\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}(\mathbf{Q}_\pi)_{jj}\big]-\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]\mathbb{E}\big[(\mathbf{Q}_\pi)_{jj}\big]\Big)=\frac{1}{s^2}\operatorname{cov}\big[(\mathbf{Q}_\pi)_{ii},(\mathbf{Q}_\pi)_{jj}\big].$$

For rescaled volume sampling this is given in the following lemma:

Lemma 4.1. *For any* $\mathbf{X}$ *and* q*, if sequence* $\pi\in[n]^s$ *is sampled from* q*-rescaled volume sampling then*

$$\operatorname{cov}\big[(\mathbf{Q}_\pi)_{ii},\,(\mathbf{Q}_\pi)_{jj}\big]=\mathbf{1}_{i=j}\frac{1}{q_i}\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]-(s-d)-\frac{(\mathbf{x}_i^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_j)^2}{q_iq_j}.$$

Since $\|\mathbf{u}_i\|^2=l_i=dq_i$ and $\mathbf{u}_i^\top(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{u}_j=\mathbf{u}_i^\top\mathbf{u}_j$, we can express matrix $\mathbf{M}$ as follows:

$$\mathbf{M}=\operatorname{diag}\Big(\frac{d\,\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]}{\|\mathbf{u}_i\|^2s^2}\|\mathbf{u}_i\|^2\Big)_{i=1}^n-\frac{s-d}{s^2}\mathbf{U}\mathbf{U}^\top-\frac{d^2}{s^2}\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2\|\mathbf{u}_j\|^2}\right)_{ij}.$$

The first term simplifies to $\frac{d}{s}\mathbf{I}$, and the second term is negative semi-definite, so

$$\lambda_{\max}(\mathbf{M})\leq\frac{d}{s}+\frac{d^2}{s^2}\left\|\left(\frac{(\mathbf{u}_i^\top\mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2\|\mathbf{u}_j\|^2}\right)_{ij}\right\|.$$

Finally, we decompose the last term into a Hadamard product of matrices, and apply a classical inequality by [AAHRJ87] (symbol "$\circ$" denotes Hadamard matrix product):

$$\begin{aligned}
\left\| \left(\frac{(\mathbf{u}_i^\top \mathbf{u}_j)^3}{\|\mathbf{u}_i\|^2 \|\mathbf{u}_j\|^2} \right)_{ij} \right\| &= \left\| \left(\frac{\mathbf{u}_i^\top \mathbf{u}_j}{\|\mathbf{u}_i\| \, \|\mathbf{u}_j\|} \right)_{ij} \circ \left(\frac{(\mathbf{u}_i^\top \mathbf{u}_j)^2}{\|\mathbf{u}_i\| \|\mathbf{u}_j\|} \right)_{ij} \right\| \\
&\leq \left\| \left(\frac{(\mathbf{u}_i^\top \mathbf{u}_j)^2}{\|\mathbf{u}_i\| \|\mathbf{u}_j\|} \right)_{ij} \right\| = \left\| \left(\frac{\mathbf{u}_i^\top \mathbf{u}_j}{\|\mathbf{u}_i\| \, \|\mathbf{u}_j\|} \right)_{ij} \circ \mathbf{U}\mathbf{U}^\top \right\| \\
&\leq \|\mathbf{U}\mathbf{U}^\top\| = 1.
\end{aligned}$$

Thus, we conclude that $\mathbb{E}[\|\frac{1}{s}\mathbf{U}^\top \mathbf{Q}_\pi \mathbf{r} - \mathbf{U}^\top \mathbf{r}\|^2] \leq (\frac{d}{s} + \frac{d^2}{s^2})\|\mathbf{r}\|^2$, completing the proof.

Proof of Lemma 4.1

We compute marginal probability of two elements in the sequence π having particular values $i, j \in [n]$:

$$Z\, P\big((\pi_{s-1}{=}i) \wedge (\pi_s{=}j)\big) = \sum_{\pi \in [n]^{s-2}} \sum_{S \in \binom{[s]}{d}} \det(\mathbf{X}^\top_{[\pi,i,j]_S} \mathbf{X}_{[\pi,i,j]_S}) \prod_{t \in [s] \setminus S} q_{[\pi,i,j]_t}.$$

We partition the set $\binom{[s]}{d}$ of all subsets of size d into four groups, and summing separately over each of the groups, we have

$$Z\, P\big((\pi_{s-1}{=}i) \wedge (\pi_s{=}j)\big) = T_{00} + T_{01} + T_{10} + T_{11}, \qquad \text{where:}$$

1. Let $G_{00} = \{S \in \binom{[s]}{d} : \; s{-}1 \notin S, \; s \notin S\}$, and following derivation in the proof of Proposition 4.1, we have

$$T_{00} = q_i\, q_j \sum_{\pi \in [n]^{s-2}} \sum_{S \in G_{00}} \det(\mathbf{X}_{\pi_S})^2 \prod_{t \in [s-2] \setminus S} q_{\pi_t} = q_i\, q_j \frac{(s{-}d{-}1)(s{-}d)}{(s{-}1)\, s}\, Z.$$

2. Let $G_{10} = \{S \in \binom{[s]}{d} : \ s-1 \in S, \ s \notin S\}$, and following derivation in the proof of Theorem 4.3, we have

$$T_{10} = q_j \sum_{\pi \in [n]^{s-1}} \sum_{S \in G_{10}} \det(\mathbf{X}_{[\pi,i]_S})^2 \prod_{t \in [s-1] \setminus S} q_{[\pi,i]_t} = l_i \, q_j \frac{(s-d)}{(s-1)\, s} \, Z.$$

3. $G_{01} = \{S \in \binom{[s]}{d} : \ s-1 \notin S, \ s \in S\}$, and by symmetry, $T_{01} = l_j \, q_i \frac{(s-d)}{(s-1)\, s} \, Z$.

4. Let $G_{11} = \{S \in \binom{[s]}{d} : \ s-1 \in S, \ s \in S\}$, and the last term is

$$\begin{aligned}
T_{11} &= \sum_{\pi \in [n]^{s-1}} \sum_{S \in G_{11}} \det(\mathbf{X}_{[\pi,i,j]_S})^2 \prod_{t \in [s] \setminus S} q_{[\pi,i,j]_t} \\
&= \binom{s-2}{d-2} \sum_{\pi \in [n]^{d-2}} \det(\mathbf{X}_{[\pi,i,j]})^2 \\
&= \binom{s-2}{d-2} (d-2)! \big(\det(\mathbf{X}^\top \mathbf{X}) - \det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i}) - \det(\mathbf{X}_{-j}^\top \mathbf{X}_{-j}) + \det(\mathbf{X}_{-i,j}^\top \mathbf{X}_{-i,j})\big) \\
&\overset{(*)}{=} \frac{d!\binom{s}{d}}{s(s-1)} \det(\mathbf{X}^\top \mathbf{X}) \Big(1 - \underbrace{(1-l_i)}_{\frac{\det(\mathbf{X}_{-i}^\top \mathbf{X}_{-i})}{\det(\mathbf{X}^\top \mathbf{X})}} - \underbrace{(1-l_j)}_{\frac{\det(\mathbf{X}_{-j}^\top \mathbf{X}_{-j})}{\det(\mathbf{X}^\top \mathbf{X})}} + \underbrace{(1-l_i)(1-l_j) - l_{ij}^2}_{\frac{\det(\mathbf{X}_{-i,j}^\top \mathbf{X}_{-i,j})}{\det(\mathbf{X}^\top \mathbf{X})}} \Big) \\
&= \frac{Z}{s(s-1)} \big(\ell_i \ell_j - \ell_{ij}^2\big),
\end{aligned}$$

where $l_{ij} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_j$, and $(*)$ follows from repeated application of Sylvester's determinant formula. We can now compute the expectation for $i \neq j$:

$$\begin{aligned}
\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii} \, (\mathbf{Q}_\pi)_{jj}\big] &= \frac{1}{q_i \, q_j} \sum_{t_1=1}^{s} \sum_{t_2=1}^{s} P\big((\pi_{s-1} = i) \wedge (\pi_s = j)\big) \\
&= \frac{s(s-1)}{q_i \, q_j} \overbrace{P\big((\pi_{s-1} = i) \wedge (\pi_s = j)\big)}^{\frac{1}{Z}(T_{00}+T_{10}+T_{01}+T_{11})} \\
&= (s-d-1)(s-d) + (s-d)\frac{l_i}{q_i} + (s-d)\frac{l_j}{q_j} + \frac{l_i l_j}{q_i \, q_j} - \frac{l_{ij}^2}{q_i \, q_j} \\
&= \Big((s-d)q_i + \frac{l_i}{q_i}\Big)\Big((s-d)q_j + \frac{l_j}{q_j}\Big) - (s-d) - \frac{l_{ij}^2}{q_i \, q_j} \\
&= \mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big] \, \mathbb{E}\big[(\mathbf{Q}_\pi)_{jj}\big] - (s-d) - \frac{l_{ij}^2}{q_i q_j}.
\end{aligned}$$

Finally, if $i = j$, then

$$\begin{aligned}\mathbb{E}[(\mathbf{Q}_\pi)_{ii}\,(\mathbf{Q}_\pi)_{ii}] &= \frac{1}{q_i^2}\sum_{t_1=1}^{s}\sum_{t_2=1}^{s} P(\pi_{t_1}{=}i \,\wedge\, \pi_{t_2}{=}i)\\ &= \frac{s(s-1)}{q_i^2}\,P(\pi_{s-1}{=}i \,\wedge\, \pi_s{=}i) + \frac{s}{q_i^2}\,P(\pi_s{=}i)\\ &= \big(\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big]\big)^2 - (s-d) - \frac{l_i^2}{q_i^2} + \frac{1}{q_i}\mathbb{E}\big[(\mathbf{Q}_\pi)_{ii}\big].\end{aligned}$$

4.4.4 Subspace embedding (proof of Theorem 4.5)

We break the sampling procedure down into two stages. First, we do leveraged volume sampling of a sequence $\pi \in [n]^m$ of size $m \geq C_0 d^2/\delta$, then we do standard size s volume sampling from matrix $(\mathbf{Q}^{1/2}_{[1..n]}\mathbf{U})_\pi$. Since rescaled volume sampling is closed under this subsampling (Proposition 4.2), this procedure is equivalent to size s leveraged volume sampling from $\mathbf{U}$. To show that the first stage satisfies the subspace embedding condition, we simply use the bound from Theorem 4.4:

Lemma 4.2. *There is an absolute constant C_0, s.t. if sequence $\pi \in [n]^m$ is generated via leveraged volume sampling of size m at least $C_0\, d^2/\delta$ from $\mathbf{U}$, then*

$$P\bigg(\lambda_{\min}\Big(\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\Big) \leq \frac{1}{2}\bigg) \leq \delta.$$

The size of m is much larger than what we claim is sufficient. However, we use it to achieve a tighter bound in the second stage. To obtain substantially smaller sample sizes for subspace embedding than what Theorem 4.4 can deliver, it is standard to use tail bounds for the sums of independent matrices. However, applying these results to joint sampling is a challenging task. Interestingly, [LJS17] showed that volume

sampling is a strongly Raleigh measure, implying that the sampled vectors are negatively correlated. This guarantee is sufficient to show tail bounds for real-valued random variables [PP14, see, e.g.,], however it has proven challenging in the matrix case, as discussed by [HO14]. One notable exception is uniform sampling without replacement, which is a negatively correlated joint distribution. A reduction argument originally proposed by [Hoe63], but presented in this context by [GN10], shows that uniform sampling without replacement offers the same tail bounds as i.i.d. uniform sampling.

Lemma 4.3. *Assume that* $\lambda_{\min}\big(\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\big) \geq \frac{1}{2}$. *Suppose that set* T *is a set of fixed size sampled uniformly without replacement from* $[m]$. *There is a constant* C_1 *s.t. if* $|T| \geq C_1\, d\ln(d/\delta)$, *then*

$$P\Big(\lambda_{\min}\Big(\frac{1}{|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \leq \frac{1}{4}\Big) \leq \delta.$$

The proof of Lemma 4.3 (provided at the end of this section) is a straightforward application of the argument given by [GN10]. We now propose a different reduction argument showing that a subspace embedding guarantee for uniform sampling without replacement leads to a similar guarantee for volume sampling. We achieve this by exploiting a volume sampling procedure introduced in Chapter 3, shown here in Algorithm 4.1. This procedure relies on iteratively removing elements from the set S until we are left with s elements. Specifically, at each step, we sample an index i from a conditional distribution, $i \sim P(i\,|\,S) = (1 - \mathbf{u}_i^\top(\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U})^{-1}\mathbf{u}_i)/(|S| - d)$. Crucially for us, each step proceeds via rejection sampling with the proposal distribution being uniform. We can easily modify the algorithm, so that the samples from the proposal

distribution are used to construct a uniformly sampled set T, as shown in Algorithm 4.2. Note that sets S returned by both algorithms are identically distributed, and furthermore, T is a subset of S, because every index taken out of S is also taken out of T.

Algorithm 4.1: Volume sampling

1: $S \leftarrow [m]$
2: **while** $|S| > s$
3: **repeat**
4: Sample i unif. out of S
5: $q \leftarrow 1 - \mathbf{u}_i^\top (\mathbf{U}^\top \mathbf{Q}_{\pi_S} \mathbf{U})^{-1} \mathbf{u}_i$
6: Sample $Accept \sim \text{Bernoulli}(q)$
7: **until** $Accept = \text{true}$
8: $S \leftarrow S - i$
9: **end**
10: **return** S

Algorithm 4.2: Coupled sampling

1: $S, T \leftarrow [m]$
2: **while** $|S| > s$
3: Sample i unif. out of $[m]$
4: $T \leftarrow T - i$
5: **if** $i \in S$
6: $q \leftarrow 1 - \mathbf{u}_i^\top (\mathbf{U}^\top \mathbf{Q}_{\pi_S} \mathbf{U})^{-1} \mathbf{u}_i$
7: Sample $Accept \sim \text{Bernoulli}(q)$
8: **if** $Accept = \text{true}$, $S \leftarrow S - i$ **end**
9: **end**
10: **end**
11: **return** S, T

By Lemma 4.3, if size of T is at least $C_1 \, d \log(d/\delta)$, then this set offers a subspace embedding guarantee. Next, we will show that in fact set T is not much smaller than S, implying that the same guarantee holds for S. Specifically, we will show that $|S \setminus T| = O(d \log(d/\delta))$. Note that it suffices to bound the number of times that a uniform sample is rejected by sampling $A = 0$ in line 7 of Algorithm 4.2. Denote this

number by R. Note that $R = \sum_{t=s+1}^{m} R_t$, where $m = |Q|$ and R_t is the number of times that $A = 0$ was sampled while the size of set S was t. Variables R_t are independent, and each is distributed according to the geometric distribution (number of failures until success), with the success probability

$$r_t = \frac{1}{t} \sum_{i \in S} \left(1 - \mathbf{u}_i^\top (\mathbf{U}^\top \mathbf{Q}_{\pi_S} \mathbf{U})^{-1} \mathbf{u}_i\right) = \frac{1}{t}\Big(t - \mathrm{tr}\big((\mathbf{U}^\top \mathbf{Q}_{\pi_S} \mathbf{U})^{-1} \mathbf{U}^\top \mathbf{Q}_{\pi_S} \mathbf{U}\big)\Big) = \frac{t-d}{t}.$$

Now, as long as $\frac{m-d}{s-d} \leq C_0\, d^2/\delta$, we can bound the expected value of R as follows:

$$\mathbb{E}[R] = \sum_{t=s+1}^{m} \mathbb{E}[R_t] = \sum_{t=s+1}^{m} \Big(\frac{t}{t-d} - 1\Big) = d \sum_{t=s-d+1}^{m-d} \frac{1}{t} \leq d\, \ln\Big(\frac{m-d}{s-d}\Big) \leq C_2\, d \ln(d/\delta).$$

In this step, we made use of the first stage sampling, guaranteeing that the term under the logarithm is bounded. Next, we show that the upper tail of R decays very rapidly given a sufficiently large gap between m and s:

Lemma 4.4. *Let $R_t \sim \mathrm{Geom}(\frac{t-d}{t})$ be a sequence of independent geometrically distributed random variables (number of failures until success). Then, for any $d < s < m$ and $a > 1$,*

$$P\big(R \geq a\ \mathbb{E}[R]\big) \leq e^{\frac{a}{2}} \left(\frac{s-d}{m-d}\right)^{\frac{a}{2}-1} \quad \textit{for} \quad R = \sum_{t=s+1}^{m} R_t.$$

Let $a = 4$ in Lemma 4.4. Setting $C = C_1 + 2a\, C_2$, for any $s \geq C\, d \ln(d/\delta)$, using $m = \max\{C_0\, \frac{d^2}{\delta},\ d + \mathrm{e}^2 \frac{s}{\delta}\}$, we obtain that

$$R \leq a\, C_2\, d \ln(d/\delta) \leq s/2, \quad \text{w.p.} \quad \geq 1 - \mathrm{e}^2\, \frac{s-d}{m-d} \geq 1 - \delta,$$

showing that $|T| \geq s - R \geq C_1\, d \ln(d/\delta)$ and $s \leq 2|T|$.

Therefore, by Lemmas 4.2, 4.3 and 4.4, there is a $1-3\delta$ probability event in which

$$\lambda_{\min}\Big(\frac{1}{|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \frac{1}{4} \quad \text{and} \quad s \leq 2|T|.$$

In this same event,

$$\lambda_{\min}\Big(\frac{1}{s}\mathbf{U}^\top\mathbf{Q}_{\pi_S}\mathbf{U}\Big) \geq \lambda_{\min}\Big(\frac{1}{s}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \lambda_{\min}\Big(\frac{1}{2|T|}\mathbf{U}^\top\mathbf{Q}_{\pi_T}\mathbf{U}\Big) \geq \frac{1}{2}\cdot\frac{1}{4} = \frac{1}{8},$$

which completes the proof of Theorem 4.5.

Proof of Lemma 4.2

Replacing vector $\mathbf{r}$ in Theorem 4.4 with each column of matrix $\mathbf{U}$, we obtain that for $m \geq C\,\frac{d}{\epsilon}$,

$$\mathbb{E}\big[\|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{U}^\top\mathbf{U}\|_F^2\big] \leq \epsilon\,\|\mathbf{U}\|_F^2 = \epsilon\,d.$$

We bound the 2-norm by the Frobenius norm and use Markov's inequality, showing that w.p. $\geq 1-\delta$

$$\|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{I}\| \leq \|\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U} - \mathbf{I}\|_F \leq \sqrt{\epsilon\,d/\delta}.$$

Setting $\epsilon = \frac{\delta}{4d}$, for $m \geq C_0\,d^2/\delta$, the above inequality implies that

$$\lambda_{\min}\Big(\frac{1}{m}\mathbf{U}^\top\mathbf{Q}_\pi\mathbf{U}\Big) \geq \frac{1}{2}.$$

Proof of Lemma 4.3

Let π denote the sequence of m indices selected by volume sampling in the first stage. Suppose that $i_1, ..., i_s$ are independent uniformly sampled indices from $[m]$,

and let $j_1, ..., j_s$ be indices sampled uniformly without replacement from $[m]$. We define matrices

$$\mathbf{Z} \stackrel{def}{=} \sum_{t=1}^{s} \overbrace{\frac{1}{sq_{i_t}} \mathbf{u}_{i_t} \mathbf{u}_{i_t}^\top}^{\mathbf{Z}_t}, \quad \text{and} \quad \widehat{\mathbf{Z}} \stackrel{def}{=} \sum_{t=1}^{s} \overbrace{\frac{1}{sq_{j_t}} \mathbf{u}_{j_t} \mathbf{u}_{j_t}^\top}^{\widehat{\mathbf{Z}}_t}.$$

Note that $\|\mathbf{Z}_t\| = \frac{d}{s\, l_i} \|\mathbf{u}_{i_t}\|^2 = \frac{d}{s}$ and, similarly, $\|\widehat{\mathbf{Z}}_t\| = \frac{d}{s}$. Moreover,

$$\mathbb{E}[\mathbf{Z}] = \sum_{t=1}^{s} \left[\frac{1}{m} \sum_{i=1}^{m} \frac{1}{sq_i} \mathbf{u}_i \mathbf{u}_i^\top \right] = s\, \frac{1}{s} \frac{1}{m} \mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U} = \frac{1}{m} \mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U}.$$

Combining Chernoff's inequality with the reduction argument described in [GN10], for any λ, and $\theta > 0$,

$$P\big(\lambda_{\max}(-\widehat{\mathbf{Z}}) \geq \lambda\big) \leq \mathrm{e}^{-\theta\lambda}\, \mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\widehat{\mathbf{Z}}))\big)\Big] \leq \mathrm{e}^{-\theta\lambda}\, \mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\mathbf{Z}))\big)\Big].$$

Using matrix Chernoff bound of [Tro12] applied to $-\mathbf{Z}_1, ..., -\mathbf{Z}_s$ with appropriate θ, we have

$$\mathrm{e}^{-\theta\lambda}\, \mathbb{E}\Big[\mathrm{tr}\big(\exp(\theta(-\mathbf{Z}))\big)\Big] \leq d\, \exp\Big(-\frac{s}{16d}\Big), \quad \text{for} \quad \lambda = \frac{1}{2}\, \lambda_{\max}\Big(-\frac{1}{m} \mathbf{U}^\top \mathbf{Q}_\pi \mathbf{U}\Big) \leq -\frac{1}{4}.$$

Thus, there is a constant C_1 such that for $s \geq C_1\, d \ln(d/\delta)$, w.p. at least $1 - \delta$ we have $\lambda_{\min}(\widehat{\mathbf{Z}}) \geq \frac{1}{4}$.

Proof of Lemma 4.4

We compute the moment generating function of the variable $R_t \sim \mathrm{Geom}(r_t)$, where $r_t = \frac{t-d}{t}$:

$$\mathbb{E}\big[\mathrm{e}^{\theta R_t}\big] = \frac{r_t}{1 - (1 - r_t)\mathrm{e}^{\theta}} = \frac{\frac{t-d}{t}}{1 - \frac{d}{t}\, \mathrm{e}^{\theta}} = \frac{t - d}{t - d\, \mathrm{e}^{\theta}}.$$

Setting $\theta = \frac{1}{2d}$, we observe that $d\mathrm{e}^{\theta} \leq d+1$, and so $\mathbb{E}[\mathrm{e}^{\theta R_t}] \leq \frac{t-d}{t-d-1}$. Letting $\mu = \mathbb{E}[R]$, for any $a > 1$ using Markov's inequality we have

$$P(R \geq a\mu) \leq \mathrm{e}^{-a\theta\mu}\,\mathbb{E}\big[\mathrm{e}^{\theta R}\big] \leq \mathrm{e}^{-a\theta\mu} \prod_{t=s+1}^{m} \frac{t-d}{t-d-1} = \mathrm{e}^{-a\theta\mu}\,\frac{m-d}{s-d}.$$

Note that using the bounds on the harmonic series we can estimate the mean:

$$\mu = d \sum_{t=s-d+1}^{m-d} \frac{1}{t} \geq d\,(\ln(m-d) - \ln(s-d) - 1) = d\,\ln\Big(\frac{m-d}{s-d}\Big) - d,$$

$$\text{so} \quad \mathrm{e}^{-a\theta\mu} \leq \mathrm{e}^{a/2}\,\exp\bigg(-\frac{a}{2}\ln\Big(\frac{m-d}{s-d}\Big)\bigg) = \mathrm{e}^{a/2}\Big(\frac{m-d}{s-d}\Big)^{-a/2}.$$

Putting the two inequalities together we obtain the desired tail bound.

4.5 Efficient algorithms for leveraged volume sampling

Rescaling volume sampling with leverage scores leads to a simple and efficient algorithm called *determinantal rejection sampling*. In this section, we present this algorithm and then construct an accelerated variant which for $s = O(d^2)$ runs in time $\widetilde{O}(nd + d^4)$.

4.5.1 Determinantal rejection sampling

Consider the following procedure: Repeatedly sample $t = O(d^2)$ indices $\pi_1, \ldots, \pi_t$ i.i.d. from $q = (\frac{l_1}{d}, \ldots, \frac{l_n}{d})$, and accept the sample with probability proportional to its volume ratio, i.e. $\frac{\det(\frac{1}{t}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{X}^\top\mathbf{X})}$. Having obtained a sample, we reduce its size further via reverse iterative sampling as described in Chapter 3 (denoted here as "VolumeSample$(\cdot,\cdot)$"). We show next that this procedure not only returns a q-rescaled volume sample, but also

exploiting the fact that q is proportional to the leverage scores, it requires (surprisingly) only a constant number of iterations of rejection sampling with high probability.

Lemma 4.5. *Let $\mathbf{X} \in \mathbb{R}^{n\times d}$ be full rank and let $\mathbf{A} \in \mathbb{R}^{d\times d}$ be a positive definite matrix. If $\pi_1, \ldots, \pi_t$ are sampled i.i.d. $\sim (\hat{l}_1, \ldots, \hat{l}_n)$, where $\hat{l}_i = \mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i$, then*

$$\det\Big(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X}\Big) \leq \det(\mathbf{A}) \qquad \textit{where} \quad \widetilde{\mathbf{Q}}_\pi = \sum_{j=1}^{t} \frac{d}{\hat{l}_{\pi_j}} \mathbf{e}_{\pi_j}\mathbf{e}_{\pi_j}^\top,$$

$$\textit{and} \quad \mathbb{E}\bigg[\frac{\det(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X})}{\det(\mathbf{A})}\bigg] \geq \Big(1 - \frac{d^2}{t}\Big)\frac{\det(\mathbf{X}^\top \mathbf{X}\,\mathbf{A}^{-1})}{(\frac{1}{d}\mathrm{tr}(\mathbf{X}^\top \mathbf{X}\,\mathbf{A}^{-1}))^d}.$$

Proof. We use the geometric-arithmetic mean inequality for the eigenvalues of matrix $\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X}\mathbf{A}^{-1}$:

$$\begin{aligned}\frac{\det(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X})}{\det(\mathbf{A})} &= \det\Big(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X}\mathbf{A}^{-1}\Big) \leq \bigg(\frac{1}{d}\mathrm{tr}\Big(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X}\mathbf{A}^{-1}\Big)\bigg)^d \\ &= \Big(\frac{1}{dt}\mathrm{tr}\big(\widetilde{\mathbf{Q}}_\pi \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top\big)\Big)^d = \Big(\frac{1}{dt}\sum_{i=1}^{t}\frac{d}{\hat{l}_i}\mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i\Big)^d = 1.\end{aligned}$$

Next, we use Proposition 4.1 to compute the expected value:

$$\begin{aligned}\mathbb{E}\bigg[\frac{\det(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X})}{\det(\mathbf{A})}\bigg] &= \sum_{\pi\in[n]^t}\Big(\prod_{i=1}^{t} q_{\pi_i}\Big)\frac{\det(\frac{1}{t}\mathbf{X}^\top \widetilde{\mathbf{Q}}_\pi \mathbf{X})}{\det(\mathbf{A})} \\ &= \Big(\frac{d}{\sum_{i=1}^n \hat{l}_i}\Big)^d \frac{1}{t^d\,\det(\mathbf{A})}\sum_{\pi\in[n]^t}\Big(\prod_{i=1}^{t} q_{\pi_i}\Big)\det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \\ &= \frac{t(t-1)\ldots(t-d+1)\det(\mathbf{X}^\top \mathbf{X})}{(\frac{1}{d}\sum_{i=1}^n \mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i)^d\, t^d\,\det(\mathbf{A})} \\ &\geq \Big(1-\frac{d}{t}\Big)^d \frac{\det(\mathbf{X}^\top \mathbf{X}\,\mathbf{A}^{-1})}{(\frac{1}{d}\mathrm{tr}(\mathbf{X}^\top \mathbf{X}\,\mathbf{A}^{-1}))^d}.\end{aligned}$$

Applying Bernoulli's inequality to the above expression concludes the proof. ∎

Note that the above lemma uses a slightly different rescaling matrix $\widetilde{\mathbf{Q}}_\pi$ than was used in the definition of q-rescaled volume sampling, however the difference is only

by a constant factor. Moreover, for the special case of $\mathbf{A} = \mathbf{X}^\top\mathbf{X}$, when $\hat{l}_i$ are the exact leverage scores of $\mathbf{X}$, then $\widetilde{\mathbf{Q}}_\pi = \mathbf{Q}_\pi$, and we easily obtain a runtime guarantee presented in the next theorem.

Algorithm 4.3:

Determinantal rejection sampling

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}, q = (\frac{l_1}{d}, \dots, \frac{l_n}{d}), s \geq d$

$t \leftarrow \max\{s,\, 4d^2\}$

repeat

  Sample $\pi_1, \dots, \pi_t$ i.i.d. $\sim (q_1, \dots, q_n)$

  Sample $\mathit{Accept} \sim \text{Bernoulli}\Big(\frac{\det(\frac{1}{t}\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X})}{\det(\mathbf{X}^\top\mathbf{X})}\Big)$

until $\mathit{Accept} = \text{true}$

$S \leftarrow \text{VolumeSample}\big((\mathbf{Q}^{1/2}_{[1..n]}\mathbf{X})_\pi, s\big)$

return π_S

Algorithm 4.4:

Fast leveraged volume sampling

Input: $\mathbf{X} \in \mathbb{R}^{n\times d}$, $s \geq d$, $\epsilon \geq 0$

Compute $\mathbf{A} = (1 \pm \epsilon)\, \mathbf{X}^\top\mathbf{X}$ [AC09]

Compute $\tilde{l}_i = (1 \pm \frac{1}{2}) l_i \quad \forall_{i\in[n]}$ [DMIMW12]

$t \leftarrow \max\{s,\, 8d^2\}$

repeat

  $\pi \leftarrow$ empty sequence

  while $|\pi| < t$

    Sample $i \sim (\tilde{l}_1, \dots, \tilde{l}_n)$

    $a \sim \text{Bernoulli}\Big((1-\epsilon)\frac{\mathbf{x}_i^\top\mathbf{A}^{-1}\mathbf{x}_i}{2\tilde{l}_i}\Big)$

    if $a = \text{true}$, **then** $\pi \leftarrow [\pi, i]$

  end

  $\widetilde{\mathbf{Q}}_\pi \leftarrow \sum_{j=1}^t d\,(\mathbf{x}_{\pi_j}^\top\mathbf{A}^{-1}\mathbf{x}_{\pi_j})^{-1}\mathbf{e}_{\pi_j}\mathbf{e}_{\pi_j}^\top$

  Sample $\mathit{Acc} \sim \text{Bernoulli}\Big(\frac{\det(\frac{1}{t}\mathbf{X}^\top\widetilde{\mathbf{Q}}_\pi\mathbf{X})}{\det(\mathbf{A})}\Big)$

until $\mathit{Acc} = \text{true}$

$S \leftarrow \text{VolumeSample}\big((\widetilde{\mathbf{Q}}^{1/2}_{[1..n]}\mathbf{X})_\pi, s\big)$

return π_S

Theorem 4.6. *Given the leverage score distribution* $q = (\frac{l_1}{d}, \dots, \frac{l_n}{d})$ *and the determinant* $\det(\mathbf{X}^\top\mathbf{X})$ *for matrix* $\mathbf{X} \in \mathbb{R}^{n\times d}$*, determinantal rejection sampling (Algorithm 4.3)*

returns sequence π_S distributed according to q-rescaled volume sampling, and w.p. at least $1-\delta$ finishes in time $O((d^2+s)d^2\ln(\frac{1}{\delta}))$.

Proof. Lemma 4.5 shows that acceptance probability from line 5 is bounded by 1. Thus, sequence π is drawn according to q-rescaled volume sampling of size t. Now, the composition property of rescaled volume sampling (Proposition 4.2) implies correctness of the algorithm. We refer to Lemma 4.5 again, and the fact that $t \geq 4d^2$ (see line 2), to observe that the expected value of the acceptance probability is at least $\frac{3}{4}$. An easy application of Markov's inequality shows that at each trial there is at least a 50% chance of it being above $\frac{1}{2}$. So, the probability of at least r trials occurring is less than $(1-\frac{1}{4})^r$. Note that the computational cost of one trial is no more than the cost of SVD decomposition of matrix $\mathbf{X}^\top\mathbf{Q}_\pi\mathbf{X}$ (for computing the determinant), which is $O(td^2)$. The cost of reverse iterative sampling (line 7) is also $O(td^2)$ with high probability. Thus, the overall runtime is $O((d^2+s)d^2r)$, where $r \leq \ln(\frac{1}{\delta})/\ln(\frac{4}{3})$ w.p. at least $1-\delta$. ■

4.5.2 Faster algorithm via approximate leverage scores

In some settings, the primary computational cost of deploying leveraged volume sampling is the preprocessing cost of computing exact leverage scores for matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, which takes $O(nd^2)$. There is a large body of work dedicated to fast estimation of leverage scores (see, e.g., [DMIMW12, Mah11]), and in this section we examine how these approaches can be utilized to make leveraged volume sampling more efficient. The key challenge here is to show that the determinantal rejection sampling

step remains effective when distribution q consists of approximate leverage scores. Our strategy, which is described in the algorithm *fast leveraged volume sampling* (Algorithm 4.4), will be to compute an approximate covariance matrix $\mathbf{A} = (1 \pm \epsilon)\mathbf{X}^\top\mathbf{X}$, i.e. such that

$$(1-\epsilon)\,\mathbf{X}^\top\mathbf{X} \preceq \mathbf{A} \preceq (1+\epsilon)\,\mathbf{X}^\top\mathbf{X}, \tag{4.8}$$

and use it to compute the rescaling distribution $q_i \sim \mathbf{x}_i^\top\mathbf{A}^{-1}\mathbf{x}_i$. We can compute matrix $\mathbf{A}^{-1}$ efficiently in time $\widetilde{O}(nd + d^3)$ using a sketching technique called Fast Johnson-Lindenstraus Transform [AC09], as described in [DMIMW12]. However, the cost of computing the entire rescaling distribution is still $O(nd^2)$. Standard techniques circumvent this issue by performing a second matrix sketch. We cannot afford to do that while at the same time preserving the sufficient quality of leverage score estimates needed for leveraged volume sampling. Instead, we first compute weak estimates $\tilde{l}_i = (1 \pm \frac{1}{2})l_i$ in time $\widetilde{O}(nd + d^3)$ as in [DMIMW12], then use rejection sampling to sample from the more accurate leverage score distribution, and finally compute the correct rescaling coefficients just for the obtained sample. The below result uses Lemma 4.5, showing that for a sufficiently small ϵ, determinantal rejection sampling can still work efficiently, while reducing the preprocessing cost to $\widetilde{O}(nd + d^3)$.

Theorem 4.7. *Given a full rank matrix $\mathbf{X} \in \mathbb{R}^{n\times d}$, an integer $s \geq d$, and $\epsilon = \frac{1}{16d}$, conditions from lines 2 and 3 of Algorithm 4.4 are satisfied with high probability and in time $\widetilde{O}(nd + d^3)$, and when they are, the procedure returns sequence π_S distributed according to q-rescaled volume sampling, and takes $O((d^2+s)d^2(\ln(\frac{1}{\delta}))^2)$ time w.p. at*

least $1-\delta$.

Proof. To establish correctness of the algorithm, we note that sequence π produced in line 12 consists of i.i.d. samples from the distribution $q_i \sim \mathbf{x}_i^\top \mathbf{A}^{-1}\mathbf{x}_i$, via rejection sampling who's acceptance probability is lower bounded by a constant. Correctness now follows from Lemma 4.5 and Proposition 4.2. Note that having produced matrix $\mathbf{A}^{-1}$, computing a single leverage score estimate $\hat{l}_i$ takes $O(d^2)$. To obtain a single sequence π of length t, the algorithm w.p. $\geq 1-\delta$ only has to compute $O(t\ln(\frac{1}{\delta}))$ such estimates, which introduces an additional cost of $O(td^2\ln(\frac{1}{\delta}))$, not exceeding the cost of other dominant procedures in leveraged volume sampling (up to the logarithmic factor). It remains to show that determinantal rejection sampling remains efficient when $\mathbf{A} = (1\pm\epsilon)\mathbf{X}^\top\mathbf{X}$. Note that this guarantee on $\mathbf{A}$ implies that the matrix $\mathbf{X}^\top\mathbf{X}\mathbf{A}^{-1}$ is a spectral approximation of identity, so we can easily bound both its determinant and trace. Thus we can apply Lemma 4.5 through the following observation which uses (4.8):

$$\frac{\det((1+\epsilon)\mathbf{X}^\top\mathbf{X}\,\mathbf{A}^{-1})}{(\frac{1}{d}\mathrm{tr}((1-\epsilon)\mathbf{X}^\top\mathbf{X}\,\mathbf{A}^{-1}))^d} \geq \frac{\det(\mathbf{I})}{(\frac{1}{d}\mathrm{tr}(\mathbf{I}))^d} = 1$$
$$\Rightarrow \quad \frac{\det(\mathbf{X}^\top\mathbf{X}\,\mathbf{A}^{-1})}{(\frac{1}{d}\mathrm{tr}(\mathbf{X}^\top\mathbf{X}\,\mathbf{A}^{-1}))^d} \geq \left(\frac{1-\epsilon}{1+\epsilon}\right)^d \geq \left(1-\frac{2\epsilon}{1+\epsilon}\right)^d \geq \left(1-\frac{1}{8d}\right)^d \geq \frac{7}{8}.$$

Since Algorithm 4.4 ensures that $t \geq 8d^2$, we can combine the above bound with the expectation bound from Lemma 4.5, obtaining that the expected acceptance probability of line 15 is at least $\frac{7}{8}\cdot\frac{7}{8} \geq \frac{3}{4}$. Thus, same reasoning as in the proof of Theorem 4.6 shows that w.p. $\geq 1-\delta$ the number of determinantal rejections is $O(\ln(\frac{1}{\delta}))$. ∎

It is worth noting that as long as preprocessing successfully produces the desired estimates $\mathbf{A}$ and $\tilde{l}_1, \ldots, \tilde{l}_n$, *fast leveraged volume sampling* produces a valid q-rescaled volume sample (and not an approximation of one), so the least-squares estimators are still exactly unbiased. Moreover, Theorems 4.4 and 4.5 can be extended to the setting where q is constructed from approximate leverage scores, so our loss bounds also hold in this case.

4.6 Experiments

We present experiments comparing leveraged volume sampling to standard volume sampling and to leverage score sampling, in terms of the total square loss suffered by the subsampled least-squares estimator. The three estimators can be summarized as follows:

$$\begin{aligned}
&\textit{volume sampling:} && \mathbf{w}^*(S) = (\mathbf{X}_S)^+\mathbf{y}_S, && P(S) \sim \det(\mathbf{X}_S^\top \mathbf{X}_S), \quad S \in \binom{[n]}{s};\\
&\textit{leverage score sampling:} && \mathbf{w}^*(\pi) = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^+\mathbf{Q}_\pi^{1/2}\mathbf{y}, && P(\pi) = \prod_{i=1}^{s} \frac{l_{\pi_i}}{d}, \qquad \pi \in [n]^s;\\
&\textit{leveraged volume sampling:} && \mathbf{w}^*(\pi) = (\mathbf{Q}_\pi^{1/2}\mathbf{X})^+\mathbf{Q}_\pi^{1/2}\mathbf{y}, && P(\pi) \sim \det(\mathbf{X}^\top \mathbf{Q}_\pi \mathbf{X}) \prod_{i=1}^{s} \frac{l_{\pi_i}}{d}.
\end{aligned}$$

Both the volume sampling-based estimators are unbiased, however the leverage score sampling estimator is not. Recall that $\mathbf{Q}_\pi = \sum_{i=1}^{|\pi|} q_{\pi_i}^{-1} \mathbf{e}_{\pi_i}\mathbf{e}_{\pi_i}^\top$ is the selection and rescaling matrix as defined for q-rescaled volume sampling with $q_i = \frac{l_i}{d}$. For each estimator we plotted its average total loss, i.e., $\frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$, for a range of sample sizes s, contrasted with the loss of the best least-squares estimator $\mathbf{w}^*$ computed from all data.

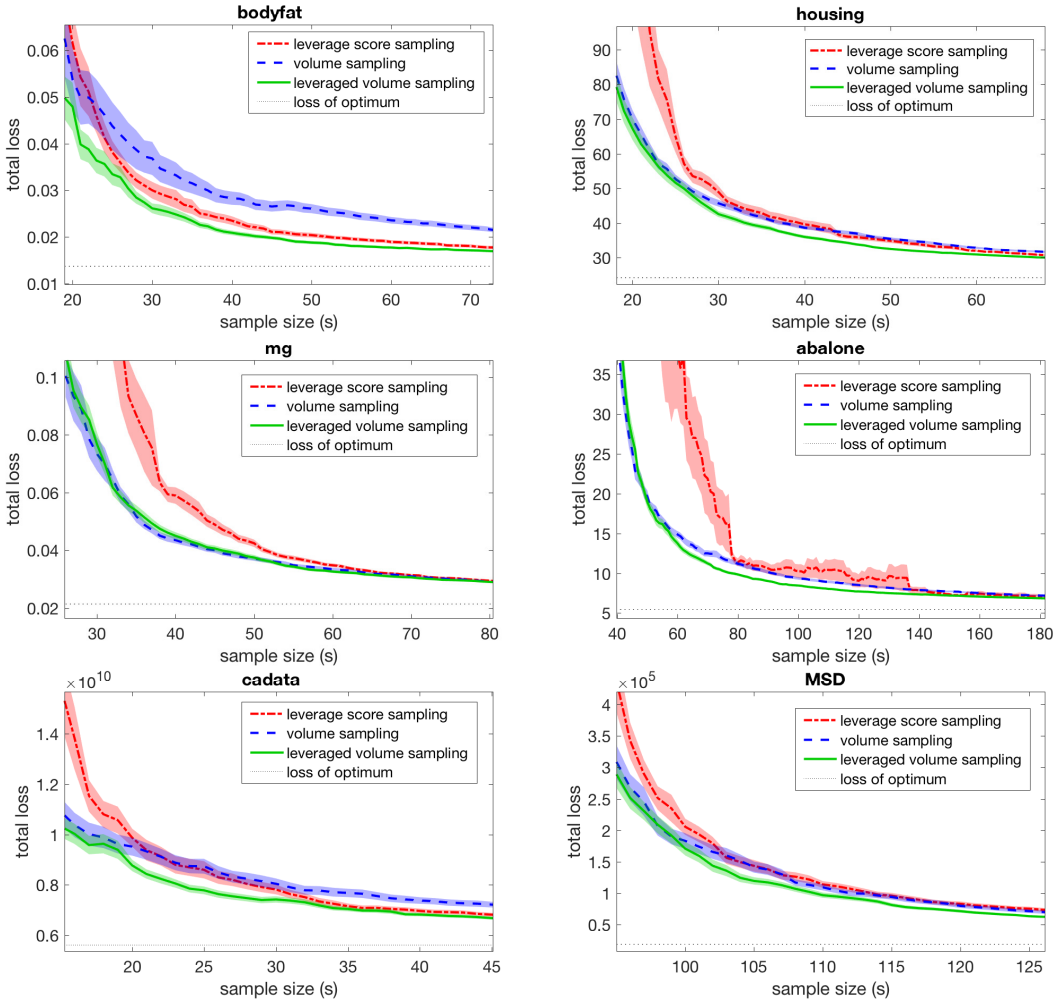

Figure 4.2: Comparison of loss of the subsampled estimator when using *leveraged volume sampling* vs using *leverage score sampling* and standard *volume sampling* on six datasets.

Dataset	Instances (n)	Features (d)
bodyfat_scale	252	14
housing_scale	506	13
mg	1,385	21
abalone	4,177	36
cpusmall_scale	8,192	12
cadata	20,640	8
MSD	463,715	90

Table 4.1: Libsvm regression datasets [CL11]. Suffix "_scale" indicates that a scaled version of the dataset was used, as explained in [CL11]. To increase dimensionality of *mg* and *abalone*, we expanded features to all degree 2 monomials, and removed redundant ones.

Plots shown in Figures 4.1 and 4.2 were averaged over 100 runs, with shaded area representing standard error of the mean. We used seven benchmark datasets from the libsvm repository [CL11] (six in this section and one in Section 4.1), whose dimensions are given in Table 4.1. The results confirm that leveraged volume sampling is as good or better than either of the baselines for any sample size s. We can see that in some of the examples standard volume sampling exhibits bad behavior for larger sample sizes, as suggested by the lower bound of Theorem 4.1 (especially noticeable on *bodyfat_scale* and *cpusmall_scale* datasets). On the other hand, leverage score sampling exhibits poor performance for small sample sizes due to the coupon collector problem, which is most noticeable for *abalone* dataset, where we can see a very sharp transition after which leverage score sampling becomes effective. Neither of the variants of volume sampling suffers from this issue.

4.7 Conclusion of the chapter

We developed a new variant of volume sampling which produces the first known unbiased subsampled least-squares estimator with strong multiplicative loss bounds. In the process, we proved a novel extension of the Cauchy-Binet formula, as well as other fundamental combinatorial equalities. Moreover, we proposed an efficient algorithm called determinantal rejection sampling, which is to our knowledge the first joint determinantal sampling procedure that (after an initial $O(nd^2)$ preprocessing step for computing leverage scores) produces its s samples in time $\widetilde{O}(d^2+s)d^2)$, independent of the data size n. The preprocessing can be reduced to $\widetilde{O}(nd+d^3)$ by rescaling with approximate leverage scores. Surprisingly the estimator stays unbiased and the loss bound still holds with only slightly revised constants.

In this chapter we focused on tail bounds. However we conjecture that expected bounds of the form $\mathbb{E}[L(\mathbf{w}^*(\pi))] \leq (1+\epsilon)L(\mathbf{w}^*)$ also hold for a variant of volume sampling of size $O(\frac{d}{\epsilon})$.

Chapter 5

Conclusions and future work

We proposed algorithms for selecting informative subsets of points from a dataset by exploring a deep connection between linear regression and volume sampling. Using a novel theoretical analysis and experimental evaluation we showed that volume sampling can be used as a subset selection technique for linear regression, offering strong statistical guarantees on the prediction error in terms of the number of responses needed. We also proposed algorithms which took this procedure from being virtually infeasible, to becoming an efficient and practical tool even for large datasets. A crucial technique which was introduced in Chapter 2 is reverse iterative sampling, a general approach for sampling from joint distributions over sets, which not only leads to efficient algorithms but also provides a theoretical framework for showing statistical guarantees. Furthermore, we developed several extensions of volume sampling. In Chapter 3 we proposed a regularized variant which offers statistical guarantees for ridge estimators when sampling fewer than dimension many responses. Then, in Chapter 4 we extended

the unbiased least squares estimator of standard volume sampling to a whole family of unbiased estimators via a rescaling technique which resulted in better loss bounds and a new type of sampling algorithm, called "determinantal rejection sampling", which further improved the efficiency of volume sampling. Many open questions and future directions still remain to be explored and we outline some of them below.

Best unbiased least squares estimator. The central question that lies at the core of this thesis can be stated as follows: Is there a way to produce an *unbiased least squares estimator* $\widehat{\mathbf{w}}$ from $O(d/\epsilon)$ responses such that its expected square loss on all data points is at most $1+\epsilon$ times the optimum (i.e., $\mathbb{E}[\widehat{\mathbf{w}}] = \mathbf{w}^*$ and $\mathbb{E}[L(\widehat{\mathbf{w}})] \leq (1+\epsilon)L(\mathbf{w}^*)$). In Chapter 2, we show an unbiased estimator which needs $O(d^2/\epsilon)$ responses to achieve the expected loss bound, whereas in Chapter 4 we construct an unbiased estimator from $O(d\log d + d/\epsilon)$ responses, which satisfies $L(\widehat{\mathbf{w}}) \leq (1+\epsilon)L(\mathbf{w}^*)$ with high probability (but not necessarily in expectation). On the other hand, a recent paper [CP17] showed that if we forego the unbiasedness requirement, then there is an estimator which achieves a $1+\epsilon$ loss bound with high probability from $O(d/\epsilon)$ responses, which suggest that a similar result may be possible for an unbiased estimator as well. However, our central question still remains open.

Reverse iterative sampling for structured data. The computational improvements in volume sampling developed in this work promise a new research frontier in subset selection, where sophisticated joint sampling techniques (such as reverse iterative sampling) can be deployed on a large scale in place of existing i.i.d. methods without

sacrificing the performance. Can the techniques proposed in this work be effectively applied in domains with structured data? For example, in graph theory related sampling approaches are used to find spectral sparsifiers [BSS09], or random spanning trees with desired properties [AB13]. Another important future direction is to explore the connection between volume sampling and submodular functions [JLB11], which offers potential extensions of the sampling techniques proposed in this thesis.

Stochastic Newton's method with unbiased steps. Recently, subsampling techniques for performing approximate Newton's method and other second-order optimization algorithms received a lot of attention (see [RM16, WRXM17]). The fundamental limitation of these approaches is that they can only show that subsampling improves computational efficiency, while not decreasing the convergence rate. On the other hand, subsampling for first-order methods (for example, going from gradient descent to stochastic gradient descent) is known to significantly improve the convergence rate of optimization (see [BB07]), both theoretically and empirically. The results obtained in this work suggest that similar results could be shown for stochastic second-order optimization. In particular, we showed that volume sampling can be used to produce an unbiased estimator of matrix pseudo-inverse from a small subset of examples. Computing the pseudo-inverse is an essential step in performing optimization using Newton's method, and the unbiasedness property allows us to use the subsampling noise to our advantage, similar as it was done for stochastic first-order methods [BB07]. Computing unbiased estimates also plays a crucial role in many ensemble methods, which are

useful in distributed settings. Thus, a promising research direction involves designing new stochastic second-order methods which will leverage the unbiased Newton steps to achieve better convergence rates in theory and efficiency in practice.

Bibliography

[AAHRJ87] T Ando, Roger A. Horn, and Charles R. Johnson. The singular values of a hadamard product: A basic inequality. 21:345–365, 12 1987.

[AB13] Haim Avron and Christos Boutsidis. Faster subset selection for matrices and applications. *SIAM Journal on Matrix Analysis and Applications*, 34(4):1464–1499, 2013.

[AC09] Nir Ailon and Bernard Chazelle. The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM Journal on computing*, 39(1):302–322, 2009.

[AD11] Anne Auger and Benjamin Doerr. *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2011.

[AM15] Ahmed El Alaoui and Michael W. Mahoney. Fast randomized kernel ridge regression with statistical guarantees. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 775–783, Montreal, Canada, December 2015.

[AZLSW17] Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal design of experiments via regret minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 126–135, Sydney, Australia, August 2017.

[BB07] Leon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. In *Proceedings of the 20th International Conference on Neural Information Processing Systems*, NIPS'07, pages 161–168, USA, 2007. Curran Associates Inc.

[BDM13] Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal coresets for least-squares regression. *IEEE Trans. Information Theory*, 59(10):6880–6892, 2013.

[BI92] Adi Ben-Israel. A volume associated with m x n matrices. *Linear Algebra and its Applications*, 167(Supplement C):87 – 111, 1992.

[BSS09] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. In *Proceedings of the 41st annual ACM symposium on Theory of computing*, STOC '09, pages 255–262, New York, NY, USA, 2009. ACM.

[BSS12] Joshua Batson, Daniel A Spielman, and Nikhil Srivastava. Twice-

ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

[BTT90] Aharon Ben-Tal and Marc Teboulle. A geometric property of the least squares solution of linear equations. *Linear Algebra and its Applications*, 139:165 – 170, 1990.

[CBL06] Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

[CDKV16] L Elisa Celis, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. How to be fair and diverse? *CoRR*, abs/1610.07183, 2016.

[CKS+18] L Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K Vishnoi. Fair and diverse dpp-based data summarization. *CoRR*, abs/1802.04023, 2018.

[CL11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[CP17] Xue Chen and Eric Price. Condition number-free query and active learning of linear families. *CoRR*, abs/1711.10051, 2017.

[CW13] Kenneth L. Clarkson and David P. Woodruff. Low rank approximation and regression in input sparsity time. In *Proceedings of the Forty-fifth*

Annual ACM Symposium on Theory of Computing, STOC '13, pages 81–90, New York, NY, USA, 2013. ACM.

[DMIMW12] Petros Drineas, Malik Magdon-Ismail, Michael W. Mahoney, and David P. Woodruff. Fast approximation of matrix coherence and statistical leverage. *J. Mach. Learn. Res.*, 13(1):3475–3506, December 2012.

[DMM06] Petros Drineas, Michael W Mahoney, and S Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*, pages 1127–1136. Society for Industrial and Applied Mathematics, 2006.

[DMM08] Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30(2):844–881, September 2008.

[DR10] Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *Proceedings of the 2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, pages 329–338, Las Vegas, USA, October 2010.

[DRVW06] Amit Deshpande, Luis Rademacher, Santosh Vempala, and Grant Wang. Matrix approximation and projective clustering via volume sampling. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, pages 1117–1126, Miami, FL, USA, January 2006.

[DW17] Michał Dereziński and Manfred K. Warmuth. Unbiased estimates for linear regression via volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 3087–3096, Long Beach, CA, USA, December 2017.

[DW18a] Michał Dereziński and Manfred K. Warmuth. Reverse iterative volume sampling for linear regression. *CoRR*, abs/1806.01969, June 2018.

[DW18b] Michał Dereziński and Manfred K. Warmuth. Subsampling for ridge regression via regularized volume sampling. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, 2018.

[DWH18] Michał Dereziński, Manfred K. Warmuth, and Daniel Hsu. Leveraged volume sampling for linear regression. *CoRR*, abs/1802.06749, 2018.

[Fed72] Valerii V Fedorov. *Theory of optimal experiments*. Probability and mathematical statistics. Academic Press, New York, NY, USA, 1972.

[GN10] David Gross and Vincent Nesme. Note on sampling without replacing from a finite collection of matrices. *arXiv preprint arXiv:1001.2738*, 2010.

[GPK16] Mike Gartrell, Ulrich Paquet, and Noam Koenigstein. Bayesian low-rank determinantal point processes. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 349–356, Boston, MA, USA, September 2016.

[GS12] Venkatesan Guruswami and Ali K. Sinop. Optimal column-based low-rank matrix reconstruction. In *Proceedings of the Twenty-third Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1207–1214, Kyoto, Japan, January 2012.

[HO14] Nicholas JA Harvey and Neil Olver. Pipage rounding, pessimistic estimators and matrix concentration. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 926–945. SIAM, 2014.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[Hol01] Lars Holst. Extreme value distributions for random coupon collector and birthday problems. *Extremes*, 4(2):129–145, 2001.

[Jan18] Svante Janson. Tail bounds for sums of geometric and exponential variables. *Statistics and Probability Letters*, 135:1 – 6, 2018.

[JLB11] Stefanie Jegelka, Hui Lin, and Jeff Bilmes. On fast approximate submodular minimization. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, pages 460–468, USA, 2011. Curran Associates Inc.

[Kan13] Byungkon Kang. Fast determinantal point process sampling with applica-

tion to clustering. In *Proceedings of the 26th International Conference on Neural Information Processing Systems*, NIPS'13, pages 2319–2327, USA, 2013.

[KT10] Alex Kulesza and Ben Taskar. Structured determinantal point processes. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 1171–1179. Curran Associates, Inc., 2010.

[KT11] Alex Kulesza and Ben Taskar. k-DPPs: Fixed-Size Determinantal Point Processes. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1193–1200, Bellevue, WA, USA, June 2011.

[KT12] Alex Kulesza and Ben Taskar. *Determinantal Point Processes for Machine Learning*. Now Publishers Inc., Hanover, MA, USA, 2012.

[LJS17] Chengtao Li, Stefanie Jegelka, and Suvrit Sra. Polynomial time algorithms for dual volume sampling. In *Advances in Neural Information Processing Systems 30*, pages 5045–5054, Long Beach, CA, USA, December 2017.

[LS15] Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In *Foundations of Computer Science (FOCS), 2015 IEEE 56th Annual Symposium on*, pages 250–269. IEEE, 2015.

[LWG+07] Jun Liao, Manfred K. Warmuth, Sridhar Govindarajan, Jon E. Ness, Rebecca P. Wang, Claes Gustafsson, and Jeremy Minshull. Engineering pro-

teinase k using machine learning and synthetic genes. *BMC Biotechnology*, 7(1):16, Mar 2007.

[Mah11] Michael W. Mahoney. Randomized algorithms for matrices and data. *Found. Trends Mach. Learn.*, 3(2):123–224, February 2011.

[MN99] Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, second edition, 1999.

[NST18] Aleksandar Nikolov, Mohit Singh, and Uthaipon Tao Tantipongpipat. Proportional volume sampling and approximation algorithms for a-optimal design. *CoRR*, abs/1802.08318, 2018.

[PP12] Kaare B. Petersen and Michael S. Pedersen. The matrix cookbook, November 2012. Version 20121115.

[PP14] Robin Pemantle and Yuval Peres. Concentration of lipschitz functionals of determinantal and other strong rayleigh measures. *Combinatorics, Probability and Computing*, 23(1):140–160, 2014.

[RM16] F. Roosta-Khorasani and M. W. Mahoney. Sub-Sampled Newton Methods I: Globally Convergent Algorithms. *ArXiv e-prints*, January 2016.

[Sar06] Tamas Sarlos. Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium*

on Foundations of Computer Science, FOCS '06, pages 143–152, Washington, DC, USA, 2006. IEEE Computer Society.

[SN09] Masashi Sugiyama and Shinichi Nakajima. Pool-based active learning in approximate linear regression. *Mach. Learn.*, 75(3):249–274, June 2009.

[Tro12] Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, Aug 2012.

[Woo14] David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

[WRXM17] Shusen Wang, Farbod Roosta-Khorasani, Peng Xu, and Michael W. Mahoney. GIANT: globally improved approximate newton method for distributed optimization. *CoRR*, abs/1709.03528, 2017.

[ZKM17] Cheng Zhang, Hedvig Kjellström, and Stephan Mandt. Determinantal point processes for mini-batch diversification. In *33rd Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, 11 August 2017 through 15 August 2017*. AUAI Press Corvallis, 2017.