A Geometric Theory of Outliers and Perturbation

by

John D. Dunagan

Bachelor of Science in Mathematics with Computer Science
Massachusetts Institute of Technology, 1998

Submitted to the Department of Mathematics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2002

Author . . . .
Department of Mathematics
May 3, 2002

Certified by . . . .
Santosh Vempala
Assistant Professor
Thesis Supervisor

Accepted by . . . .
Chair, Applied Mathematics

Accepted by . . . .
Tomasz S. Mrowka
Chairman, Department Committee on Graduate Students

A Geometric Theory of Outliers and Perturbation

by

John D. Dunagan

Submitted to the Department of Mathematics
on May 3, 2002, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

We develop a new understanding of outliers and the behavior of linear programs under perturbation. Outliers are ubiquitous in scientific theory and practice. We analyze a simple algorithm for removal of outliers from a high-dimensional data set and show the algorithm to be asymptotically good. We extend this result to distributions that we can access only by sampling, and also to the optimization version of the problem. Our results cover both the discrete and continuous cases. This is joint work with Santosh Vempala.

The complexity of solving linear programs has interested researchers for half a century now. We show that an arbitrary linear program subject to a small random relative perturbation has good condition number with high probability, and hence is easy to solve. This is joint work with Avrim Blum, Daniel Spielman, and Shang-Hua Teng. This result forms part of the *smoothed analysis* project initiated by Spielman and Teng to better explain mathematically the observed performance of algorithms.

Thesis Supervisor: Santosh Vempala
Title: Assistant Professor

Acknowledgments

The greatest source of contentment and purpose in my life is my wife, Heather. I also thank my family for the nurturing they have provided me over many years and continue to provide.

My mathematical education has long been in good hands. Before my teenage years, my parents had the foresight to suggest that some of their college math books would make good reading. The three summers I spent in the TIP program at Duke University contributed uniquely to both my social and mathematical growth. The middle school and high school math faculty at Greenhill School were both kind and encouraging. I particularly appreciate the year Dr. Barbara Currier spent teaching a class of first three students, then two, then one (me).

At MIT, site of both my undergraduate and graduate studies, my cup overflowed with opportunity. The faculty here have spent many hours giving me advice both mathematical and not. I am indebted to the advising I received from my earliest days at the hands of Hartley Rogers and Tom Leighton, the opportunity to work several summers under Bonnie Berger, the wonderful courses taught by Mike Sipser, Dan Spielman and David Karger, and the fruitful collaboration with Dan Spielman, Shang-Hua Teng, and Avrim Blum.

The most significant influence on my mathematical development has been my advisor, Santosh Vempala. From him I have not just learned material – I think about math differently because of our many hours together, and the change has been wholely for the good. The opportunity to work with him has been a blessing.

Contents

1 Introduction **11**

1.1 Outliers 11

1.1.1 Outliers in High-Dimensional Space 12

1.1.2 Organization of Outliers Results 14

1.2 Perturbations to Linear Programs 15

1.2.1 The Model of Perturbation 16

1.2.2 Renegar's Condition Number for Linear Programming 16

1.2.3 Algorithms using the Condition Number 19

1.2.4 Organization of Perturbation Results 19

2 Outliers **21**

2.1 Algorithms for Outlier Removal 21

2.2 Outlier Removal over Arbitrary Support 22

2.3 Outlier Removal over Discrete Support 27

2.4 Efficiency 35

2.4.1 Point sets 35

2.4.2 Arbitrary distributions 37

2.5 A Matching Lower Bound 42

2.6 An Approximation Algorithm 45

2.7 Standard Deviations from the Mean 46

2.8 An Implementation 50

2.9 Miscellanea 50

2.9.1 Interpolating Between Discrete and Arbitrary Support 51

2.9.2 Discrete Support: Rationals versus Integers 51

2.9.3 Some Further Thoughts on Optimization 51

2.9.4 Robust Statistics 52

2.10 Some Properties of Matrices 54

3 Perturbations **57**

3.1 Notation, Definitions, Main Result 57

3.1.1 Definition of Condition Number for Linear Programming 58

3.1.2 Smoothed Analysis of the Condition Number 59
3.2 Primal Condition Number 60
3.2.1 Primal number, feasible case 62
3.2.2 Primal number, infeasible case 64
3.2.3 Primal number, both cases 69
3.3 Dual Condition Number 70
3.4 Combining the Primal and Dual Analyses 76
3.5 Future Avenues of Investigation 77
3.6 The Perceptron Algorithm 78
3.6.1 Algorithm Definition and Analysis 78
3.6.2 The Input Mapping 79
3.6.3 Wiggle Room 80
3.6.4 Smoothed Analysis of the Perceptron Algorithm 80
3.7 Alternative Models of Perturbation 82
3.7.1 The Original Spielman-Teng Model 82
3.7.2 Zero-Preserving Perturbations 82
3.7.3 Non-Gaussian Perturbations 83
3.8 Technical Matters 84
3.8.1 A Bound on the Sum of Gaussian Random Variables 84
3.8.2 An Application of the Brunn-Minkowski Theory 85
3.8.3 Small Boundaries are Easily Missed 87

List of Figures

1-1 Defining Outliers 13

2-1 Lower Bound Constructions 43
2-2 Weaker Support Counterexample 51

3-1 Worst case $\boldsymbol{K}$ for lemma 29. 86

Chapter 1

Introduction

In this thesis, we study outliers and linear programs from a geometric setting. Outliers and linear programs are fundamental to machine learning and combinatorial optimization, two fields connected in a vibrant and ongoing dialogue. We develop similar geometric and probabilistic techniques to illuminate the two subjects.

The theory of outliers presented here was jointly developed with my thesis advisor, Santosh Vempala. Much of the theory originally appeared in a conference paper and subsequent journal version[DV 01]. The theory of perturbations to linear programs was developed over the course of two papers, one with Avrim Blum[BD 02], and one with Daniel Spielman and Shang-Hua Teng[DST 02]. However, this work on perturbations also benefitted from the comments of my advisor.

1.1 Outliers

Applied mathematicians often want to state that some property is typical of a data set. We commonly label as *outliers* the data points where the given property does not hold. In a machine learning context, outliers can dramatically slow the convergence of a learning algorithm, or even cause it to converge to a suboptimal hypothesis. In a practical setting, our understanding of a phenomenon might be enhanced by considering the outlier-free subset of the data, or the outliers themselves might constitute the phenomena of interest. In the classic setting of a scientist collecting data from an experiment that is sometimes contaminated by an infrequent external process, an outlier-free subset of the data is the desired object. The study of this external process would be the study of the outliers themselves. In the field of robust statistics, the goal is to construct statistics of a data set that will not be unduly influenced by the presence of a small contaminating process. Often, robust statistics are weighted averages of the examples, where outlying examples are given less weight.

To discuss the problem of finding outliers, we need a precise definition of an outlier. In the case of one-dimensional data, it is common to call something an outlier if it is far away from most of the data, where the measure of distance is normalized by some measure of the scatter of the data set. In particular, we define a point x to be a γ^2-outlier if x is more than γ standard deviations from the mean of the data set.

A natural algorithm for identifying and removing outliers is to fix γ, look for every γ^2-outlier, and remove those points. However, the remaining data set might still have outliers. The mean and standard deviation of the remaining data set may have changed, and points that were not outliers with respect to the original data set may now be outliers with respect to the remaining data set.

If we had instead proposed to define an outlier in terms of the standard deviation of the initial data set even after removing some points, the definition would have serious limitations. Most pressingly, a single outlier might "mask" the presence of other outliers. To see this, suppose we have one point a small distance away from the origin, a second point very far away, and many points right at the origin. Then the second point would be correctly identified as an outlier, but the first point might not be so identified. It is straightforward to construct a model for noise in our input data that leads to this problem. A second drawback to the alternative definition is that it does not support an application to learning theory that we will discuss later in this thesis. A third drawback from our viewpoint is that the relevant mathematics for the alternative definition are already well understood.

The definition we have adopted makes no mention of a hypothesis that can vary, even though we initially spoke of outliers as being points which fail some hypothesis. This invariance to the particular hypothesis that we will eventually learn is a desirable property when it is possible; it allows us to separate the tasks of learning and outlier removal. It will turn out that our definition of an outlier aids in the learning of linear separators, the most important hypothesis class in machine learning, in a manner that is invariant to the particular hypothesis found at the end.

An obvious next candidate algorithm is to apply the first natural algorithm iteratively, identifying and removing outliers repeatedly until we are done. When this iterative algorithm terminates, it clearly results in a γ^2-outlier free subset of the data. It just remains to bound the amount of the data set thrown away by the algorithm.

In this thesis we show that a natural generalization of this iterative algorithm is asymptotically optimal for the n-dimensional outlier removal problem. Before proceeding to discuss the algorithm, we discuss the n-dimensional problem in more detail.

1.1.1 Outliers in High-Dimensional Space

We will assume throughout that the data consists of points (or a distribution) in n-dimensional Euclidean space, hereafter abbreviated by $\mathcal{R}^n$. In figure 1-1, the top picture depicts the definition of a γ^2-outlier that we adopted for one-dimensional data: the data points are the solid circles, and the mean, along with the mean plus or minus one standard deviation, are the hash marks. The leftmost point is 1.86 standard deviations away from the mean.

In some settings it may be desirable to consider standard deviations from the mean, and in other settings it may be desirable to consider squared distance from a fixed reference point. For most of our discussion of outliers, we will measure squared distance from a fixed reference point, which we take without loss of generality to be the origin. This will simplify the exposition, and it is in this form that we need the result in order to apply it to the learning problem of [BFKV 99]. In section 2.7, we will translate our results back to the setting where we measure standard deviations from the mean of the data, and prove analogous theorems for this case.

The following generalization of the one-dimensional outlier definition to higher dimensions was used in [BFKV 99]. Let P be a set of points in $\mathcal{R}^n$.

Definition 1 (β-Outlier) *A point x in P is called a β-outlier if there exists a vector w such that the squared length of x along w is more than β times the average squared length* *of P along w, i.e. if*

$$(w^T x)^2 > \beta \mathbf{E}_{x \in P}[(w^T x)^2]$$

Note that $(w^T x)^2$ is the squared distance along w from the origin, and clearly $\beta = \gamma^2$ in the one-dimensional case. In figure 1-1, the bottom two pictures show how different points may be the furthest outliers for different choices of w. In each graph, the solid circles are the data points, the line is the direction w, and the hash marks along the line are the projections of the data points onto the line.

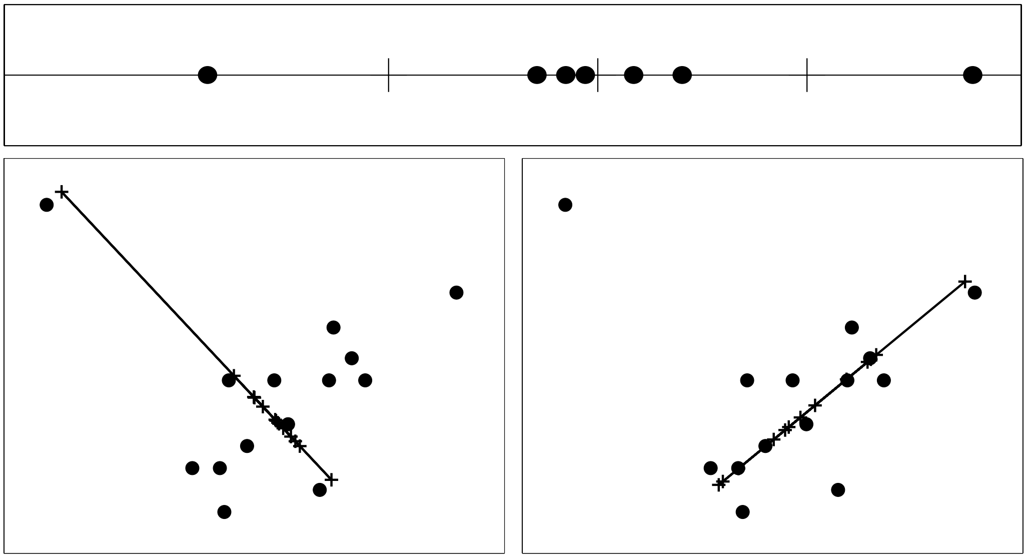

Figure 1-1: Defining Outliers

This definition of an outlier in $\mathcal{R}^n$ has a long history in statistics and machine learning. An equivalent definition using terminology from the field of machine learning is "a point is a γ^2-outlier if it has Mahalanobis distance greater than γ." A statistician might say "after normalizing by the covariance of the data, the point is more than γ away from the origin." The constructive procedure for identifying outliers in section 2.1 shows the equivalence of our definition to these two other definitions.

The first problem we address is the following: does there exist a small subset of (a point set) P whose removal ensures that the remaining set has no outliers? More precisely, what is the smallest β such that on removing a subset consisting of at most an ϵ fraction of P, the remaining set has no β-outliers (*with respect to the remaining set*)?

Our main result about outlier removal is that β can be made quite small as a function of ϵ. Let $\mathcal{Z}_b^n$ denote the set of n-dimensional b-bit integers, $\{1, \ldots, 2^b\}^n$.

Theorem 1 (Outlier Removal over Integer Support) [1] *Let μ be a probability distribution on $\mathcal{Z}_b^n$. Then for every $\epsilon > 0$, there exists S and*

$$\beta = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right)$$

[1] *An early version of this work[DV 01] claimed a slightly different version of theorem 1 with an insufficiently strong hypothesis.*

such that
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{(w^T x)^2 : x \in S\} \leq \beta \mathbf{E}[(w^T x)^2 : x \in S]$ *for all* $w \in \mathcal{R}^n$

The hypothesis of theorem 1 requires that the probability distribution under consideration be discretized in a certain way. Both the modern computers in widespread use today and the Turing machine model that has provided a basis for the mathematical study of computation represent numbers using finite precision, suggesting that discrete probability distributions are important both practically and theoretically. We prove a similar theorem for the continuous case, but handling the discrete case requires additional mathematical insights.

1.1.2 Organization of Outliers Results

The proof of theorem 1 (section 2.3) is constructive. Before proving theorem 1, we will prove a similar theorem about distributions with arbitrary support (theorem 2, section 2.2). Although the hypothesis on the support of the distribution in theorem 2 is much weaker, we need an additional assumption. The proofs of theorems 1 and 2 make use of the same principal idea.

In section 2.1, we describe (two variants of) an algorithm for outlier removal. The theorems can be proven using either variant. Although the theorems are not obvious, the algorithm is extremely simple. To convince the reader of this, we include a matlab implementation of the algorithm in section 2.8.

For a point set with m points ($m > n$) the algorithm runs in $O(m^2 n)$ time. In section 2.4 we show that the algorithm can also be used on an unknown distribution if it is allowed to draw random samples from the distribution. The number of samples required is $\tilde{O}(\frac{n^2 b}{\epsilon})$ and the running time is $\tilde{O}(\frac{b^2 n^5}{\epsilon \delta^4})$ for accuracy $(1 + \delta)$.

One variant of our algorithm is identical to the algorithm of [BFKV 99], the immediate inspiration for our work. The bound on β in theorems 1 and 2 improves on the previous best bound of $O(\frac{n^7 b}{\epsilon})$ given in [BFKV 99]. There it was used as a crucial component in the first polytime algorithm for learning linear threshold functions in the presence of random noise. Due to the high value of β, the bound on the running time of the learning algorithm, although polynomial, is a somewhat prohibitive $\tilde{O}(n^{28})$. In contrast, our theorem implies an improved bound of $\tilde{O}(n^5)$ for learning linear thresholds from arbitrary distributions in the presence of random noise. Further, our bound on β is asymptotically the best possible. This is shown in section 2.5 by an example where for any $\epsilon < \frac{1}{2}$, a bound on β better than $\Omega(\frac{n}{\epsilon}(b - \log \frac{1}{\epsilon}))$ is not possible.

Our main theorem gives an extremal bound on β. A natural follow-up question is the complexity of achieving the best possible β for any particular distribution. Given a distribution μ and a parameter ϵ, we want to find a subset of probability at most ϵ whose removal leaves an outlier-free set with the smallest possible β. We show this question to be NP-hard even in the one-dimensional case by a reduction to subset-sum. In section 2.6, we prove that our algorithm achieves a $(\frac{1}{1-\epsilon})$-approximation to the best possible β for any given ϵ.

In some cases, it may be desirable to translate the data set so that the origin coincides with the mean, rather than having a fixed origin. We prove the following corollary for standard deviations from the mean in section 2.7. Let μ be a probability distribution on Z_b^n. Then for any $\epsilon > 0$, there exists a $(1 - \epsilon)$ fraction of the distribution such that along every direction, no point is further away from the mean than $O(\sqrt{\frac{n}{\epsilon}(b + \log \frac{n}{\epsilon})})$ standard deviations in that direction. We also give a $(\frac{1-\epsilon}{1-3\epsilon})$-approximation algorithm for the corresponding optimization problem.

In section 2.9, we present some further observations about outlier removal, including a connection to robust statistics. In section 2.10, we prove some technical properties of matrices that are used elsewhere in the thesis.

1.2 Perturbations to Linear Programs

Linear programs are the earliest and greatest success of combinatorial optimization[Chv 83, pp7-9]. The field of combinatorial optimization sprang in large part from the empirical demonstration that linear programs could be used to guide the more efficient allocation of resources in a variety of settings[Chv 83, pp7-9]. We will briefly discuss the history of algorithms for solving linear programs, and then describe our contribution.

The analysis of algorithms for solving linear programs has been a subject of investigation for over fifty years. Dantzig both suggested the use of linear programs for optimization and proposed the first algorithm for solving linear programs, the simplex algorithm[Dan 51]. Many alternative pivot rules for use with the simplex algorithm have since been proposed, as well as alteranative rules for choosing a starting vertex. The simplex algorithm is combinatorial in nature. Shortly after Dantzig's initial work, Agmon proposed the perceptron algorithm for solving linear programs[Agm 54]. The perceptron algorithm can be shown to quickly solve a relaxation of the linear programming problem, but as the relaxation is tightened, the perceptron algorithm takes an increasing amount of time to converge. The perceptron algorithm uses the geometry of the linear program in a way that the simplex algorithm does not.

Around the same time that polynomial time gained wide currency within the theoretical computer science and applied math community as a measure of algorithmic efficiency, Klee and Minty showed that the simplex algorithm may require time exponential in the number of variables and constraints[KM 72]. A few years later, Khachiyan proved that the ellipsoid algorithm solved linear programs in polynomial time[Kha 79]. Karmarkar later developed the interior point method for solving linear programs[Kar 84], and this has led to a tremendous body of subsequent work.

Analysis of both the ellipsoid and interior point methods relied upon assumptions about the bit-size of the data representation, and a number of people in the community, such as Lenore Blum[Blu 90], suggested that geometric measures independent of the bit size could be used to bound the performance of interior point methods. Renegar initiated this line of work, translating the theory of interior point methods to the setting of numerical and functional analysis by introducing the concept of *condition number* of a linear program[Ren 94, Ren 95a, Ren 95b]. Since then, a number of other people have further developed the theory of condition numbers for linear programs, and developed algorithms that can be analyzed in terms of the condition number [CP 01, FE 00a, FE 00b, FE 01, FN 01, FV 99, FV 00, Ver 96].

All of the previously cited analysis has been of the worst-case behavior of algorithms for solving linear programs. Condition numbers represent a refinement of bit-size for the purpose of describing problem difficulty, but there is no a-priori bound on the condition number of a linear program given one as input.

A parallel line of inquiry arose to address the empirically observed fact that although the simplex algorithm has poor worst-case complexity, in practice it runs remarkably quickly, requiring a number of iterations nearly linear in the minimum of the number of constraints or dimensions. Towards this end, a number of authors investigated the expected running time of the simplex algorithm on a given distribution of linear programs[Adl 83, AKS 87, AM 85, Bor 77, Bor 80, Hai 83, Meg 86, Mur 80, Sma 82, Sma 83, Tod 86, Tod 91]. In one model of a random linear program, we generate the linear program by starting with an arbitrary linear program and then randomly flipping the direction of every constraint. In the other widely considered model, studied by Borgwardt, Smale, and Megiddo[Bor 77, Bor 80, Sma 82, Sma 83, Meg 86], each constraint is drawn from a spherically symmetric distribution.

1.2.1 The Model of Perturbation

In 2001, Spielman and Teng proposed a new model[ST 01]. In this new model, which they dubbed *smoothed analysis*, they consider a small random perturbation to an arbitrary linear program. Although similar to the model of Borgwardt, this model has an appealing feature that the previous models do not: by varying the size of the perturbation, we may interpolate between average-case and worst-case results. The other appealing feature of the smoothed analysis model is best explained by a change in belief within our field about the nature of reality.

Over the past few decades, the field of theoretical computer science has become acutely aware that typical problem instances bare little resemblance to the random instances commonly generated by our mathematically tractable distributions. The task of characterizing the distribution of problem instances encountered in practice is now recognized as a formidable one. A further discussion of this issue can be found in [ST 01], but we will try to summarize: typical instances do not look like random instances; they *might* look like an arbitrary instance subject to a small random perturbation. By considering a small random perturbation to an arbitrary instance, the smoothed analysis model seeks to better describe the observed performance of algorithms on real-world instances.

In their initial work, Spielman and Teng showed the simplex algorithm with shadow vertex pivot rule to have polynomial smoothed complexity. This was both a significant mathematical accomplishment, and a validation that the model, though significantly more restrictive than previous models of a random linear program, still admitted interesting average-case results.

In this thesis, we analyze a geometric quantity of the perturbed linear program. This geometric quantity is the *condition number*, and it can be used to bound the performance of the perceptron algorithm, the ellipsoid algorithm, and the many interior point methods.

We hope that this work both elucidates the smoothed analysis model and possibly explains the distribution of condition numbers encountered in practice. We do not claim that the model of a perturbed linear program considered here is the best possible description of the distribution of problem instances encountered in practice. However, we have great hope that this is a useful step in better describing the distribution of problem instances encountered in practice, and the analysis of algorithms on such distributions.

1.2.2 Renegar's Condition Number for Linear Programming

We begin by discussing condition numbers in general, and then Renegar's condition number for linear programming in particular. We then discuss smoothed analysis in some detail.

Condition numbers are ubiquitous in numerical analysis and scientific computing. For many computational tasks with matrices, the ratio of the maximum and minimum eigenvalues of the matrix is a good condition number. For other tasks, such as solving a discretized partial differential equation for given boundary conditions, a different condition number may be defined. A condition number typically has two uses,

1. to estimate the sensitivity of the problem's answer to error in the input, and
2. to bound the number of iterations required by an iterative method to achieve a given degree of accuracy.

Analysis of algorithms using condition numbers may be interpreted as a *parameterized* worst-case complexity analysis. For many iterative methods, the maximum number of iterations is bounded by some function of the condition number, although the actual number of iterations may be less. Thus condition number is a refinement of *input size* as a measure of problem difficulty. Additionally, the condition number is typically bounded by some function of the input size, where the input size includes both the number of input parameters and the bit size required to represent these parameters, and so condition number bounds typically imply worst-case complexity bounds in the standard Turing model of computation.

Another reason for the study of condition numbers is that:

Numerical analysis is the study of algorithms for the problems of continuous mathematics.[2]

For a continuous input domain, it may be unnatural to discretize the input in the problem definition. Condition numbers are well-defined for arbitrary real-valued inputs, where measuring the input size in bits may not be possible. The fields of numerical analysis and scientific computing consider such problems and inputs, and condition numbers have been a pervasive underpinning of research in these fields.

Renegar [Ren 94] introduced a condition number for linear programs. In this work, he suggested that the study of condition numbers for linear programming was a natural outgrowth of the central role iterative solvers, particularly interior point methods, had assumed in the study of algorithms for convex programming. A large body of further work, detailed below, has developed on bounding the number of iterations required to solve a given linear program as a function of the condition number. This analysis has included both new bounds on old methods and the development of new algorithms.

Our work addresses the question of "what are likely values for the condition number?" In particular, for a natural model of noise in the input data, we show that the condition number is likely to be low. This addresses a question outside the scope of previous work on the condition number for linear programs. The great body of work on how condition number influences running time is an extensive foundation, and we hope to build another layer underneath, on how noisy data leads to bounded condition number.

In [ST 01], Spielman and Teng showed that for an arbitrary linear program, a small random relative perturbation of that program is solved by the simplex algorithm (with the shadow vertex pivot rule) in polynomial time with very high probability. They also expressed the hope in [ST 01] that their result might explain the observed good performance of the simplex algorithm in practice: if your linear program is defined by a constraint matrix drawn from noisy data, it will probably be one that is easily solved by the simplex algorithm.

[2] Lloyd Trefethen, November 1992 SIAM News.

The smoothed complexity model seeks to interpolate between worst-case and average-case complexity analysis. By letting the size of the random perturbation to the data (i.e., the variance of the noise) become large, one obtains the traditional average-case complexity measure. By letting the size of the random perturbation go to zero, one obtains the traditional worst-case complexity measure. In between, one obtains new theoretical results that may also be practically meaningful.

The examples given above and the work in this chapter pertain to the smoothed analysis of algorithms for linear programming. We use a two-step approach:

1. Bound the running time of an algorithm in terms of a condition number.

2. Perform a smoothed analysis of this condition number.

Step 1. has already been done (see subsection 1.2.3). Our main theorem accomplishes step 2.

We do not wish to give the impression that smoothed analysis is only meaningful for linear programming, or even convex optimization problems. Recall that different problems (matrix inversion, solving a partial differential equation, etc.) have different condition numbers. Typically these condition numbers are defined to be (for any given problem instance) the maximum ratio of the magnitude of change in the output to the magnitude of change in the input. Many of these condition numbers have the property that the condition number is low if the smallest relative change to the input data necessary for the problem to be ill-posed is large. (Loosely speaking, a problem instance is ill-posed if an arbitrarily small further change to the input data may yield an arbitrarily large change in the answer. The linear programming condition number we consider here is defined to be the distance to ill-posedness, and can then be shown to bound the magnitude of change in the output due to change in the input.) For such condition numbers it may be the case that, from any initial instance, a small random perturbation to that instance is quite likely to yield a new instance that is not too close to ill-posedness. One exciting aspect of [ST 01] is that it shows that the simplex algorithm fits into this general framework. This thesis shows the same thing for the linear programming condition number. This phenomenon may be very common (the condition number for matrix inverion is addressed in [ST 02]).

To give a preview of our results, we state a rough version of our main theorem without constants or logarithmic factors.

Statement 1 (Smoothed Complexity of Renegar's Condition Number) *For an arbitrary linear program defined by an appropriately scaled n-by-d constraint matrix subject to a Gaussian perturbation of variance σ^2, with probability at least $1-\delta$ over the random perturbation, Renegar's condition number C satisfies*

$$C = \tilde{O}(\frac{n^2 d^{3/2}}{\sigma^2 \delta})$$

A precise version of this statement is theorem 5.

As an example of what kind of conclusion we derive on the overall performance of algorithms, we mention that a particular interior point method [FM 00] only requires $O(\sqrt{n+d}\ln(C/\epsilon))$ iterations to come within ϵ of the optimal solution, and each iteration requires only an approximate matrix inversion computable in $O((n+d)^{2.5})$ time. Thus the smoothed complexity is $O((n+d)^3 \ln(nd/(\sigma\epsilon)))$ for this particular method to come within ϵ of the optimal solution.

1.2.3 Algorithms using the Condition Number

Since Renegar's initial papers [Ren 94, Ren 95a, Ren 95b] on condition numbers for linear programs, there has been a large body of subsequent work. The running time of a number of algorithms for optimization has been analyzed in terms of their dependence on the condition number [FN 01, FV 00]. The notion of condition number has even inspired new algorithms for optimization [FE 00a, FE 00b]. Additionally, some variants of Renegar's original condition number have also been studied [FE 01].

In section 3.6, we describe a parameter known as the *wiggle room* of a linear program. It is well-known within the machine learning community that the running time of the perceptron algorithm may be bounded in terms of the wiggle room. In section 3.6, we show that the wiggle room is exactly the primal condition number by another name, and hence the running time of the perceptron algorithm for linear programming problems may similarly be bounded in the smoothed complexity model. The observation that wiggle room exactly corresponds to primal condition number has occurred to others[3] but we do not think it is widely known. The goal of section 3.6 is therefore both to publicize this connection, and to illustrate the use of a condition number in analyzing the running time of an algorithm for linear programming.

Our theorem on the smoothed complexity of the condition number implies a smoothed complexity of the perceptron alorithm that is polynomial, something that is not true under traditional worst-case analysis. The perceptron algorithm was only the second algorithm (after the simplex algorithm) for which an exponential (or worse) running time was shown to improve to a polynomial running time in the smoothed complexity model. Robert Freund pointed out that numerous other simple algorithms, such as [FE 00b], have a similar dependance on the condition number.

Part of the reason for the volume of work on condition number is that every linear programming formulation requires a separate condition number analysis. This point is made by [Ren 94, Ren 95a, Ren 95b, Ver 96, CP 01] in their work developing interior point methods that have good dependence on the condition number. In addition to bounding the time necessary to optimize, there has been work on quickly estimating the condition number [FV 99], a well-known question for the condition numbers of other problems. Also, the notion of condition number for linear programs has been extended to semi-definite programs [FN 01], but we will not elaborate further on this topic.

1.2.4 Organization of Perturbation Results

In section 1.2.2, we define Renegar's condition number for linear programming, a geometric quantity of a linear program, and we state the exact results we will prove. In section 3.2 and 3.3, we analyze the condition numbers of the primal and dual problems respectively. In section 3.4, we combine these analyses to characterize the smoothed complexity of Renegar's condition number. In section 3.5, we discuss some possible future avenues of investigation.

In section 3.6, we describe a classical machine learning algorithm, the perceptron algorithm, and show how it has complexity polynomial in the primal condition number. This result makes use of the well-known characterization of the running time of the perceptron algorithm in terms of the *wiggle room* of the linear program. The perceptron algorithm is

[3] Rob Freund, personal communication.

one of many algorithms that have a smaller cost per iteration than an interior point method, but whose running time depends polynomially on the condition number, rather than on the log of the condition number.

In section 3.7 we discuss several alternative models of perturbation. In section 3.8 we develop some technical results that the main body of the work uses as a black box.

Chapter 2

Outliers

The first question we address is that of detecting outliers. Since our definition of a γ^2-outlier involves all directions, it might not be obvious that this can be done in finite time. Even if we were only interested in a finite set of directions, it might not be obivous that this can eb done efficiently.

2.1 Algorithms for Outlier Removal

In order to detect outliers, we use a linear transformation. Let $M = \mathbf{E}[xx^T]$ where x is a sample drawn according to the probability distribution μ. If M is positive definite, there exists a matrix A such that $M = A^2$. Consider the transformed space $z = A^{-1}x$. This transformation preserves outliers: if z is a β-outlier in direction w in the transformed space, the corresponding $x = Az$ is a β-outlier in direction $w' = A^{-1}w$ in the untransformed space, and vice versa. The transformed distribution is in *isotropic* position [LKS 95], and we will refer to the transformation as *rounding.* Such transformations have previously been used in the design of algorithms to make geometric random walks more efficient [LKS 97]. If M does not have full rank, it is still positive semi-definite, and we instead round μ in the span of M. For those familiar with the definitions of Mahalanobis distance or normalizing by the covariance of the data set, this transformation shows the equivalence between our definition of an outlier and these two other definitions.

For an isotropic distribution, any point x that is an outlier for some direction w is also an outlier in the direction x. This follows from the fact that an isotropic distribution has $\mathbf{E}[(w^Tx)^2] = 1$ for every w such that $|w| = 1$, and that the projection of the point x on to a direction w is greatest when $w = x/|x|$. Thus, outlier identification is easy for isotropic distributions.

The first algorithm has the following simple form: while there are β-outliers, remove them; stop when there are no outliers. In the description below, μ is the given distribution and $\beta = \gamma^2$, where the exact value of β is specified in the proofs of theorems 1 and 2.

Algorithm 1 (Restriction to Ellipsoids):

1. Round μ. If there exists x such that $|x| > \gamma$, let $S = \{x : |x| \leq \gamma\}$. Retain only points in S.

2. Repeat until the condition is not met.

Algorithm 1 is identical to the outlier removal algorithm of [BFKV 99]. The following variant of the above algorithm will be significantly easier to analyze. Whereas in the previous algorithm we removed outliers in every direction in one step, in Algorithm 2 we only remove outliers in one direction per step.

Algorithm 2 (Restriction to Slabs):

1. Round μ. If there exists a unit vector w such that $\max\{(w^Tx)^2\} > \gamma^2$, let $S = \{x : (w^Tx)^2 \leq \gamma^2\}$. Retain only points in S.
2. Repeat until the condition is not met.

2.2 Outlier Removal over Arbitrary Support

We will prove the following theorem about outlier removal over a distribution with arbitrary support before proceeding to prove theorem 1. We refer to conditions (I, II) in the hypothesis of theorem 2 as the *full-dimensional* condition. In theorem 1 we will remove this condition, replacing it only by a condition on the support of the distribution.

Theorem 2 (Outlier Removal over Arbitrary Support) *Let μ be a probability distribution on $\mathcal{R}^n$ satisfying*

(I) $\forall$ unit vector $\hat{w}$, $\quad \int(\hat{w}^Tx)^2 d\mu \quad \leq \quad R^2$

(II) $\forall$ unit vector $\hat{w}$, $\quad \forall S : \mu(S) \geq 1-\bar{\epsilon}$, $\quad \int_S(\hat{w}^Tx)^2 d\mu \quad \geq \quad r^2$

Then for every ϵ such that $0 < \epsilon \leq \bar{\epsilon}$, there exists S and

$$\beta = O\left(\frac{n}{\epsilon}\ln\frac{R}{r}\right)$$

such that
(i) $\mu(S) \geq 1-\epsilon$
(ii) $\max\{(w^Tx)^2 : x \in S\} \leq \beta\mathbf{E}[(w^Tx)^2 : x \in S]$ for all $w \in \mathcal{R}^n$

To prove the theorem, we analyze the set S returned by either algorithm. This set S is clearly β-outlier free. It remains to show that we do not discard too much of the distribution. The main idea of the proof is to show that in every step the volume of an associated dual ellipsoid increases. By bounding the total growth of the dual ellipsoid volume over the course of the algorithm, we will deduce that no more than a certain fraction of the original probability mass is thrown out before the algorithm terminates.

Towards this end, we will need some definitions. For a matrix M such that $M = A^2$, define the ellipsoids $E(M)$ and $W(M)$ as

$$E(M) = \{x : |A^{-1}x| \leq 1\} \quad \text{and} \quad W(M) = \{x : |Ax| \leq 1\}.$$

We will refer to $E(M)$ and $W(M)$ as the primal inertial ellipsoid and the dual ellipsoid respectively. For any subset S of $\mathcal{R}^n$, we denote by M_S the matrix given by

$$M_S = \sum_{x\in S}\mu(x)xx^T = \mathbf{E}[xx^T : x \in S]\Pr[x \in S]$$

In other words, M_S is the M obtained after restricting μ to S (zeroing out points outside of S, not renormalizing the distribution). We denote this restricted probability distribution directly by μ_S. Throughout this chapter, μ_S will denote a restriction of μ to the subset of space S, never a new and unrelated distribution. The useful property attained by rounding with respect to μ_S (the restriction of the original distribution to S) is that

$$\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = 1$$

for every unit vector w, where the expectation and probability are with respect to x drawn from μ. We will actually prove theorem 2 with $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$ in place of $\mathbf{E}[(w^T x)^2]$. Note that this is a stronger statement than the original theorem. Let $x \in \mu_S$ denote $x \in S : \mu(x) > 0$, and let $span(\mu_S)$ denote the span of $\{x \in \mu_S\}$.

We will also need the following elementary facts about ellipsoids: the volume of a full-dimensional ellipsoid is given by the product of the axis lengths times the volume of the unit ball, which we will denote by $f(n)$. The ellipsoid $\{x : |A^{-1}x| \leq 1\}$ has axes given by the singular vectors of A. The axis lengths of $W(M)$ and $E(M)$ are given by the singular values of A^{-1} and A, and so they are reciprocals. It follows that $Vol(W(M))Vol(E(M)) = (f(n))^2$, a function solely of the dimension.

Lemma 1 relates the dual volume growth to the loss of probability mass, and lemma 2 upper bounds the total dual volume growth.

Lemma 1 (Restriction to a Slab) *Let γ be fixed, and let μ be a full-dimensional isotropic distribution. Suppose $\exists w, |w| = 1$ such that*

$$\max\{(w^T x)^2\} > \gamma^2 \mathbf{E}[(w^T x)^2]$$

Let $S = \{x : (w^T x)^2 \leq \gamma^2\}$ and $p = \Pr[x \notin S]$. Then

$$Vol(W(M_S)) \geq e^{p\gamma^2/2} Vol(W(M))$$

Proof: Let $a^2 = \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]$. Starting from the identity

$$\mathbf{E}[(w^T x)^2] = \mathbf{E}_{x \in S}[(w^T x)^2] \Pr[x \in S] + \mathbf{E}_{x \notin S}[(w^T x)^2] \Pr[x \notin S]$$

and using that $(w^T x)^2 \geq \gamma^2$ for all x not in S, we get that $1 \geq a^2 + \gamma^2 p$, which implies

$$a^2 \leq 1 - \gamma^2 p \leq e^{-\gamma^2 p}$$

We now construct a vector w' of length $1/a$ belonging to the dual ellipsoid of μ_S. Letting $w' = w/a$ suffices since w is a unit vector by assumption and

$$a^2 = \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = w^T M_S w$$

$$\Rightarrow \quad 1 = w'^T M_S w' \quad \Rightarrow \quad w' \in W(M_S)$$

We also show that every $v \in W(M)$ also belongs to $W(M_S)$. We have

$$M_S = M - \sum_{x \notin S} \mu(x) x x^T.$$

Hence,

$$v^T M_S v = v^T M v - \sum_{x \notin S} \mu(x) v^T x x^T v$$

$$= v^T M v - \sum_{x \notin S} \mu(x)(v^T x)^2 \leq v^T M v \leq 1$$

implying that $v \in W(M_S)$ (the last step is from the assumption that $v \in W(M)$). The length of a point on the boundary of an ellipsoid lower bounds the length of the longest axis. Since at least one axis of the dual ellipsoid has length $1/a$, and all the other axes have length at least 1, $Vol(W(M_S)) \geq (1/a)f(n)$ while $Vol(W(M)) = f(n)$, implying the dual volume grows by at least a factor of $e^{\gamma^2 p/2}$. This concludes the proof of lemma 1. □

Note that if we desire to apply the lemma to analyze the result of a later iteration of Algorithm 2, where μ_T goes to $\mu_{T \cap S}$, we simply replace the starting identity by

$$\mathbf{E}_{x \in T}[(w^T x)^2] \Pr[x \in T] = \mathbf{E}_{x \in T \cap S}[(w^T x)^2] \Pr[x \in T \cap S] + \mathbf{E}_{x \in T \setminus S}[(w^T x)^2] \Pr[x \in T \setminus S]$$

The analysis and conclusion remain the same.

Lemma 2 (Dual Volume Growth) *Let μ be a distribution satisfying*

(I) $\forall$ unit vector $\hat{w}$, $\quad \int (\hat{w}^T x)^2 d\mu \quad \leq \quad R^2$

(II) $\forall$ unit vector $\hat{w}$, $\quad \forall S : \mu(S) \geq 1 - \bar{\epsilon}$, $\quad \int_S (\hat{w}^T x)^2 d\mu \quad \geq \quad r^2$

For any S^*, let μ_{S^*} be the restriction of μ to S^*. Assume $\mu(S^*) \geq 1 - \bar{\epsilon}$. Then

$$Vol(W(M)) \geq \frac{f(n)}{R^n}$$

$$Vol(W(M_{S^*})) \leq \frac{f(n)}{r^n}$$

Proof: First we lower bound the initial dual volume, $Vol(W(M))$. Consider any vector v of length at most $1/R$. We have

$$v^T M v = \mathbf{E}[(v^T x)^2] = \int (v^T x)^2 d\mu \leq (v^2 R^2) \leq 1$$

so v belongs to the dual ellipsoid. Thus the dual ellipsoid initially has volume at least $f(n)/R^n$.

Next we upper bound $Vol(W(M_{S^*}))$. Consider any vector v of length more than $1/r$. Then

$$v^T M_{S^*} v = \int_{S^*} (v^T x)^2 d\mu \geq (v^2 r^2) > 1$$

Thus v is not in $W(M_{S^*})$, and thus the ultimate volume of the dual ellipsoid is no more than the volume of the sphere of radius $1/r$, yielding the claimed upper bound. □

In the proof of theorem 2 below, μ_{S^*} will be the final distribution resulting from application of either algorithm. Using lemmas 1 and 2, we prove that Algorithm 2 terminates with $S = S^*$ satisfying theorem 2.

Proof of Theorem 2: Let $\beta = 4\frac{n}{\epsilon}(\ln \frac{R}{r} + 1)$. Suppose that the algorithm terminates with subset S^* after having thrown out no more than ϵ' of the original probability mass. Then we have that for every w,

$$\max\{(w^T x)^2 : x \in S^*\} \leq \gamma^2 \mathbf{E}[(w^T x)^2 : x \in S^*] \Pr[x \in S^*]$$

We remind the reader again that normalizing μ_{S^*} so that it is a probability distribution on points from μ, rather than with points outside of S^* replaced by zeros, increases the right-hand side of this inequality by the factor $1/\mu(S^*)$, but does not increase the left-hand side. Thus the inequality will still be true even if we normalize μ_{S^*}. We thus achieve a β-outlier free subset with

$$\beta = \gamma^2 = 4\frac{n}{\epsilon}(\ln\frac{R}{r} + 1)$$

It now remains to show that $\epsilon' \leq \epsilon$, i.e. that we do not throw out more of the probability mass than claimed. Suppose that during the i^{th} iteration of the algorithm through step 1, a p_i fraction of the original points are thrown out. Then the total amount thrown out is $\sum p_i$. By lemma 1, the total amount of dual volume increase is $\prod_i e^{p_i\gamma^2/2} = e^{\frac{\gamma^2}{2}\sum p_i}$. Comparing this to our bound on the total increase in the dual volume from lemma 2 yields

$$e^{\frac{\gamma^2}{2}\sum p_i} \leq (\frac{R}{r})^n = e^{n\ln\frac{R}{r}}$$

$$\Rightarrow \frac{1}{2}\gamma^2\epsilon' = \frac{1}{2}(4\frac{n}{\epsilon}\ln\frac{R}{r})\epsilon' \leq n\ln\frac{R}{r}$$

$$\Rightarrow \epsilon' \leq \epsilon/2$$

The one remaining catch is showing that $\epsilon' \leq \bar{\epsilon}$, since we relied on this in applying lemma 2 above. By slight overloading of notation, we let ϵ' denote the cumulative probability mass that has been removed at any point during the algorithm. Suppose for the purpose of establishing a contradiction that in iteration j, $\epsilon' \leq \bar{\epsilon}$, but then in iteration $j+1$, $\epsilon' > \bar{\epsilon}$. Then on step j, we can apply lemma 2, and from the analysis above, $\epsilon' \leq \epsilon/2 \leq \bar{\epsilon}/2$. However, in any single iteration, the maximum probability mass the algorithm might throw out is $1/\gamma^2$, as can be seen from the proof of lemma 1:

$$a^2 \leq 1 - \gamma^2 p \quad \Rightarrow \quad 0 \leq 1 - \gamma^2 p \quad \Rightarrow \quad p \leq 1/\gamma^2$$

Thus in one step ϵ' increase by at most $\epsilon/[4n(\ln(R/r)+1)] \leq \frac{\bar{\epsilon}}{2}$, and so on step $j+1$, we still have $\epsilon' \leq \bar{\epsilon}$. This concludes the proof of theorem 2. □

We now give an alternate proof of theorem 2 using the construction given by Algorithm 1. We begin by proving an analogue to lemma 1.

Lemma 3 (Restriction to an Ellipsoid) *Let γ be fixed, and let μ be a full-dimensional isotropic distribution. Let $S = \{x : (x^Tx) \leq \gamma^2\}$ and $p = \Pr[x \notin S]$. Then*

$$Vol(W(M_S)) \geq e^{p\gamma^2/2}Vol(W(M))$$

Proof: First we establish the tradeoff for a radially symmetric distribution, and then we show that a radially symmetric distribution is the worst case for the tradeoff we want.

Let μ' be a radially symmetric distribution, and define M', S, and p as above. We then calculate the increase in $Vol(W(M'))$. Let $a^2 = \mathbf{E}_{\mu'}[(w^Tx)^2 : x \in S]\Pr[x \in S]$ for any $w, |w| = 1$. From the center of an n-dimensional sphere of radius γ, the projection of the sphere on to any direction is sharply concentrated around $\gamma/\sqrt{n}$, and the squared expectation is exactly γ^2/n. Using the identity

$$\mathbf{E}[(w^Tx)^2] = \mathbf{E}_{x\notin S}[(w^Tx)^2]\Pr[x \notin S] + \mathbf{E}_{x\in S}[(w^Tx)^2]\Pr[x \in S]$$

as in the proof of lemma 1, but now for any w, we deduce $1 \geq a^2 + \gamma^2 p/n$, and thus

$$a^n \leq \left(1 - \frac{\gamma^2 p}{n}\right)^{n/2} \leq e^{-\gamma^2 p/2}$$

As in the proof of lemma 1, we observe that $W(M'_S)$ includes a vector of length $1/a$ in the direction of w. Since this is now true for every w, the dual ellipsoid volume increases by at least a factor of $(1/a)^n$. This shows that in the case of a radially symmetric distribution,

$$Vol(W(M_S)) \geq e^{p\gamma^2/2} Vol(W(M))$$

Now we show that a radially symmetric distribution is the worst case for the tradeoff we want. Suppose there were some isotropic, full-dimensional distribution μ for which the statement of the lemma was not true. We construct a new isotropic, full-dimensional and radially symmetric distribution μ' for which the statement is also false.

We begin by noting that every point thrown out from μ is also thrown out from any rotation of μ – this just follows from the fact that μ is isotropic. Let μ' be the expectation of μ under a random rotation. That is, μ' is a radially symmetric distribution such that the probability of choosing x from μ' at distance less than r from the origin is exactly the same as the probability of choosing x from μ at distance less than r from the origin, for every r. Let M' correspond to μ'.

Consider an axis direction w_i of $E(M_S)$, $|w_i| = 1$. We have $a_i^2 = \mathbf{E}[(w_i^T x)^2 : x \in S] \Pr[x \in S]$. For $E(M'_S)$, denote the axis length for any axis (also just the radius of $E(M'_S)$) by $\bar{a}$. We find from the construction of μ' that

$$\bar{a}^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[(w_i^T x)^2 : x \in S] \Pr[x \in S] = \frac{1}{n}\sum_{i=1}^{n} a_i^2$$

One way to visualize this equality is to take μ and simply consider $\tilde{\mu}$ achieved by averaging over rotations of the axes of μ onto the other axes of μ; since this is a discrete set of rotations, it is clear that the squared axis lengths of $\tilde{\mu}$ are just the arithmetic averages of the squared axis lengths of μ. Then we can make $\tilde{\mu}$ into μ' by taking a continuous set of rotations, without affecting the axis lengths from $\tilde{\mu}$.

We now consider the volume of $E(M'_S)$. We have

$$Vol(E(M'_S)) = f(n)\prod_{i=1}^{n} \bar{a} = f(n)\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2}\right)^n \geq f(n)\prod_{i=1}^{n} a_i = Vol(E(M_S)$$

using the arithmetic mean-geometric mean inequality. This implies that $Vol(W(M_S)) \geq Vol(W(M'_S))$. This concludes the proof of lemma 3.

□

Finally, we prove that Algorithm 1 terminates with S satisfying theorem 2.

Proof of Theorem 2: As in the proof of theorem 2 using Algorithm 2, let $\beta = 4\frac{n}{\epsilon}(\ln \frac{R}{r}+1)$. Lemma 2 still holds. The rate of increase in the dual volume as we throw out probability mass (lemma 3) is the same as before (lemma 1). The only thing we need to address is what we called "the one remaining catch" in the proof using Algorithm 2. Our bound on the amount of probability mass that can be thrown out in a single step is no longer $1/\gamma^2$, but is now n/γ^2. However, $n/\gamma^2 = \epsilon/[4(\ln(R/r)+1)] \leq \frac{\bar{\epsilon}}{2}$ just as before. This concludes the analysis of Algorithm 1. □

The following connection shows that the success of either algorithm implies that they both succeed. If our criterion for a point x to be a β-outlier in a direction w were instead that

$$(w^T x)^2 > \beta \mathbf{E}[(w^T x)^2 : x \in P] \Pr[x \in P]$$

then Algorithms 1 and 2 both throw out the exact same points, and so must yield the same bound on β as a function of ϵ. To see this, note that any β-outlier under this definition remains a β-outlier as further points are removed, and so will have to be removed itself eventually. Also, no point is ever removed unless it currently is a β-outlier. Thus the two algorithms throw out exactly the same set of points in the end under this alternative definition of an outlier. In section 2.6, we develop this observation into an approximation algorithm for the problem of outlier removal using the standard definition of a β-outlier (not this alternative definition).

We pause to stress what we have gained by allowing some points of the distribution to be removed. If we force $\epsilon = 0$, then even under the hypothesis of theorem 2, β may be unbounded. Even a radially symmetric distribution satisfying the hypothesis with support in $\{B_R \setminus B_{r\sqrt{n}}\}$, where B_R denotes the ball of radius R, might have β as large as

$$\beta = \frac{R^2}{r^2}$$

By allowing $\epsilon > 0$, we have achieved

$$\beta = 4\frac{n}{\epsilon}(\ln\frac{R}{r} + 1)$$

2.3 Outlier Removal over Discrete Support

While theorem 2 might suffice for many applications, it is indeed possible that during outlier removal on an arbitrary set, the full-dimensional condition might be violated (indeed, the dimensionality of the remaining set might decrease). In this section we prove the following theorem, which shows that for distributions over integers, the full-dimensional condition is entirely unnecessary.

Theorem 1 (Outlier Removal over Discrete Support) *Let μ be a probability distribution on $\mathcal{Z}_b^n$. Then for every $\epsilon > 0$, there exists S and*

$$\beta = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right)$$

such that
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{(w^T x)^2 : x \in S\} \leq \beta \mathbf{E}[(w^T x)^2 : x \in S]$ *for all* $w \in \mathcal{R}^n$

The proof of this theorem presents two difficulties that were not present in the proof of theorem 2. First, μ might initially lie entirely on a lower-dimensional subspace, or μ might lie on a lower-dimensional subspace after the removal of a few points. Secondly, even if the distribution does not lie on a lower-dimensional subspace, we do not have the same lower bound on the smallest singular value of the distribution (singular value of the matrix

M associated with μ). While we insisted in the hypothesis of theorem 2 that the smallest singular value must be at least $1/r$, which will be roughly equivalent to 2^{-b} in the discrete case, it may be that the smallest singular value is actually 2^{-nb}, as the following example makes clear.

Example 1 *Let $B = 2^b$, and let each row of the matrix below represent a point in space. Denote the first $n-1$ rows by $\{v_i\}_{i=1}^{n-1}$ and denote the last row by p.*

$$\begin{bmatrix} B & -1 & & & \\ & B & -1 & & \\ & & B & -1 & \\ & & & \cdots & \\ 1 & & & & \end{bmatrix}$$

This set of points is clearly full-dimensional, and in most directions the singular values are on the order of B. However, in the direction $w = [B^{-n}, B^{-n+1}, \ldots, B^{-1}, 1]$, we find that $(w^T v_i)^2 = 0$ while $(w^T p)^2 = B^{-n} = 2^{-nb}$. Since $w > 1$, the singular value is actually slightly less than 2^{-nb}.

Example 1 shows that even disregarding issues of the distribution not being full-dimensional, we *cannot use theorem 2* to treat the distribution with integer support unless we are willing to settle for $\beta = \tilde{O}(\frac{n^2 b}{\epsilon})$. In extending our techniques to prove theorem 1, we will show that although one singular value may be small, they are not all small simultaneously in an appropriate *amortized* sense.

The first thing we shall define is a potential function that generalizes the dual ellipsoid volume we used in the proof of theorem 2. This potential function will account for the distribution μ being concentrated on a lower dimensional subspace, or even the possibility that μ is simply quite close to a lower dimensional distribution. We begin by defining the α-*core* of a distribution to be that subset of the distribution which lies on a subspace spanned by every large subset of the distribution. It will help to define the indicator function of E to be

$$\chi^E = \begin{cases} 1 & \text{if } E \text{ is true} \\ 0 & \text{if } E \text{ is false} \end{cases}$$

where E is a logical statement. The α-core is then given by

Definition 2 (α-core) *Define the* α-core *of μ_S to be μ_T, where $T \subset S$ is chosen to be maximum such that*

$$\forall w \in span(\mu_T) \text{ such that } w \neq 0, \quad \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha$$

We now establish some characteristics of the α-core, including that the α-core is well-defined.

Lemma 4 (Characterization of α-core)

(i) For any μ_T,

$$\forall w \in span(\mu_T) \text{ such that } w \neq 0, \quad \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha \qquad (a)$$

if and only if

$$\forall w, \quad w^T x \neq 0 \textit{ for some } x \in \mu_T \ \Rightarrow \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha \qquad (b)$$

(ii) $Q \subset S \implies \alpha$*-core*$(\mu_Q) \subset \alpha$*-core*$(\mu_S)$

(iii) $Q \subset S \implies \alpha$*-core*$(\mu_Q) = \alpha$*-core*$(\mu_Q \cap \alpha$*-core*$(\mu_S))$

(iv) Suppose that $\mu_T = \alpha$*-core*(μ_S)*, and that dim*$($*span*$(\mu_S)) = k$*, dim*$($*span*$(\mu_T)) = k'$*. Then* $\mu(S \setminus T) \leq (k - k')\alpha$

Proof: We first establish (i). Let μ_T be arbitrary. Assume (b) does not hold, i.e., there exists $x \in \mu_T$ and a direction w such that $w^T x \neq 0$ and $\sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. Writing

$$w = w_1 + w_2, \quad w_1 \in \textit{span}(\mu_T), \quad w_2 \perp \textit{span}(\mu_T)$$

we find $w^T x = w_1^T x \neq 0$, while $\sum_{x \in \mu_T} \chi^{\{w'^T x \neq 0\}} \mu(x) = \sum_{x \in \mu_T} \chi^{\{w_1^T x \neq 0\}} \mu(x) < \alpha$. Since $w_1 \in \textit{span}(\mu_T)$, (a) does not hold.

Now suppose that (b) does hold. If $w \in \textit{span}(\mu_T)$, then $w^T x \neq 0$ for some $x \in \mu_T$, and hence (b) implies that $\sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \alpha$, so (a) holds too.

To show (ii), we give an algorithm for constructing α-core(μ_S):

1. If there exists $x \in \mu_S$ and a direction w such that $w^T x \neq 0$ but $\sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$, remove x from μ_S.

2. Repeat until there does not exists such an x.

To argue the correctness of this algorithm, it suffices to show that if x meets the criterion of step 1, then x cannot be in any μ_R, $R \subset S$ such that $\forall w, w^T x \neq 0 \implies \sum_{x \in \mu_R} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. But this is obvious, since the w associated with x in step 1 satisfies $w^T x \neq 0$ and yet $\sum_{x \in \mu_R} \chi^{\{w^T x \neq 0\}} \mu(x) \leq \sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$. Therefore any x identified in step 1 cannot be in α-core(μ_S). Since the algorithm stops when it has arrived at μ_T satisfying (b), and no point has been removed that could be in α-core(μ_S), $\mu_T = \alpha$-core(μ_S). This establishes that the α-core is well-defined.

Now consider the order in which points are identified in step 1 when the algorithm is applied to μ_S. Considering points in the same order (and omitting points that are in μ_S but not in μ_Q), the algorithm run on μ_Q would always remove the points as well, simply because $\sum_{x \in \mu_{Q'}} \chi^{\{w^T x \neq 0\}} \mu(x) \leq \sum_{x \in \mu_{S'}} \chi^{\{w^T x \neq 0\}} \mu(x)$, where Q' and S' are Q and S minus the points that the algorithm has removed prior to the iteration under consideration.

We prove (iii) using (ii).

$$\begin{aligned} \alpha\text{-core}(\mu_Q) &\subset \alpha\text{-core}(\mu_S) \Rightarrow \\ \mu_Q \cap \alpha\text{-core}(\mu_Q) &\subset \mu_Q \cap \alpha\text{-core}(\mu_S) \Rightarrow \\ \alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_Q)) &\subset \alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_S)) \end{aligned}$$

We note that α-core$(\mu_Q \cap \alpha$-core$(\mu_Q)) = \alpha$-core(μ_Q). Now

$$\begin{aligned} \mu_Q \cap \alpha\text{-core}(\mu_S) &\subset \mu_Q \Rightarrow \\ \alpha\text{-core}(\mu_Q \cap \alpha\text{-core}(\mu_S)) &\subset \alpha\text{-core}(\mu_Q) \end{aligned}$$

Combining these yields α-core$(\mu_Q) = \alpha$-core$(\mu_Q \cap \alpha$-core$(\mu_S))$.

To see (iv), construct μ_T from μ_S in the following greedy manner. If $dim(span(\mu_T)) < dim(span(\mu_S))$, then

$$\exists w \in span(\mu_S) \quad \text{such that} \sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) < \alpha$$

If w were not $\perp$ to $span(\mu_T)$, we could write

$$w = w_1 + w_2, \quad w_1 \perp span(\mu_T), \quad w_2 \in span(\mu_T)$$

and then argue $\sum_{x \in \mu_S} \chi^{\{w^T x \neq 0\}} \mu(x) \geq \sum_{x \in \mu_T} \chi^{\{w^T x \neq 0\}} \mu(x) = \sum_{x \in \mu_T} \chi^{\{w_2^T x \neq 0\}} \mu(x) \geq \alpha$. Hence $w \perp span(\mu_T)$. Remove every point $x \in \mu_S$ such that $w^T x \neq 0$ (a less than α fraction of the total probability mass), and note that this causes $dim(span(\mu_S))$ to drop by at least 1. Therefore this construction can be iterated at most $(k - k')$ times, and hence $\mu(S \setminus T) \leq (k - k')\alpha$. □

We can now define the potential ϕ of a distribution (or a subset of a distribution).

Definition 3 (Potential Function: ϕ) *Let μ_T be the α-core of μ_S. Let $\phi(\mu_S)$ be $Vol(W(M_T))$, the volume of the dual ellipsoid of μ_T. If μ_T is not full dimensional, but instead lies in a space of dimension k, let $\phi(\mu_S)$ be $Vol_k(W(M_T))$, the k-dimensional volume (within the span of μ_T) of the dual ellipsoid of μ_T.*

We now prove upper and lower bounds on $\phi(\mu_S)$, analogous to lemma 2, for the case that the α-core of μ_S is full-dimensional. Although a tighter version of this lemma may be possible, the analysis here is sufficient to show the asymptotic result of theorem 1.

Lemma 5 (Bounds on ϕ) *Denote the α-core of μ_S by μ_T and suppose that μ_T is full-dimensional (and hence $\mu_T = \mu_S$). Then*

$$\phi(\mu_S) \geq (2^b \sqrt{n})^{-n} f(n)$$

$$\phi(\mu_S) \leq (n/\alpha)^n f(n)$$

Proof: We lower bound ϕ by showing that for any vector v satisfying $|v| \leq \frac{2^{-b}}{\sqrt{n}}$, v is in the dual ellipsoid. Using that no element x of μ has length greater than $\sqrt{n}2^b$, we find that

$$v^T M_S v = \sum_{x \in \mu_S} (x^T v)^2 \mu(x) \leq \sum_{x \in \mu_S} x^2 v^2 \mu(x) \leq 1$$

The claimed lower bound now follows from the fact that $W(M_S)$ contains a ball of radius $\frac{2^{-b}}{\sqrt{n}}$.

To upper bound $\phi(\mu_S)$, we will use that μ_T is full-dimensional. Because the volume of an ellipse is equal to the product of the axis lengths times a factor that depends only on the dimension, we have that $\phi(\mu_S) = f(n)/Det(M_S)$ where $M_S = \sum_{x \in \mu_S} x x^T \mu(x)$. We now show that we can decompose M_S into a set of simpler components plus some extra points,

$$M_S = \sum_i \lambda_i M_i + \sum_y y y^T$$

where each M_i is a positive definite nxn matrix of integers and $\sum \lambda_i \geq \alpha/n,\ \lambda_i \geq 0$.

To see this decomposition, begin by picking any point $x_1 \in \mu_S$. Now pick any point $x_2 \in \mu_S$ such that $x_2 \notin span(x_1)$. Now pick any point $x_3 \notin span(x_1, x_2)$. Continuing, we can always make such a choice by considering any direction w perpendicular to the span of the previous points — any point with non-zero inner product with this w, guaranteed to exist by the definition of α-core, lies off the span of the previous points. This first set of points $\{x_j\}_{j=1}^n$ yields $M_1 = \sum_j x_j x_j^T$ with $\lambda_1 = \min_j \mu(x_j)$. To form M_2, we must restrict ourselves to picking points from $\{\mu_S \setminus \lambda_1 M_1\}$ (using slight overloading of notation). By the definition of α-core, as long as $\sum \lambda_i < \alpha/n$, we will always be able to form a new M_i because we have subtracted off less than an α fraction of the probability mass from the distribution thus far. The process can be seen to terminate in a finite number of steps because the support of μ_S is initially a finite number of points, and at every step the cardinality of the support decreases by at least one. The $\{y\}$ which we referred to as "extra points" above are simply the points remaining in μ_S when this operation, having formed a sufficient number of M_i, comes to an end.

Note that each matrix M_i satisifies $Det(M_i) \geq 1$ because it is the sum of the products of many integer terms and it is positive (because M_i is positive definite). We now show that we may ignore the y terms in establishing a lower bound for $Det(M_S)$. Another consequence of M_i being positive definite is that $M_i = A_i A_i^T$ for some A_i. Since the determinant of M_S is the product of the eigenvalues, and each eigenvalue e_j is equal to $\sum_i \lambda_i (A_i^T w_j)^2 + \sum_y (y^T w_j)^2$ for some unit vector w_j, $Det(\sum_i \lambda_i A_i A_i^T) \leq Det(\sum_i \lambda_i A_i A_i^T + \sum yy^T)$.

We have from fact 2 in section 2.10 (and since the geometric mean is at least the min) that for $\sum \lambda_i' = 1$,

$$Det(\sum_i \lambda_i' M_i) \geq \min_i \{Det(M_i)\}$$

The last step is to write $Det(\xi M) = \xi^n Det(M)$, which implies $Det(M_S) \geq (\alpha/n)^n$. This yields the claimed upper bound on $\phi(\mu_S)$. □

Note that lemma 5 implies that the log of the ratios between the upper and lower bounds on ϕ is at most $n(b + 1.5(\log \frac{n}{\alpha}))$. (The relevant setting of α for the proof of theorem 1 will be $\alpha = \epsilon/(3n)$.) This compares favorably with the corresponding ratio in the continuous case, $n \ln \frac{R}{r}$, and suggests that we have not introduced much slack while extending our techniques to amortize over the singular values.

We now address the issue of dimension dropping. We refer to *non-monotone growth* in the title of lemma 6 because now ϕ may drop when we remove some of the distribution. To see this, consider example 1 again: ϕ is initially about $f(n)$, but after removing the point p, ϕ becomes roughly $2^{-b(n-1)} f(n)$. In the proof of theorem 2, we bounded the drop in probability mass by bounding the increase in the volume of the dual ellipsoid. Because ϕ may decrease greatly during the course of the algorithm (when the α-core drops in dimension), a bound on ϕ's final value is no longer enough to bound the drop in probability mass. Happily, we can still bound the growth of ϕ in the following sense:

Lemma 6 (Non-Monotone Growth of ϕ) *Over the course of either algorithm on distribution μ, let $(\Delta\phi)_i$ denote the relative increase in ϕ while α-core(μ) spans a subspace of dimension i (or 1 if α-core(μ) is never concentrated on a subspace of dimension i). Then*

$$\prod_i (\Delta\phi)_i \leq 2^{n(b + 3\log\frac{n}{\alpha} + 1)}$$

Proof: Suppose that initially the α-core of μ is full-dimensional, and that $\prod_i (\Delta\phi)_i = V$. Under a simplifying assumption, we construct a distribution μ' such that the α-core of μ' is full-dimensional and $\phi(\mu')/\phi(\mu) \geq V$. (If the result of applying the oulier removal algorithm to μ is μ_S that has full-dimensional α-core, then $\mu' = \mu_S$ and there is nothing to do.) By lemma 5, $\phi(\mu')$ and $\phi(\mu)$ cannot differ by a factor of more than $2^{n(b+1.5\log\frac{n}{\alpha})}$, and thus this suffices to prove the bound on V. We then remove the simplifying assumption. We defer the issue that the α-core of μ might not initially be full-dimensional to the very end of the proof.

Suppose that the algorithm goes from μ_R of dimension $(i+1)$ to μ_S of dimension i, and then runs for a while to produce μ_T (still of dimension i). The simplifying assumption we mentioned above is that the dimension of the α-core has only fallen by 1 on this step. For ease of exposition, assume that each distribution is equal to its α-core. This is without loss of generality because ϕ is defined in terms of the α-core, and so the points outside the α-core are irrelevant for this lemma. We will construct $\mu'_{S'}$ and $\mu'_{T'}$ of dimension $(i+1)$ such that $\phi(\mu'_{S'}) \geq \phi(\mu_R)$ and $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. Then we will have

$$\phi(\mu'_{T'}) = \frac{\phi(\mu_T)}{\phi(\mu_S)}\phi(\mu'_{S'}) \geq (\Delta\phi)_i \phi(\mu_R)$$

Applying this construction iteratively over all the dimensions yields μ' of dimension n satisfying $\phi(\mu')/\phi(\mu) \geq V$.

Let us now construct $\mu'_{S'}$ and $\mu'_{T'}$. Define $p_j = \mu(x_j)$ for all $x_j \in \mu_{R\setminus S}$ and let $P = \sum_j p_j$. Then

$$M_R = M_S + \sum_j p_j x_j x_j^T = \sum_i \frac{p_j}{P}(M_S + P x_j x_j^T)$$

Define X_j to be $(M_S + P x_j x_j^T)$. By fact 2 (section 2.10),

$$Det(\sum \lambda_j X_j) \geq \min\{Det(X_j)\}, \quad \sum \lambda_j = 1$$

there exists j such that $Det(X_j) \leq Det(M_R)$. Denote this particular x_j by x, and let

$$\mu'_{S'} = \{\mu_S + x \text{ with weight } \alpha\}$$

$$\mu'_{T'} = \{\mu_T + x \text{ with weight } \alpha\}$$

Note that $P \geq \alpha$, and so $Det(M'_{S'}) \leq Det(X_j)$. Thus $\phi(\mu'_{S'}) \geq \phi(\mu_R)$.

We now show $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. Rotate the distributions so that span(μ_S) and span(μ_T) are equal to the first i coordinate axes, and x lies in the span of the first $i+1$ coordinate axes. Denote the vector formed from the first i coordinates of x by $x[1\ldots i]$, and the $(i+1)^{st}$ coordinate of x by $x[i+1]$. Then the distance of x to span(μ_S) is just $x[i+1]$, and this is also the distance of x to span(μ_T). We have $\phi(\mu_T) = f(i)/Det(M_T)$, while

$$\phi(\mu'_{T'}) = f(i+1)/Det\left(\begin{bmatrix} M_T & 0 \\ 0 & 0 \end{bmatrix} + \alpha^2 \begin{bmatrix} x[1\ldots i]x[1\ldots i]^T & x[i+1]x[1\ldots i] \\ x[i+1]x[1\ldots i]^T & x[i+1]^2 \end{bmatrix}\right)$$

where the upper left matrix block $(M_T + \alpha^2 x[1\ldots i]x[1\ldots i]^T)$ is ixi. For any matrix A, subtracting a scalar multiple of some row of A from another row of A does not change the determinant of A. To calculate $\phi(\mu'_{T'})$, we subtract $x[l]/x[i+1]$ times the last row of the matrix from the l^{th} row for every $l \leq i$. This yields

$$Det\left(\begin{bmatrix} M_T & 0 \\ 0 & 0 \end{bmatrix} + \alpha^2 \begin{bmatrix} 0 & 0 \\ x[i+1]x[1\dots i]^T & x[i+1]^2 \end{bmatrix}\right) = Det(M_T)\alpha^2 x[i+1]^2$$

Therefore $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = (\alpha x[i+1])^{-2}\frac{f(i+1)}{f(i)}$. An identical calculation yields an identical result for $\frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. This shows that $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$.

We now remove the simplifying assumption and extend this construction to the case that at some step the α-core falls in dimension by more than 1. If μ_R and μ_S differ by k dimensions, we construct $\mu'_{S'}$ by adjoining k points from $\mu_{R\setminus S}$, each with weight α/k. We now show how to find these points. Since μ_R is an α-core, and $span(\mu_S)$ is a subspace of k dimensions less, we can use the construction of lemma 5 to write

$$M_R = M_S + \sum_i \lambda_i A_i A_i^T + \sum yy^T, \qquad \sum_i \lambda_i = \Lambda \geq \frac{\alpha}{k}$$

where each A_i is a set of k points such that $span(\{\mu_S + A_i\}) = span(\mu_R)$. As above,

$$Det(M_R) \geq Det(M_S + \sum_i \lambda_i A_i A_i^T) \geq \min_i\{Det(M_S + \Lambda A_i A_i^T)\} \geq \min_i\{Det(M_S + \frac{\alpha}{k} A_i A_i^T)\}$$

Let A denote the A_i realizing this minimum and let

$$\mu'_{S'} = \{\mu_S + A \text{ with weight } \frac{\alpha}{k}\}$$

$$\mu'_{T'} = \{\mu_T + A \text{ with weight } \frac{\alpha}{k}\}$$

We have $\phi(\mu'_{S'}) \geq \phi(\mu_R)$ by construction. It remains to show $\frac{\phi(\mu'_{T'})}{\phi(\mu_T)} = \frac{\phi(\mu'_{S'})}{\phi(\mu_S)}$. We do this by by showing that the previous calculation (giving this fact under the simplifying assumption) can be repeated k times. Let ${\mu'}^{(l)}_{S'}$ denote $\{\mu_S +$ first l points of A with weight $\frac{\alpha}{k}\}$ and define ${\mu'}^{(l)}_{T'}$ similarly. Then the previous calculation yields

$$\frac{\phi({\mu'}^{(1)}_{T'})}{\phi(\mu_T)} = \frac{\phi({\mu'}^{(1)}_{S'})}{\phi(\mu_S)}, \quad \frac{\phi({\mu'}^{(2)}_{T'})}{\phi({\mu'}^{(1)}_{T'})} = \frac{\phi({\mu'}^{(2)}_{S'})}{\phi({\mu'}^{(1)}_{S'})}, \quad \dots \quad \Rightarrow \frac{\phi({\mu'}^{(k)}_{T'})}{\phi({\mu'}_{T'})} = \frac{\phi({\mu'}^{(k)}_{S'})}{\phi({\mu'}_{S'})}$$

This concludes the construction of μ' which is full-dimensional and at least an α/n-core.

We now turn to the case that μ is initially only k dimensional, where $k < n$. In this case, we adjoin any $(n-k)$ points, each with weight α, to form $\bar{\mu}$, where $\bar{\mu}$ is full-dimensional. Then $\bar{\mu}$ may not be a probability distribution, but it has total weight at most $1 + n\alpha$, and so

$$\phi(\bar{\mu}) \geq (2^b\sqrt{n}\sqrt{1+n\alpha})^{-n} f(n)$$

by the same construction as in the lower bound of lemma 5. The iterative construction above (without the simplifying assumption) yields $\bar{\mu}'$ such that $\phi(\bar{\mu}') \leq (n^2/\alpha)^n f(n)$, and so

$$\phi(\bar{\mu}')/\phi(\bar{\mu}) \leq (\frac{n^2}{\alpha}2^b\sqrt{n+n^2\alpha})^n \leq (2n^3 2^b/\alpha)^n \leq 2^{n(b+3\log\frac{n}{\alpha}+1)}$$

This concludes the proof of the lemma. □

We now prove that Algorithm 2 applied to a distribution μ over the b-bit integers yields S satifying theorem 1.

Proof of Theorem 1: Let $\alpha = \epsilon/(3n)$ and let $\beta = \gamma^2 = 6\frac{n}{\epsilon}(b + 4\log\frac{n}{\epsilon} + 3)$. The only time that the drop in probability mass due to action by the algorithm does not lead to an increase in ϕ is when either the algorithm causes the dimension of the α-core to drop, or the algorithm removes probability mass that lies outside the α-core.

We first consider the fraction of the distribution that is not part of the α-core. Initially, $dim(span(\mu))$ is at most n. Suppose $dim(span(\alpha\text{-core}(\mu))) = k$. Then by lemma 4:(iv), at most an $\alpha(n-k)$ fraction of μ lies outside of the α-core of μ. Points only leave the α-core when they are removed by the algorithm, or when the algorithm takes μ_S to $\mu_{S'}$ in one step and

$$dim(\alpha\text{-core}(\mu_S)) = k_1, \quad dim(\alpha\text{-core}(\mu_{S'})) = k_2, \quad k_2 < k_1$$

In the latter case, the probability mass lost from the α-core (and not removed by the algorithm) is given by

$$\{\mu_{S'} \cap \alpha\text{-core}(\mu_S)\} \setminus \{\alpha\text{-core}(\mu_{S'})\}$$

Since $S' \subset S$, by lemma 4:(iii), this is the same as

$$\{\mu_{S'} \cap \alpha\text{-core}(\mu_S)\} \setminus \{\alpha\text{-core}(\mu_{S'} \cap \alpha\text{-core}(\mu_S))\}$$

and by lemma 4:(iv) this is no more than $\alpha(k_1 - k_2)$. Since the cumulative drop in dimension of the α-core is no more than n dimensions, no more than an αn fraction of the distribution ever leaves the α-core (without being removed by the algorithm) over the course of the algorithm.

We now bound the amount of the distribution removed by the algorithm on steps in which the α-core drops in dimension. In the proof of theorem 2, we showed that in any single step, Algorithm 2 throws out no more than a $1/\gamma^2$ fraction of the distribution. Since there are no more than n steps where the α-core drops in dimension, we throw out no more than an n/γ^2 fraction in this way. This yields that at most an $n\alpha + n/\gamma^2 \leq 2\epsilon/3$ fraction of the probability mass that we throw away does not contribute to increasing ϕ.

We now proceed exactly as we did in the proof of theorem 2 for Algorithm 2. Every time we remove p_i of the probability mass from the α-core and the α-core does not drop in dimension, we have from lemma 1 that ϕ must increase by $e^{p_i\gamma^2/2}$. If we throw out an ϵ' fraction of μ, at most a $2\epsilon/3$ fraction does not contribute to ϕ increasing, so by application of lemma 1

$$\prod (\Delta\phi)_i \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})}$$

Lemma 6 then yields

$$2^{n(b+4\log\frac{n}{\epsilon}+3)} \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})}$$

$$\Rightarrow n(b + 4\log\frac{n}{\epsilon} + 3) \geq \frac{\gamma^2}{2}(\epsilon' - \frac{2\epsilon}{3})$$

$$\Rightarrow 1 \geq \frac{3}{\epsilon}(\epsilon' - \frac{2\epsilon}{3}) \Rightarrow \epsilon' \leq \epsilon$$

This concludes the proof of theorem 1 using Algorithm 2. □

We now prove theorem 1 using Algorithm 1. As we noted previously, this may be obtained as a corollary of the success of Algorithm 2, but a direct proof raises an additional issue that we explore below. The resolution of this issue leads to a bound on β with smaller leading constant.

Proof of Theorem 1: Let $\alpha = \epsilon/(3n)$ and let $\beta = \gamma^2 = 3\frac{n}{\epsilon}(b + 4\log\frac{n}{\epsilon} + 3)$. The only new issue is bounding the amount of probability mass removed by the algorithm on steps in which the α-core drops in dimension. We might remove up to an n/γ^2 fraction in a single step, but our asymptotic bound would not stand up if we could remove up to an n^2/γ^2 fraction of the probability mass over the course of the algorithm.

Suppose the α-core falls by k dimensions in one step of the algorithm. Rather than considering all the points outside $S = \{x : |x| \leq \gamma \text{ after rounding}\}$ as being removed at once, imagine instead that the probability mass on every point is uniformly decreased. Then, ϕ increases continuously except for at most k discrete time steps, when the dimension of the α-core drops. Apart from the steps on which the α-core drops, ϕ increases as a function of the probability mass removed exactly as implied by lemma 3. Every time the α-core drops by i dimensions, at most an $i\alpha$ amount of probability mass leaves the α-core (by lemma 4:(iv)). Therefore at most a $k\alpha$ amount of probability mass is removed without an increase in ϕ. In this thought experiment, no probability mass in the α-core is removed by the algorithm without an increase in ϕ. Thus at most an $n\alpha = \epsilon/3$ amount of probability mass is removed without an increase in ϕ. We apply lemma 3 and lemma 6 as before to obtain

$$\prod(\Delta\phi)_i \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})}$$

$$\Rightarrow 2^{n(b+4\log\frac{n}{\epsilon}+3)} \geq e^{\frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})}$$

$$\Rightarrow n(b + 4\log\frac{n}{\epsilon} + 3) \geq \frac{\gamma^2}{2}(\epsilon' - \frac{\epsilon}{3})$$

$$\Rightarrow 1 \geq \frac{3}{2\epsilon}(\epsilon' - \frac{\epsilon}{3}) \Rightarrow \epsilon' \leq \epsilon$$

This concludes the proof of theorem 1 using Algorithm 1. □

2.4 Efficiency

In this section we describe polynomial time versions of both algorithms. The computational model is to allow multiplications and additions in unit time.

2.4.1 Point sets

Suppose the distribution μ is specified explicitly as a set of m points with weights corresponding to probabilities. Then we can achieve exactly the stated value of β with either algorithm deterministically. The running time for either algorithm is given by the time to compute M $(O(mn^2))$, the time to round the distribution $(O(n^3 + mn^2))$, the time to find an outlier $(O(mn))$, and the need to repeat the whole process up to m times. This yields a time bound of $O(m^2n^2 + mn^3)$.

In the above discussion we made the worst case assumption that only one data point was thrown out in each iteration of rounding and looking for outliers. In the case that a single data point is throw out, rounding the distribution can be done more efficiently. If the distribution is initially isotropic, and v of probability p is removed, then $M' = I - pvv^T$

gives the new inertial ellipsoid. We can factor M'^{-1} symbolically as

$$M'^{-1} = BB^T = \left(I - \left(1 - \frac{1}{\sqrt{1 - v^T vp}}\right) \frac{vv^T}{v^T v}\right)^2$$

where we have chosen B to be symmetric. To verify this calculation, note that

$$BM'B^T = (I - bvv^T)(I - pvv^T)(I - bvv^T) = [I - (2b - b^2v^2)vv^T][I - pvv^T]$$

where $b = \frac{1}{v^2}\left(1 - \frac{1}{\sqrt{1-v^2p}}\right)$ and we have used that the matrices commute. We calculate

$$2b - b^2v^2 = \frac{1}{v^2}\left((2 - \frac{2}{\sqrt{1 - pv^2}}) - (1 - \frac{2}{\sqrt{1 - pv^2}} + \frac{1}{1 - pv^2})\right)$$

$$= \frac{1}{v^2}\left(1 - \frac{1}{1 - pv^2}\right) = \frac{-p}{1 - pv^2}$$

Plugging this in completes the verification

$$[I - (2b - b^2v^2)vv^T][I - pvv^T] = [I + \frac{p}{1 - pv^2}vv^T][I - pvv^T]$$

$$= [I + (\frac{p}{1 - pv^2} - p - \frac{p^2v^2}{1 - pv^2})vv^T] = I$$

If the old distribution was $\{x\}$, the new isotropic distribution is $\{Bx\}$, where our formula for B yields

$$Bx = x - \left(1 - \frac{1}{\sqrt{1 - v^T vp}}\right) \frac{v(v^T x)}{v^T v}$$

which is computable in time $O(n)$ for any point x. Another explanation for this formula is that we are just correcting the inertial ellipsoid in the direction of v; this type of update step is sometimes referred to as a *rank-1 update*. Using this observation, we can compute M from scratch once ($O(mn^2)$), round the distribution from scratch once ($O(n^3 + mn^2)$), and then find an outlier ($O(mn)$) and reround using our formula above ($O(mn)$) a total of at most m times. This yields the improved time bound of $O(m^2n + mn^2 + n^3)$. If we throw away less than an ϵ fraction of the point set, the time bound is just $O(\epsilon m^2n + mn^2 + n^3)$.

If we specialize our analysis to $\mathcal{Z}_b^n$ and the case that the distribution has full-dimensional α-core throughout the algorithm, we can obtain a running time with a different dependance on the relevant parameters. Suppose that on some step of Algorithm 1 with parameter β we remove all β-outliers and ϕ (equivalently, the dual ellipsoid volume) increases by a factor of no more than $(1 + \delta)$ — then the remaining data set is $(1 + \delta)\beta$-outlier free. Because we may have removed many points, we cannot use the technique just developed above, and our time bound is $O(mn^2 + n^3)$ per iteration. However, by our upper and lower bounds on ϕ, there are at most $\log_{(1+\delta)} 2^{\tilde{O}(nb)} = \tilde{O}(\frac{nb}{\delta})$ iterations where ϕ increases by $(1+\delta)$ or more. The final bound on the running time is then $\tilde{O}(\frac{mn^3b+n^4b}{\delta})$ to obtain a $(1 + \delta)\beta$-outlier free set.

2.4.2 Arbitrary distributions

Now suppose that we are not given μ explicitly, but rather only the ability to sample from μ. For ease of exposition, we will refer only to the case that the support of μ is in $\mathcal{Z}_b^n$. The outlier-free restriction of μ will be specified as the part of μ contained in an ellipsoid. The algorithm for distributions is:

1. Get a set $P = \{x_1, \dots, x_m\}$ of m samples from μ.

2. Run the outlier removal algorithm on the discrete point set P with parameter Γ^2.

3. Let P' be the outlier-free subset of P. Then the outlier-free restriction of P is given by $\Gamma E(M')$, where $M' = \frac{1}{m}\sum_{x_i \in P'} x_i x_i^T$. The outlier-free restriction of μ is given by $(1+\delta)\Gamma E(M')$, where $\delta \in (0, 1/4)$ is an accuracy parameter.

The main theorem of this section is the following.

Theorem 3 (Sample Complexity) *Let*

$$m = O\left(\frac{\gamma^2}{\delta^2}\left(n \log \frac{n}{\delta} + \log \frac{n(b + \log n)}{\delta}\right)\right) = \tilde{O}\left(\frac{n\gamma^2}{\delta^2}\right)$$

Then with high probability, either outlier removal algorithm run with parameter $\Gamma^2 = (1+\delta)^2\gamma^2$ returns an ellipsoid $T = \Gamma E(M')$ satisfying
(i) $\mu((1+\delta)T) \geq 1 - \epsilon$
(ii) $(1+\delta)T$ has no $(1+\delta)^{O(1)}\gamma^2$-outliers
where (γ^2, ϵ) is achieved by the deterministic omniscient algorithm (omniscient in that it knows the distribution exactly).

For the remainder of this section, assume that the deterministic omniscient algorithm with parameter γ^2 finds a subset S such that $\mu(S) \geq 1 - \epsilon$, and μ_S has no γ^2-outliers. The statement "μ_S has no γ^2-outliers", or simply "S has no γ^2-outliers" (since μ is implicit), is exactly that

$$\forall w, \quad \max\{(w^T x)^2 : x \in S\} \leq \gamma^2 \mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] = \gamma^2 \sum_{x \in S} (w^T x)^2 \mu(x)$$

The max is not over $x \in \mu_S$, but rather $x \in S$. This is an important subtlety. Since S and T constructed by the algorithm are always convex, whenever we have $\forall w, \max\{(w^T x)^2 : x \in S\} \leq \max\{(w^T x)^2 : x \in T\}$, we will be able to conclude that $S \subseteq T$. If we defined the max over $x \in \mu_S$, we would only be able to conclude that $\mu(T \setminus S) = 0$, which would increase the length of the proof.

We know that $\gamma^2 = \tilde{O}(\frac{bn}{\epsilon})$ is always achievable, but in some cases we may do better. Our bound on running time is proved for arbitrary values of γ^2.

Suppose that at some step we can estimate $E(M)$ to within $1 \pm \delta$ in every direction. Let $\Gamma^2 = (1+\delta)^2\gamma^2$. Then every point that we perceive to be a Γ^2-outlier will be at least a γ^2-outlier with respect to the true distribution, and so removing them does not throw away any point that the deterministic algorithm keeps. Similarly, if we perceive the distribution to have no Γ^2-outliers, the true distribution will have no $(1+\delta)^2\Gamma^2$-outliers. Before removing outliers, we may not have that our working estimate of M, $\bar{M}$, is within $1 \pm \delta$ of M. However, whenever we are wrong by more than $1 + \delta$, there is some true outlier with respect to the original distribution that we throw out even using our flawed estimate $\bar{M}$. This line of reasoning (made rigorous) will allow us to find a $(1+\delta)^{O(1)}\gamma^2$-outlier-free subset in space, where γ^2 is achieved by the deterministic version of the algorithm. In lemma 7 we show this for a particular direction in a particular iteration. In lemma 8 we extend this to all iterations, and in the proof of theorem 3 we extend this to all directions and all iterations, at every step bounding the sample complexity.

Lemma 7 (Outlier Detection, One Iteration) *Fix a direction w. Let S be a subset of space. Let our number of samples be*

$$m = O(\frac{\gamma^2}{\delta^2})$$

and consider the sample distances in direction w given by $\{w^T x_i\}$. Let y denote the true variance of S and $\bar{y}$ denote the sample variance,

$$y = \sum_{x \in S} (w^T x)^2 \mu(x) \qquad \bar{y} = \frac{1}{m} \sum_{x_i \in S} (w^T x_i)^2$$

Then with constant probability
(i) $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y \Rightarrow (1-\delta) y \leq \bar{y} \leq (1+\delta) y.$
(ii) $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y$ *and* $T = \{x : (w^T x)^2 \leq \Gamma^2 \bar{y}\} \Rightarrow S \subset T.$

Proof: Property (i) says that we do correctly estimate the variance of an outlier-free restriction of the distribution, and property (ii) assures us that any outlier-free restriction of the distribution has no probability mass past Γ^2 times the sample variance (i.e., we can always safely throw away probability mass using the sample variance). Both claims are for a fixed direction w. Note that S is assumed to be γ^2-outlier-free in the hypotheses of both (i) and (ii). Lemma 8 will not rely upon part (ii) explicitly, but it will involve a similar argument.

Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$, or 0 if $x_i \notin S$. Without loss of generality, assume $\max\{(w^T x)^2 : x \in S\} = 1$ (by an appropriate scaling). First we show (i). Since μ_S has no γ^2-outliers, we have $y \geq \frac{1}{\gamma^2}$. Applying the Chernoff bound to determine the probability that $\bar{y}$ is not a good estimate for y, we have

$$\Pr[|m\bar{y} - my| \geq \delta m y] \leq e^{-\delta^2 m y/3}$$

This occurs with constant probability for $m = O(\frac{\gamma^2}{\delta^2})$.

Now we show (ii). Let T be as above, and again assume $\max\{(w^T x)^2 : x \in S\} = 1$ without loss of generality. If S has no γ^2-outliers, then $y \geq \frac{1}{\gamma^2}$, and we would have found $\bar{y}$ to be an accurate estimate by the analysis in the previous paragraph. In this case, $(1-\delta)y \leq \bar{y} \Rightarrow y \leq (1+\delta)^2 \bar{y}$, and S has no γ^2-outliers implies $\max\{(w^T x)^2 : x \in S\} \leq \gamma^2 y \leq \Gamma^2 \bar{y}$. This then implies $S \subseteq T$. □

Lemma 8 (Outlier Detection, Many Iterations) *Fix w. Assume S is full-dimensional. Let*

$$m = O\left(\frac{\gamma^2}{\delta^2} \log \frac{n(b + \log n)}{\delta}\right) = \tilde{O}\left(\frac{\gamma^2}{\delta^2}\right)$$

Then with constant probability either outlier removal algorithm restricted to w with parameter Γ^2 produces a subset of space

$$T = \{x : (w^T x)^2 \leq t\}$$

for some value t such that
(i) For any subset of space S that has no γ^2-outliers along w, $S \subseteq T$.
(ii) $(1+\delta)T$ has no $(1+\delta)^8\gamma^2$-outliers along w.

Proof: By "either outlier removal algorithm restricted to w", we simply mean the one-dimensional version of the two algorithms. Consider S achieved by the deterministic omniscient version of the algorithm (restricted to w). Since our outlier removal algorithm only throws away probability mass when necessary, this S is the largest possible restriction that is γ^2-outlier free. Define y and $\bar{y}$ as in lemma 7. By lemma 7 part (i), we have that $\bar{y}$ is a good approximation to y. This ensures that with good probability, we identify S as Γ^2-outlier-free, and so (i) is proved. It remains to show that, if our algorithm for some reason chooses a substantially larger set T, then $(1+\delta)T$ has no $(1+\delta)^8\gamma^2$-outliers.

Define $T_\alpha = \{x : (w^T x)^2 \leq \alpha\}$. Suppose $\exists \alpha$ such that T_α has no Γ^2-outliers. Then $T_{(1+\delta)\alpha}$ has no $(1+\delta)^2\Gamma^2$-outliers. This follows from the fact that

$$\max\{(w^T x)^2 : x \in T_{(1+\delta)\alpha}\} \leq (1+\delta)^2 \max\{(w^T x)^2 : x \in T_\alpha\}$$

and $\mathbf{E}[(w^T x)^2 : x \in T_\alpha] \Pr[x \in T_\alpha]$ is a monotonically increasing function of α.

Suppose we estimate that some set $T = T_t$ has no Γ^2-outliers (in which case the algorithm might return T as an answer). Then our sample also leads us to calculate that T_α has no $(1+\delta)^2\Gamma^2$-outliers for $\alpha \in [t, (1+\delta)t]$ by the same reasoning as in the preceding paragraph. For every t, we will show that for some nearby (within a factor of $(1+\delta)$) value of α, we estimate the sample variance of the restriction of μ to T_α with sufficient accuracy. We proceed to analyze what values of α we need to consider.

Assume without loss of generality that w is a unit vector. An easy upper bound on $\max(w^T x)^2$ is $2^b\sqrt{n}$. To develop a lower bound, we will need to use the assumption that S is full-dimensional. For any μ_S, we can write $\max(w^T x)^2 \geq \mathbf{E}[(w^T x)^2]$. By decomposing μ_S in the manner of lemma 5, we can obtain the stronger statement $\max(w^T x)^2 \geq \mathbf{E}_{\{y_j\}_{j=1}^n}[(w^T y_j)^2]$ where the probability distribution on the $\{y_j\}$ is uniform and the $\{y_j\}$ are full-dimensional. The term $\mathbf{E}[(w^T y_j)^2]$ is lower bounded by the smallest singular value of the $\{y_j\}$. We have previously shown that the product of the singular values of such a distribution is at least n^{-2n}. Since no individual singular value is more than $2^b\sqrt{n}$, we have that the smallest is at least $n^{-2n}\ 2^{n(b+.5\log n)} = 2^{\tilde{O}(nb)}$. Therefore we can restrict our attention to $\alpha = (1+\delta)^k$ for k an integer and union bound over the at most $\log_{(1+\delta)} 2^{\tilde{O}(nb)} = O(\frac{n(b+\log n)}{\delta})$ possible values for k.

We now show that if we estimate T_α to have no $(1+\delta)^2\Gamma^2$-outliers, then with good probability T_α actually has no $(1+\delta)^6\Gamma^2$-outliers with respect to the true distribution, and by our reasoning above, since there is an α within $(1+\delta)$ of t, $T_{(1+\delta)t}$ is $(1+\delta)^8\Gamma^2$ outlier-free.

We do this by showing that if T_α has a $(1+\delta)^6\Gamma^2$-outlier, then with good probability our sample shows T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier. Let X_i be the random variable representing the squared distance of x_i along the direction w, $X_i = (w^T x_i)^2$, or zero if $x_i \notin T_\alpha$. Without loss of generality, assume $\alpha = 1$. Define y and $\bar{y}$ as in lemma 7 (but with

T_α in place of S). Then by assumption on T_α, $y = \mathbf{E}[X_i] \leq \frac{1}{(1+\delta)^6\Gamma^2}$. The condition that our samples show T_α to have at least a $(1+\delta)^2\Gamma^2$-outlier is $\bar{y} = \frac{1}{m}\sum X_i \leq \frac{1}{(1+\delta)^2\Gamma^2}$. We apply the Chernoff bound,

$$\Pr[\bar{y} > (1+\Delta)y] < e^{-\Delta^2 my/3}$$

where we have stated the Chernoff bound for the case that $\Delta < 1$. Let $\Delta = \frac{1}{y(1+\delta)^2\Gamma^2} - 1$ (this yields the event that $\bar{y} > \frac{1}{(1+\delta)^2\Gamma^2}$ in our probability calculation). If $\Delta < 1$, then

$$\Delta^2 y \quad = \quad \left(\frac{1}{(1+\delta)^2\Gamma^2} - y\right)^2 \frac{1}{y} \quad \geq \quad \left(\frac{4\delta}{(1+\delta)^6\Gamma^2}\right)^2 \frac{1}{y} \quad \geq \quad \frac{\delta^2}{\Gamma^2}$$

and the probability we do not correctly identify the furthest outlier is at most $e^{-\Delta^2 my/3} = O(1)$ for $m = O(\frac{\Gamma^2}{\delta^2})$. If $\Delta \geq 1$, then

$$\Delta y = \frac{1}{(1+\delta)^2\Gamma^2} - y \geq \frac{\delta}{\Gamma^2}$$

and the applicable alternate form of the Chernoff bound

$$\Pr[\bar{y} > (1+\Delta)y] < e^{-\Delta my/3}$$

yields that $e^{-\Delta my/3} = O(1)$ for the same setting of m.

Since there are only $O(\frac{n(b+\log n)}{\delta})$ different values of α to consider, $m = O(\frac{\Gamma^2}{\delta^2}\log\frac{n(b+\log n)}{\delta})$ allows us to union bound over all the possible values of α. This shows that with constant probability, if we estimate T to have no Γ^2-outliers (in which case our algorithm might return T), then $(1+\delta)T$ has no $(1+\delta)^8\Gamma^2$-outliers. This implies (ii). □

We extend the analysis of lemmas 7 and 8 from a fixed direction to all directions and argue the correctness of the entire algorithm by proving theorem 3.

Proof of Theorem 3: Let S be the ellipsoid found by the deterministic algorithm (i.e. the outlier-free subset of points lies in this ellipsoid). Assume initially that S is full-dimensional. Rather than considering the original space, consider the transformed space where S is the unit sphere.

Consider the many directions w given by a δ'-grid in the unit cube, $\delta' = \frac{\delta}{6n}$. We form this grid by choosing every w such that the coordinates of w lie in $\{0, \delta', 2\delta', \ldots, 1\}$. By our choice of m, we can apply lemma 8 part (i) to each of these $(\frac{6n}{\delta})^n$ directions simultaneously and then union bound. Then with good probability, for every w in the δ'-grid, $\max\{(w^Tx)^2 : x \in T\} \geq \max\{(w^Tx)^2 : x \in S\}$ (i.e., in this direction T contains S). We now show that for an arbitrary direction w, $(1+\delta)T$ contains S.

Consider an arbitrary unit vector w. By rounding every coordinate of w up or down to an integer multiple of δ' we obtain a point on the δ'-grid. The set of all possible such roundings forms a box surrounding w, and some (not neccessarily unique) subset of n of these points, which we denote $\{w_i\}$, satisfy that w is in the convex cone of the $\{w_i\}$. Since w is a unit vector, each w_i has length $|w_i| \in (1 \pm \delta'\sqrt{n})$, and so $\hat{w}_i = w_i/|w_i|$ is within $2\delta'\sqrt{n}$ of w. Define $T(y)$ to be the distance to the boundary of T along the direction y. Since T is convex and $T(w_i) \geq 1$, the quantity $T(w)$ is lower bounded by the minimum distance of points on the convex hull of the $\{\hat{w}_i\}$ to the origin. Since w is within $2\delta'\sqrt{n}$ of each $\{\hat{w}_i\}$, so is the projection of w to their convex hull. Since the point on the convex hull is at most $2\delta'\sqrt{n}$ away from $\hat{w}_i$ for any i, $T(w) \geq 1 - 2\delta'\sqrt{n} \geq 1 - \delta/3$. Since S is within 1 of the origin everywhere, $(1+\delta)T$ contains S. This concludes the proof of (i).

Now consider (ii). Since $S \subset (1+\delta)T$, T is full-dimensional as well. For every w in our δ'-grid, we have that $(1+\delta)T$ is $(1+\delta)^8\Gamma^2$-outlier-free along w by lemma 8 part (ii). As before, consider the transformed space in which $(1+\delta)T$ is the unit sphere. Let $R = E(M_T)$ be the actual inertial ellipsoid of μ_T. Let w be an arbitrary unit vector and define $\{w_i\}$ as before. We have that $R(w_i) \geq \frac{1}{(1+\delta)^8\Gamma^2}$ and we reason as above that $R(w) \geq \frac{1-\delta/3}{(1+\delta)^8\Gamma^2} \geq \frac{1}{(1+\delta)^9\Gamma^2}$. Therefore $(1+\delta)T$ is $(1+\delta)^9\Gamma^2$-outlier-free.

We now remove the assumption that S is full-dimensional. Suppose S is not full-dimensional, but rather spans a subspace ζ. It suffices to consider $w \in \zeta$. For such a w, the projection of the associated $\{w_i\}$ to ζ yields $\{w'_i\}$ that are within $\delta'\sqrt{n}$ of the $\{w_i\}$ (because they don't move further than the distance to w upon projection). We can compare the $\{w'_i\}$ and w just as we did the $\{w_i\}$ and w previously. Because the max along w'_i is within a factor $(1-\delta/3)$ of the max along w_i, and the max along w'_i was lower bounded in lemma 8, the max along w_i is similarly lower bounded even though $w_i \notin \zeta$ (the change in the lower bound on the max is asymptotically negligible). Therefore we can apply lemma 8 part (i) to w_i. Thus $T(w_i) \geq 1 - \delta/3$, and so $T(w) \geq 1 - \frac{2\delta}{3}$. Thus $(1+\delta)T$ contains S. This establishes part (i).

We can extend the proof of part (ii) to the case that T is not full-dimensional in an identical manner. This concludes the proof of theorem 3. □

Corollary 1 (Running Time) *The algorithm runs in time* $\tilde{O}(\frac{b^2n^5}{\epsilon\delta^4})$.

Proof: We have from section 2.2 that $\beta = \gamma^2$ is at most $\tilde{O}(bn/\epsilon)$, and so we never need more than $m = \tilde{O}(\frac{bn^2}{\epsilon\delta^2})$ samples. Plugging in this value for m to our bounds from section 2.4.1 yields that the algorithm runs in time $\tilde{O}(\frac{b^2n^5}{\epsilon\delta^4})$, which is the bound we referred to in the introduction. In this time we achieve a $(1+\delta)^{O(1)} = 1+O(\delta)$ approximation to the optimal value of β. □

We now pose a related problem: Suppose that we are not given the parameter γ^2, but rather only ϵ, and asked to find the appropriate γ^2. Lemma 9 will show that we can at any point determine within a factor of $(1+\delta)$ how much of the probability mass is within a fixed ellipsoid. Since $\gamma^2 \in [1, \tilde{O}(\frac{bn}{\epsilon})]$, there are at most $\log_{(1+\delta)} \tilde{O}(\frac{bn}{\epsilon}) = O(\frac{\log(\frac{bn}{\epsilon})}{\delta})$ values of γ^2 to consider (with a loss of at most a factor of $(1+\delta)$ in the value we find for γ^2). Therefore we can simply try them all, estimating for each one whether this γ^2 requires us to throw away more than a $(1+\delta)\epsilon$ fraction of the distribution.

Thus, if the parameters (γ^2, ϵ) are achievable for the deterministic algorithm, and we are only given ϵ, we can find a subset of space T satisfying parameters $((1+O(\delta))\gamma^2, (1+O(\delta))\epsilon)$. Our asymptotic running time is still $\tilde{O}(\frac{b^2n^5}{\epsilon\delta^4})$.

Lemma 9 (Probability Mass Location) *Let E be an ellipsoid. Let our number of samples $\{x_i\}_{i=1}^m$ be $m = O(\frac{1}{\epsilon\delta^2})$. Then with constant probability, if we estimate a $(1+\delta)\epsilon$ fraction of our samples to be outside of E, at most a $(1+\delta)^2\epsilon$ fraction is outside of E, and at least an ϵ fraction is outside of E.*

Proof: Round E. Let Y_i be a random variable, $Y_i = 1$ iff $x_i^2 > 1$. Let $\bar{y} = \frac{1}{m}\sum Y_i$ and $y = \mathbf{E}[y_i]$. The event that we estimate a $(1+\delta)\epsilon$ fraction of the sample to be outside E

when less than an ϵ fraction truly lies outside E, is $y < \epsilon,\ \bar{y} \geq (1+\delta)\epsilon$. We can upper bound the probability of this event using the Chernoff bound

$$\Pr[\sum Y_i \geq m(1+\Delta)\mathbf{E}[Y_i]] \leq e^{-\Delta^2 m \mathbf{E}[Y_i]/3}$$

where $\Delta = \frac{(1+\delta)\epsilon}{y} - 1$. Then

$$\Delta^2 y = (\frac{(1+\delta)\epsilon}{y} - 1)((1+\delta)\epsilon - y) \geq (\frac{(1+\delta)\epsilon}{\epsilon} - 1)((1+\delta)\epsilon - \epsilon) = \delta^2\epsilon$$

and so the upper bound on the probability is constant for $m = O(\frac{1}{\epsilon\delta^2})$. If $\Delta \geq 1$, in which case the alternate form of the Chernoff bound is applicable, we find $\Delta y \geq \delta\epsilon$, and so the number of samples is still sufficient.

A similar calculation for the event that $y > (1+\delta)^2\epsilon,\ \bar{y} \leq (1+\delta)\epsilon$ using

$$\Pr[\sum Y_i \leq m(1-\Delta)\mathbf{E}[Y_i]] \leq e^{-\Delta^2 m \mathbf{E}[Y_i]/3}$$

involves setting $\Delta = 1 - \frac{(1+\delta)\epsilon}{y}$, which yields

$$\Delta^2 y = (y - (1+\delta)\epsilon)(1 - \frac{(1+\delta)\epsilon}{y}) \geq \delta^2\epsilon$$

and similarly for the alternate form of the Chernoff bound if $\Delta \geq 1$. Therefore the probability of significantly underestimating the amount of probability mass outside E is at most a constant for the same value of m. □

One consequence of the theorems in this section is that a sample of size $\tilde{O}(\frac{n^2 b}{\epsilon})$ is enough to estimate the inertial ellipsoid of any distribution on Z_b^n (after removing at most an ϵ fraction) and thus bring it into nearly isotropic position.

2.5 A Matching Lower Bound

We show that for any $\epsilon < 1/4$ there exists a distribution μ with support $\subset \mathcal{Z}_b^n$ such that, for any S satisfying $\mu(S) \geq 1-\epsilon$, there exists w such that

$$\max\{(w^T x)^2 : x \in S\} \geq \bar{\beta}\mathbf{E}[(w^T x)^2 : x \in S]\Pr[x \in S] \geq \frac{\bar{\beta}}{2}\mathbf{E}[(w^T x)^2 : x \in S]$$

where $\bar{\beta} = \Omega(\frac{n}{\epsilon}(b - \log\frac{1}{\epsilon}))$. Based on the comparison between our upper and lower bounds on β in the case that we can't throw out more than half the distribution

$$O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right) \quad vs. \quad \Omega\left(\frac{n}{\epsilon}(b - \log\frac{1}{\epsilon})\right)$$

we describe theorem 1 as asymptotically optimal.

We motivate the construction of the worst case distribution by constructing three simpler distributions, each of which proves a weaker lower bound. The strong lower bound will follow from examining a distribution that is a composite of the three distributions showing the weaker lower bounds.

To prove the first weak lower bound, let μ be the uniform distribution on the one-

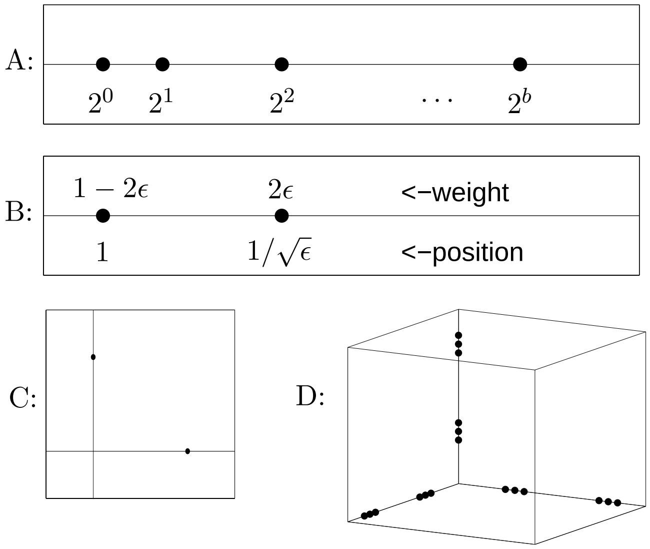

Figure 2-1: Lower Bound Constructions

dimensional points $\{2^0, 2^1, ...2^b\}$. An illustration of this μ is given in figure 2-1, part A. We claim that for any $\epsilon < \frac{1}{4}$, the best achievable (i.e. smallest) β satisfies $\beta = \Omega(b)$. The proof is simple: suppose the largest data point we keep is 2^k. Then (ignoring the factor w since we are in one dimension), $\max\{x^2 : x \in S\} = 2^{2k}$, while $\mathbf{E}[x^2 : x \in S] \leq \frac{2^0 + ... 2^{2k}}{(b+1)(1-\epsilon)} = O(\frac{2^{2k}}{b})$. Since $\beta = \frac{\max\{\cdot\}}{\mathbf{E}[\cdot]}$, we find $\beta = \Omega(b)$.

To prove the next weak lower bound, we construct a distribution as in figure 2-1, part B. Let μ be the probability distribution on one-dimensional points given by $\mu(1) = 1 - 2\epsilon, \mu(\frac{1}{\sqrt{\epsilon}}) = 2\epsilon$. Then for $\epsilon < \frac{1}{4}$, neither point can be thrown away. Thus $\max\{x^2 : x \in S\} = \frac{1}{\epsilon}$, while $\mathbf{E}[x^2 : x \in S] = 3 - 2\epsilon$, yielding $\beta = \Omega(\frac{1}{\epsilon})$.

For the third weak lower bound, we let μ be a distribution on n-dimensional space. In particular, let μ be the uniform distribution on n points, one on each coordinate axis, each one at unit distance from the origin, as illustrated in figure 2-1, part C. For $\epsilon < \frac{1}{2}$, we do not throw away any points on at least $n/2$ of the axes. Then for w a unit vector along one of the axes where the point is not thrown away, we have $\max\{(w^T x)^2 : x \in S\} = 1$, $\mathbf{E}[(w^T x)^2 : x \in S] \leq \frac{4}{n}$, and thus $\beta = \Omega(n)$.

The composite construction that we use to prove our strong lower bound in illustrated in figure 2-1, part D. We obtain the composite distribution by taking the distribution of part A, and making two copies that are weighted and translated as the two points are that compose the distribution of part B. We then place a copy of this new one-dimensional distribution along each axis, as in the distribution of part C. We now restate this construction formally and proceed to analyze it.

Fix n, ϵ and $b' = \frac{b}{2} - \frac{1}{4}\log\frac{1}{\epsilon}$. Let μ be a copy of the following distribution along each axis. Let there be $2b'$ points at distances

$$2^0, 2^1, \ldots, 2^{b'-1}, \frac{2^{b'}}{\sqrt{\epsilon}}, \frac{2^{b'+1}}{\sqrt{\epsilon}}, \ldots \frac{2^{2b'-1}}{\sqrt{\epsilon}}$$

and consider the distribution that places a $(1-2\epsilon)$ fraction of the probability mass uniformly on the first b' points and a 2ϵ fraction uniformly on the remaining b' points. This distribution satisifes that the maximum bit length along an axis is $\log \frac{2^{2b'}}{\sqrt{\epsilon}} = b$.

There are many ways of choosing a subset S of this distribution, but we can quickly restrict the set of possible choices. First we show that it never helps to treat the different axes asymmetrically. Suppose that this statement is not true. We begin by noting that for a distribution concentrated on the axes and fixed S, the vector w that maximizes

$$\frac{\max\{(w^T x)^2 : x \in S\}}{\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S]}$$

always occurs on an axis — to see this, note that the rounding transformation need only scale the axes, the maximizing w after rounding is in the direction of some point (i.e., along an axis), and therefore the maximizing w before rounding is also along an axis. Let μ_1 be a distribution concentrated on the axes and symmetric on each axis on which it is possible to throw out an ϵ fraction of the distribution and achieve parameter $\bar{\beta}$. Further suppose that this ϵ is the minimum ϵ such that this $\bar{\beta}$ is achievable, and the only S achieving $\bar{\beta}$ is asymmetric. Let axis i be an axis that this maximum outlier occurs on, and suppose that along axis i we throw out an ϵ_i fraction of the total distribution. If $\epsilon_i \leq \epsilon/n$, then let S' be the subset of μ_1 where we throw out the same points along every axis that we threw out along axis i in S. Then we have $\epsilon' = n\epsilon_i \leq \epsilon$, and yet S' achieves $\bar{\beta}$ along each axis, contradicting the assumption that there was no symmetric subset we could throw out achieving the same $(\epsilon, \bar{\beta})$. If $\epsilon_i > \epsilon/n$, then there is some other axis j such that along axis j we throw out an $\epsilon_j < \epsilon_i$ fraction of the probability distribution, but achieving $\bar{\beta}_j \leq \bar{\beta}$ along that axis (i.e. $\max\{x_j : x \in S\} \leq \bar{\beta}_j \mathbf{E}[x_j^2 : x \in S] \Pr[x \in S]$). Constructing S'' by taking S and replacing our choice of points to throw out along axis i with the points thrown out along axis j then yields a contradiction because $\epsilon'' < \epsilon$. Thus we can restrict our attention to S symmetric.

For any direction w along an axis, the projection onto w of any point on the other $n-1$ axes is 0, so we obtain

$$\mathbf{E}[(w^T x)^2] = \frac{1}{n}\mathbf{E}[x^2, \mu \text{ one-dimensional}]$$

We ignore the factor of n for the rest of the proof and restrict our attention to a single coordinate axis. Suppose the furthest point kept by S achieving parameters $(\epsilon, \bar{\beta})$ is the point with exponent k. By our choice of distribution, we cannot have thrown out more than half the points with a $\frac{1}{\sqrt{\epsilon}}$ factor, and so we have $\max\{x^2 : x \in S\} = \frac{2^{2k}}{\epsilon}$, $k > b'$. Calculating the expectation

$$\mathbf{E}[x^2 : x \in S] \Pr[x \in S] \leq \frac{1-2\epsilon}{2b'}(2^0 + 2^2 + \ldots + 2^{2b'-2}) + \frac{2\epsilon}{2b'}\frac{1}{\epsilon}(2^{2b'} + 2^{2b'+2} + \ldots + 2^{2k})$$

$$\leq \frac{2^{2b'-1}}{2b'} + \frac{2^{2k+1}}{b'} \leq \frac{2^{2k+2}}{b'}$$

yields that $\bar{\beta} = \frac{\max[\cdot]}{\mathbf{E}[\cdot]} = \frac{\max[\cdot]}{\mathbf{E}[\cdot]\Pr[\cdot]}\Pr[\cdot] \geq \frac{b'}{4\epsilon}(1-\epsilon) \geq \frac{b'}{8\epsilon}$ for the one-dimensional case. Thus our lower bound in the n-dimensional case is

$$\bar{\beta} \geq \frac{n}{16\epsilon}(b - \log\frac{1}{\epsilon})$$

2.6 An Approximation Algorithm

We showed earlier in the paper that for any distribution μ, and any ϵ we can achieve $\beta = O(\frac{n}{\epsilon}(b + \log \frac{n}{\epsilon}))$. A question that naturally arises is how well we can do on a particular distribution compared to the best possible on that particular distribution. Formally, given μ and ϵ, we seek S minimizing β subject to the constraints that

(i) $\mu(S) \geq 1 - \epsilon$

(ii) $\forall w, \max\{(w^T x)^2 : x \in S\} \leq \beta \mathbf{E}[(w^T x)^2 : x \in S]$

This is really a bicriteria approximation problem with parameters (β, ϵ). Note that in this case, we are looking for the *normalized* probability distribution to be β-outlier free. We show this problem to be NP-hard even for one-dimensional data by a reduction from the subset-sum problem. We then exhibit a $(\frac{1}{1-\epsilon}, 1)$-approximation algorithm for this task in the case that we are given the distribution explicitly. If we can only sample from the distribution μ, our algorithm yields a $(\frac{1}{1-\epsilon} + \delta, 1 + \delta)$-approximation for any constant $\delta > 0$ with high probability.

The subset-sum problem is: given $p_i \in (0,1), i \in \{1, ...n\}$, find I maximizing $\sum_{i \in I} p_i$ subject to the constraint that $\sum_{i \in I} p_i \leq 1$. To form a corresponding instance (μ, ϵ) of the outlier removal problem, let $P = \sum_i p_i, \epsilon = \frac{1}{2P}$, and let μ be given by

- a point at 1 with probability mass $\frac{1}{2}$
- $\forall i$, a point at 0 with probability mass $p'_i = \frac{p_i}{2P}$

Let S be a possible solution to this instance of the outlier removal problem. Since $P > 1$ (otherwise the subset-sum problem is trivial), the point at 1 cannot be removed, and hence $\max_{x \in S} = 1$. If we remove probability mass ϵ' of the points at 0, $\mathbf{E}[x^2 : x \in S] = \frac{(1)\frac{1}{2} + (0)(\frac{1}{2} - \epsilon')}{1 - \epsilon'} = \frac{1}{2 - 2\epsilon'}$. Thus the ratio $\frac{\max[\cdot]}{\mathbf{E}[\cdot]} = 2 - 2\epsilon'$, and minimizing this subject to $\epsilon' \leq \epsilon$ is exactly the problem of finding the optimal solution I to the subset-sum problem.

We now prove a lemma that enables the approximation result.

Lemma 10 (Preservation of Outliers) *Let μ be a distribution. Any β-outlier for μ is at least a $\beta(1-\epsilon)$-outlier with respect to any subset S satisfying $\mu(S) \geq 1 - \epsilon$.*

Proof: Let x be a β-outlier in the original distribution. Then for some w, $(w^T x)^2 > \beta \mathbf{E}[(w^T x)^2]$ For any S, we have $\mathbf{E}[(w^T x)^2 : x \in S] \Pr[x \in S] \leq \mathbf{E}[(w^T x)^2]$ and so x satisfies $(w^T x)^2 > \beta(1 - \epsilon)\mathbf{E}[(w^T x)^2 : x \in S]$ □

The approximation algorithm is simply either algorithm described in section 2.4, with error parameter δ in the case that we are sampling from μ. We could determine the optimal β for a fixed ϵ through a binary search. Suppose the value β_{OPT} is achievable by the restriction of μ to some S satisfying $\mu(S) \geq 1 - \epsilon$. Anytime our algorithm sees a point that is a β'-outlier with respect to the unnormalized distribution, $\beta' > \frac{\beta_{OPT}}{1-\epsilon}$, we know that this cannot be a $(\leq \beta_{OPT})$-outlier under any restriction of μ by lemma 10. So this point will have to be thrown out by the optimal solution. Thus running our algorithm with $\beta = \frac{\beta_{OPT}}{1-\epsilon}$ forces us to throw away no points that the optimal solution does not also throw away. This yields that we achieve a $\frac{1}{1-\epsilon}$-approximation in the case of an explicitly provided distribution. As before, the running time is $O(m^2 n)$ for $m > n$.

The outlier removal algorithm in fact finds an approximation to β for *every* ϵ in one pass. The algorithms of section 2.1 can be used to define an *outlier ordering* of a point set, namely, the first point that is an outlier, the second point, etc. Now to approximate the best possible β for a particular value of ϵ we simply remove the initial ϵ fraction of the points in the outlier ordering one at a time, and then look back to see the lowest value of β achieved by any $\epsilon' < \epsilon$.

2.7 Standard Deviations from the Mean

We prove a variant of our theorem that shows we can find a large subset of the original probability distribution where no point is too many standard deviations away from the mean.

Corollary 2 (Standard Deviations from the Mean) *Let μ be a probability distribution on $\mathcal{Z}_b^n$. Let S be a subset of space. Denote by $\mu(S)$ the probability that x chosen according to μ is in S. Let $\bar{x} = \mathbf{E}[x : x \in S]$ and $\sigma_w^2 = \mathbf{E}[(w^T(x-\bar{x}))^2 : x \in S]$. Then for every $\epsilon > 0$, there exists S and*

$$\beta = O\left(\frac{n}{\epsilon}(b + \log\frac{n}{\epsilon})\right)$$

such that
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{w^T(x-\bar{x}) : x \in S\} \leq \sqrt{\beta}\sigma_w$ for all $w \in \mathcal{R}^n$

Proof: The proof of the corollary is much like the proof of theorem 1. The appropriately modified outlier removal algorithm for constructing S is simply to translate the data set so that the origin coincides with the mean before each removal step. We can easily show that translating μ so that the origin coincides with the mean never decreases the volume of the dual ellipsoid of μ. We then explain how a variation on our potential function ϕ, and the upper and lower bounds on ϕ, imply that the modified algorithm does not throw out more than an ϵ fraction of the data set.

To analyze the volume of the dual ellipsoid, consider a fixed direction w and let $\frac{1}{r^2} = \mathbf{E}[(w^Tx)^2]$ (r is the length of the dual ellipsoid in this direction). If we translate the origin to a value z along w, then have $\frac{1}{r^2} = \mathbf{E}[(w^T(x-z))^2]$. Single variable calculus shows that the value maximizing r is $z = \mathbf{E}[w^Tx/|w|]$, which is just the mean. Thus translating our origin to $\bar{x}$ maximizes the length of the dual ellipsoid in every direction simultaneously. Thus the tradeoff between drop in probability mass and growth of the dual ellipsoid shown in lemmas 1 and 3 also holds for the modified algorithm.

To describe our modified ϕ, we need to define the *α-affine-core* of a distribution μ_S to be μ_T where $T \subset S$ is chosen to be maximum subject to the requirement that the affine hull of $\{\mu_T$ minus an α fraction of $\mu_T\}$ is not of lower dimension than the affine hull of μ_T for any choice of the α fraction. Under this definition, an appropriately modified version of lemma 4 is still true. Define $\phi'(\mu_S)$ to be an appropriately modified ϕ, $\phi'(\mu_S) = Vol(W(M_T))$ where μ_T is the α-affine-core of μ_S. We now explain how to derive upper and lower bounds on ϕ' analogous to lemma 5 in the case that the α-affine-core of μ_S is full-dimensional.

The lower bound is immediately implied by the argument above that translating the origin to the mean does not decrease the dual volume. To derive the upper bound, consider a set A of $n+1$ points $\{a_i\}$ whose affine hull is full-dimensional. In lemma 5, we argued that $Det(AA^T)$ was a positive integer, not zero by choice of A, and thus $Det(AA^T) \geq 1$.

Letting $\bar{a} = \frac{1}{n+1}\sum_i^{n+1} a_i$, we must lower bound $Det(\sum_i^{n+1}(a_i - \bar{a})(a_i - \bar{a})^T)$. Writing

$$(n+1)^{2n}Det(\sum_i^{n+1}(a_i - \bar{a})(a_i - \bar{a})^T) = Det(\sum_i^{n+1}((n+1)a_i - (n+1)\bar{a})((n+1)a_i - (n+1)\bar{a})^T)$$

we have that the second term is the positive non-zero determinant of an integer matrix, and hence the original determinant is at least $\frac{1}{(n+1)^{2n}}$. Because the origin corresponding to the mean of a set of points maximizes the dual volume, this bound holds for all possibilities for the origin. The upper bound on ϕ' is then $(\frac{n}{\alpha})^n(n+1)^{2n}f(n)$.

To prove a statement analogous to lemma 6 for the cumulative drop in ϕ', we revisit the construction of $\mu'_{T'}, \mu'_{S'}$ from μ_R, μ_S, μ_T. Define these objects just as in the proof of lemma 6. We have that $Det(M'_{S'}) \leq Det(M_R)$ when the origin is the mean of μ_R, and so $\phi'(\mu'_{S'}) \geq \phi'(\mu_R)$ to at least the same extent. We now calculate $\frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)}$. Letting the origin correspond to the mean of μ_T, we have $\phi'(\mu_T) = f(i)/Det(M_T)$ where $M_T = \sum_{y \in T} yy^T\mu(y)$. The mean of $\mu'_{T'}$ is given by $\bar{x} = \frac{\alpha x}{\alpha + \mu(T)}$. Then

$$M'_{T'} = \sum_{y \in T}(y - \bar{x})(y - \bar{x})^T\mu(y) + \alpha^2(x - \bar{x})(x - \bar{x})^T =$$

$$\left(\sum_{y \in T} yy^T\mu(y) - \sum_{y \in T}\bar{x}y^T\mu(y) - \sum_{y \in T} y\bar{x}^T\mu(y) + \bar{x}\bar{x}^T\mu(T)\right) + \alpha^2(x - \bar{x})(x - \bar{x})^T =$$

$$\sum_{y \in T} yy^T\mu(y) + \bar{x}\bar{x}^T\mu(T) + \alpha^2(x - \bar{x})(x - \bar{x})^T = \sum_{y \in T} yy^T\mu(y) + \frac{\mu(T)^2\alpha^2 + \mu(T)\alpha^2}{(\mu(T) + \alpha)^2}xx^T$$

Performing the same analysis using Gaussian elimination as we did previously and then computing the ratio yields

$$\frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)} = \frac{f(i+1)}{f(i)}\frac{(\mu(T) + \alpha)^2}{\mu(T)\alpha^2(1 + \mu(T))x[i+1]^2}$$

$$\Rightarrow \frac{\phi'(\mu'_{T'})}{\phi'(\mu_T)}\frac{\phi'(\mu_S)}{\phi'(\mu'_{S'})} = \frac{(\mu(T) + \alpha)^2}{\mu(T)\alpha^2(1 + \mu(T))}\frac{\mu(S)\alpha^2(1 + \mu(S))}{(\mu(S) + \alpha)^2}$$

We will now assume that we never remove more than an ϵ fraction of the probability mass. This is not circular reasoning — just as in the proof of theorem 2 using algorithm 2, the upper bound on ϕ' under this assumption will imply that we never remove more than an $\epsilon/2$ fraction of the probability mass, and since we never remove more than an $\epsilon/2$ fraction on any one step, the assumption will always hold. Using this assumption, we calculate

$$\frac{(\mu(T) + \alpha)^2}{(\mu(S) + \alpha)^2} \geq \frac{(1 - \epsilon)^2}{1^2} \geq \frac{1}{4}, \quad \frac{\mu(S)\alpha^2(1 + \mu(S))}{\mu(T)\alpha^2(1 + \mu(T))} \geq 1$$

Multiplying these factors together over the at most n steps in the iterative construction yields an additional cumulative factor of at most 2^{2n}, which is negligible. Combining this bit of additional slack with the new bound on ϕ' in the full dimensional case and the possibility that we only have an (α/n)-affine-core (as at the end of the proof of lemma 6), we finally arrive at a bound on the total cumulative drop in ϕ' of

$$2^{n(b+3\log\frac{n}{\alpha}+3)}$$

This immediately implies the claimed value for β in corollary 2. □

We now show that the $\frac{1}{1-\epsilon}$-approximation algorithm of section 2.6 naturally extends to a $\left(\frac{1-\epsilon}{1-3\epsilon}\right)$-approximation algorithm in the setting where we measure outlierness with respect to the mean, rather than a fixed origin. To establish this, it suffices to prove the following analogue of lemma 10.

Lemma 11 (Outlier Preservation Variant) *Let μ be a distribution. As in Corollary 2, measure outlierness by squared distance from the mean rather than from a fixed origin. Suppose x_0 is a β-outlier for μ, and no other point is a β'-outlier for $\beta' > \beta$. Then x_0 is at least a $\beta\frac{1-3\epsilon}{1-\epsilon}$-outlier with respect to any subset S satisfying $\mu(S) \geq 1-\epsilon$.*

Proof: As in the proof of lemma 10, consider a unit vector w such that $(w^T x_0)^2 > \beta\mathbf{E}[(w^T x)^2]$, and let $\beta = \gamma^2$. The difference between this bound and the bound of lemma 10 will result from the mean possibly moving closer to x_0 after removing other points $\{x_i\}$. Without loss of generality, let the mean of μ be the origin, and let $\mathbf{E}[(w^T x)^2] = 1$.

Suppose that to reach S we remove points $\{x_i\}$ of total probabilty mass $\epsilon' \leq \epsilon$. Then

$$\mathbf{E}[(w^T x)^2 : x \in S]\Pr[x \in S] = 1 - \sum_i (w^T x_i)^2 \mu(x_i)$$

$$\Rightarrow \mathbf{E}[(w^T x)^2 : x \in S] = (1 - \sum_i (w^T x_i)^2 \mu(x_i))/(1-\epsilon')$$

We calculate the new mean as

$$\mathbf{E}[(w^T x) : x \in S]\Pr[x \in S] = 0 - \sum_i (w^T x_i)\mu(x_i)$$

$$\Rightarrow \mathbf{E}[(w^T x) : x \in S] = (0 - \sum_i (w^T x_i)\mu(x_i))/(1-\epsilon')$$

Therefore the new distance of x_0 to the mean is $\left(\gamma - (0 - \sum_i (w^T x_i)\mu(x_i))/(1-\epsilon')\right)$. We calculate

$$\gamma'^2 = \frac{\text{distance}^2}{\text{variance}} = \frac{\left(\gamma + \frac{\sum_i (w^T x_i)\mu(x_i)}{1-\epsilon'}\right)^2}{\left(\frac{1-\sum_i (w^T x_i)^2\mu(x_i)}{1-\epsilon'}\right)} = \frac{((1-\epsilon')\gamma + \sum_i (w^T x_i)\mu(x_i))^2}{(1-\epsilon')(1-\sum_i (w^T x_i)^2\mu(x_i))}$$

Let $\bar{x} = \frac{\sum w^T x_i \mu(x_i)}{\epsilon'}$, the average of the points. Then removing $\bar{x}$ of weight ϵ' changes the numerator by the same amount, and $\bar{x}^2\epsilon' \leq \sum (w^T x_i)^2 \mu(x_i)$, so the denominator cannot decrease. The derivation of $\bar{x}^2\epsilon \leq \sum x_i^2 \mu(x_i)$ follows from

$$(\sum \lambda_i x_i)^2 \leq \sum \lambda_i x_i^2, \quad \sum \lambda_i = 1, \quad \lambda_i \geq 0$$

which follows from

$$(\frac{1}{2}a + \frac{1}{2}b)^2 \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$$

along the same lines that fact 2 follows from fact 1 in section 2.10. Now we have shown we may consider removing only a single point $\bar{x}$ of weight ϵ' in order to lower bound ${\gamma'}^2$. We may view this as a constrained maximization problem over $\bar{x}$, with constraints $|\bar{x}| \leq \gamma$, and $\bar{x}^2\epsilon' \leq 1$. The expression for $f(\bar{x}) = {\gamma'}^2$ is

$$
{\gamma'}^2 = \frac{((1-\epsilon')\gamma + \epsilon'\bar{x})^2}{(1-\epsilon')(1-\epsilon'\bar{x}^2)}
$$

If the constraint $\bar{x}^2\epsilon' \leq 1$ were tight, then the variance of the distribution after removing $\bar{x}$ would be 0, which would imply ${\gamma'}^2 = 1$. If the constraint $|\bar{x}| \leq \gamma$ were tight, we would have

$$
{\gamma'}^2 = \frac{((1-\epsilon')\gamma - \gamma\epsilon')^2}{(1-\epsilon')(1-\gamma^2\epsilon')} = \gamma^2 \frac{(1-2\epsilon')^2}{1-\epsilon'} \frac{1}{1-\gamma^2\epsilon'} \geq \gamma^2 \left(\frac{1-2\epsilon'}{1-\epsilon'}\right)^2 \geq \gamma^2 \left(\frac{1-3\epsilon'}{1-\epsilon'}\right)
$$

If neither constraint is tight, we may solve the unconstrained optimization problem by setting $\frac{df(\bar{x})}{d\bar{x}} = 0$ to find the local maximum, and then evaluating $f(\bar{x})$ at this maximum.

$$
f(\bar{x}) = {\gamma'}^2 = \frac{1}{1-\epsilon} \frac{u(\bar{x})^2}{v(\bar{x})}
$$

$$
\begin{gathered}
\frac{df(\bar{x})}{d\bar{x}} = \frac{1}{1-\epsilon}\left(\frac{2u(\bar{x})u'(\bar{x})}{v(\bar{x})} - \frac{u(\bar{x})^2 v'(\bar{x})}{v(\bar{x})^2}\right) = 0 \quad \Rightarrow \\
2v(\bar{x})u'(\bar{x}) - u(\bar{x})v'(\bar{x}) = 0 \quad \Rightarrow \\
2(1-\epsilon'\bar{x}^2)(\epsilon') - ((1-\epsilon')\gamma + \epsilon'\bar{x})(-2\epsilon'\bar{x}) = 0 \quad \Rightarrow \\
(1-\epsilon'\bar{x}^2) + ((1-\epsilon')\gamma + \epsilon'\bar{x})\bar{x} = 0 \quad \Rightarrow \\
1 + \gamma\bar{x} - \epsilon'\gamma\bar{x} = 0 \quad \Rightarrow \quad \bar{x} = -\frac{1}{(1-\epsilon')\gamma}
\end{gathered}
$$

$$
\begin{gathered}
f(\bar{x}) = \frac{((1-\epsilon')\gamma - \frac{1}{(1-\epsilon')\gamma}\epsilon')^2}{(1-\epsilon')(1-\frac{1}{(1-\epsilon')^2\gamma^2}\epsilon')} = \frac{((1-\epsilon')^2\gamma^2 - \epsilon')^2}{(1-\epsilon')((1-\epsilon')^2\gamma^2 - \epsilon')} = \\
\gamma^2 \frac{(1-\epsilon')^2 - \frac{\epsilon'}{\gamma^2}}{(1-\epsilon')} \geq \gamma^2\left(\frac{1-3\epsilon'}{1-\epsilon'}\right)
\end{gathered}
$$

which proves the lemma. □

2.8 An Implementation

Let `X` be an $n \times m$ matrix whose columns are the points of our distribution. Let `m, n, beta, epsilon` be the values for m, n, β, ϵ, and let the boolean variable `done` indicate whether we are finished removing outliers. In the case that `X` is full dimensional throughout the algorithm (a common case), a complete implementation is given by the following matlab code:

```
%% requires X,m,epsilon,beta
done = 0
while(~done)
  done = 1
  M = X*X’/m
  Y = M^(-.5)*X %% Y is isotropic version of X
  for i = 1:m, %% remove current outliers
    if Y(:,i)’*Y(:,i) > beta,  X(:,i)=0, done = 0, end
  end
end
```

Adding in the code for `X` not full dimensional yields

```
%% requires X,m,n,epsilon,beta
done = 1
while(done)
  done = 0

  M = X*X’/m %% round X
  [V,D] = eig(M^(.5))
  P = zeros(n,n)
  for i=1:n,
    if(D(i,i) ~= 0)
      P = P + V(:,i)*V(:,i)’/D(i,i)
    end
  end
  Y = P*X %% Y is isotropic version of X lying in span(X)

  for i = 1:m, %% remove current outliers
    if Y(:,i)’*Y(:,i) > beta,  X(:,i)=0, done = 1, end
  end
end
```

As of the Spring of 2002, a java applet illustrating the outlier removal algorithm is available at
`http://theory.lcs.mit.edu/~jdunagan/`

2.9 Miscellanea

We present here some further thoughts on outliers that did not fit into the development of the material earlier in the thesis. We begin by showing that one manner of interpolating between the hypotheses on μ used in theorems 1 and 2 is not sufficient to establish a bound on β.

2.9.1 Interpolating Between Discrete and Arbitrary Support

Suppose the distribution μ has support in $B_R \setminus B_{r\sqrt{n}}$. We show that such a μ may still have unbounded β. The construction is in two dimensions.

Let $R = 2, r = 1/2$. Let μ be the uniform distribution over the set of points $\{(\pm 1, \pm 2^{-j})\}_{j=1}^{k}$ where k is a parameter. This μ is illustrated in figure 2-2. We have clearly satisfied the hypothesis on the support of μ.

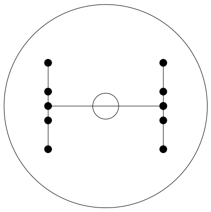

Figure 2-2: Weaker Support Counterexample

Considering the direction $w = (0, 1)$, the best possible β achievable without removing more than half the distribution is $\Omega(k)$, just as in the construction of distribution A in section 2.5. Since k is unbounded by the conditions on the support of the distribution, β is similarly unbounded.

2.9.2 Discrete Support: Rationals versus Integers

A natural question is whether theorem 1 can be extended to distributions whose support is the set of b-bit rationals. It is easy to see that the answer is yes for $\beta = \tilde{O}(\frac{n^2 b}{\epsilon})$. This follows from the fact that the absolute value of the determinant of any full-rank nxn matrix of b-bit rationals is at least $2^{-\tilde{O}(n^2 b)}$.

I am unaware of an example of such a matrix with non-zero determinant of absolute value less than $2^{-\tilde{O}(nb)}$. If this considerably tighter lower bound were to hold in general, it would straightforwardly imply that $\beta = \tilde{O}(\frac{nb}{\epsilon})$ is achievable for b-bit rationals as well. I hypothesize that if the weaker lower bound is correct, then there exists some distribution showing a similar lower bound on the best achievable β of $\tilde{O}(\frac{n^2 b}{\epsilon})$.

2.9.3 Some Further Thoughts on Optimization

The optimization version of our algorithm may be described as replacing the original objective function by a simpler objective function that is nonetheless close to the original. This raises the question of whether trying to optimize against the original objective function is something we would really want to do in practice, or whether the simpler objective function should be considered the true object of interest. For the remainder of this section, we will suppose that the original objective function (normalized β) is indeed the object of interest.

We do not fully understand the complexity of this optimization problem, but we present some observations that may be helpful for its future study. Our hardness reduction is from the problem subset-sum, which is known to admit a fully polynomial approximation scheme, yet we only present an approximation guarantee of $\frac{1}{1-\epsilon}$ in the fixed origin case, and an inferior guarantee in the case of standard deviations from the mean. Our current approximation algorithms can be seen to have worst-case behaviour given by the claimed ratios; the derivations of the approximation ratios give explicit constructions of distributions leading to this performance. We leave it as an open question whether a substantially better degree of approximation is possible.

In one-dimension, we can show that the case of unweighted points is polynomial-time solvable. This is because the best possible set of points to remove always consists of some part points closest to the origin and the rest points furthest from the origin. We can enumerate all such subsets in time $O(m^2)$. In n-dimensions, we do not know such a characterization even for the unweighted case. We leave the complexity of the unweighted version of the optimization problem in n-dimensions as another open question.

2.9.4 Robust Statistics

In robust statistics, the choice of the median as the quintessential robust statistic is commonly motivated by describing it as a “robust version of the mean.” In particular, it is noted that for any data set, the mean of the data set can be changed by an arbitrary amount simply by moving one of the data points to infinity. In contrast, the median does not “go to absurdity,” as the literature commonly puts it, until at least half of the data has been so changed by an adversary.

This criterion of robustness is clearly not enough: we also want the statistic to have some relationship to the data. Otherwise, the number 0 would always be a good robust statistic, as it is impervious to a malicious change to the entire data set.

Many robust statistics do not have a simple relationship to their non-robust predecessor. For example, the robust covariance measure of Donoho and Stahel[MY 95] weights the observed data points by a function of the median absolute deviation of all the data points from some origin before computing the traditional covariance of the newly weighted data.

We propose a methodology for constructing robust statistics that is, to the best of our knowledge, new: let the robust mean be the mean on an outlier-free subset of the data. Define other robust statistics similarly. One advantage of this approach is that it only takes a single sentence to define it, and the ease of implementation that we saw in section 2.8 is preserved. Exhaustively analyzing this statistic is not a goal of this thesis, but we present some further discussion here.

In the PAC model that has achieved such prominence in machine learning, it is asked only that a hypothesis be computed that is valid on at least a $1-\epsilon$ fraction of the data set in time polynomial in $\frac{1}{\epsilon}$. Because we cannot hope to know what happens on an unknown distribution with probability more than $1-\epsilon$ unless we draw more than $\frac{1}{\epsilon}$ samples, this is a very reasonable constraint. The assumption that our data does not consist of arbitrary reals, but is instead discretized, as we assumed in theorem 1, intuitively seems to take advantage of the fact that modern statistics relies on computers. As a rule of thumb, numbers on computers are represented by binary integers, or some variant thereof (i.e., floating point format).

These two observations provide some justification for our proposal to define a robust statistic as the statistic over an outlier-free subset of the data. A $1-\epsilon$ fraction of the data set is all we can hope to speak about confidently if we have only $\frac{1}{\epsilon}$ samples, and the β on this subset will be bounded under some conditions. We do not claim that the b-bit property holds for all data that is collected in practice, but we believe it suggests that a small value of β may commonly arise in practice. Also, not all robust statistics are efficiently computable, nor do they necessarily have polynomially bounded sample complexity — the general methodology we propose here preserves the ease with which the non-robust statistic could be computed.

We now give some background for our theorem on robust estimators. For a one-dimensional data set, define a δ-median to be a point such that at least a δ fraction of the data lies to the left of the point and at least a δ fraction to the right. In n-dimensions, call a point a δ-median if, for every direction w, it satisfies the definition of the one-dimensional δ-median under projection to w.

Using Helly's theorem, one can prove that $\frac{1}{n+1}$-medians exist for any n-dimensional data set (or distribution). Such a point is called a *centerpoint.* Centerpoints were proposed by Donoho and Gasko[DG 92] as a robust estimator for high-dimensional data. Donoho and Gasko showed centerpoints to have a *high breakdown point*, which is a technical criterion of "robustness" that we shall not discuss further here.

Teng et al [CEMST 93] gave the first polynomial time algorithm for computing an approximate center point (polynomial in d). Their algorithm produces $\Omega(\frac{1}{n^2})$-medians. We show that the algorithm of section 2.7 produces $\frac{1}{2\gamma^2(1-\epsilon)}$-medians. For a distribution on $\mathcal{Z}_b^n$, this yields $\tilde{\Omega}(\frac{1}{nb})$-medians.

Theorem 4 (A Robust Mean) *Let μ be a distribution, let $\bar{x} = \mathbf{E}[x : x \in S]$, and suppose S satisfies*
(i) $\mu(S) \geq 1 - \epsilon$
(ii) $\max\{(w^T(x-\bar{x}))^2 : x \in S\} \leq \gamma^2 \mathbf{E}[(w^T(x-\bar{x}))^2 : x \in S]$ for all $w \in \mathcal{R}^n$
Then $\bar{x}$ is a $\frac{1}{2\gamma^2(1-\epsilon)}$-median.

Proof: Suppose initially that $\mu(S) = 1$. Without loss of generality, consider a particular direction given by the unit vector w, and assume that $w^T\bar{x} = 0$ and $\mathbf{E}[(w^Tx)^2] = 1$. Since we are restricting our attention to w for the rest of the proof, we may define $y_i = w^Tx_i$. Let $\{y_i\}$ denote the distribution μ on S, and let I denote the index set. We partition I and define $\delta^{\pm}$ via

$$I^- = \{i : x_i < 0\} \qquad I^+ = \{i : x_i \geq 0\}$$
$$\delta^- = \sum_{i \in I^-} \mu(x_i) \qquad \delta^+ = \sum_{i \in I^+} \mu(x_i)$$

Then we have

$$\sum_{i \in I^-} x_i\mu(x_i) + \sum_{i \in I^+} x_i\mu(x_i) = 0 \qquad \sum_{i \in I} x_i^2\mu(x_i) = 1$$

Using that $|x_i| \leq \gamma$, we obtain

$$1 = \sum_{i \in I} x_i^2\mu(x_i) \leq \sum_{i \in I} \gamma|x_i|\mu(x_i) = \gamma(\sum_{i \in I^+} x_i\mu(x_i) - \sum_{i \in I^-} x_i\mu(x_i))$$
$$= \gamma(2\sum_{i \in I^+} x_i\mu(x_i)) \leq 2\gamma^2\delta^+$$

From this we conlude that $\delta^+ \geq \frac{1}{2\gamma^2}$, and similary for δ^-. Dropping the assumption that

$\mu(S) = 1$ turns our lower bound into $\frac{1}{2\gamma^2(1-\epsilon)}$. □

We speculate that there are more theorems to be proved about this method for constructing robust statistics, but we stop here for now.

2.10 Some Properties of Matrices

The proof in section 2.3 relied on fact 2, which we speculate to be well-known. We present the proof of this fact here since it uses techniques that are otherwise not necessary in the rest of section 2.3.

Fact 1 *For X, Y positive definite*

$$Det((X+Y)/2) \geq \sqrt{Det(X)Det(Y)}$$

Proof: This statement is equivalent to (clearing denominators and squaring twice)

$$Det(XY) \leq Det^2((X+Y)/2)$$

which is equivalent to

$$\begin{aligned}
1 &\leq \frac{Det^2((X+Y)/2)}{Det(XY)} \\
&= Det(\frac{1}{4}(X+Y))Det(X^{-1})Det(X+Y)Det(Y^{-1}) \\
&= Det(\frac{1}{4}(X+Y)(X^{-1})(X+Y)(Y^{-1})) \\
&= Det(\frac{1}{4}(I+YX^{-1})(XY^{-1}+I)) \\
&= Det(\frac{1}{4}(YX^{-1}+2I+XY^{-1})) \\
&= Det(\frac{1}{4}(A+2I+A^{-1}))
\end{aligned}$$

where we let $A = YX^{-1}$ at the very end. Also let $B = \frac{A+2I+A^{-1}}{4}$. We have reduced to the case of showing that $Det(B) \geq 1$. We will show the stronger claim that every eigenvalue of B is at least 1. Consider an arbitrary (eigenvector, eigenvalue)-pair of A, (e, λ). Then

$$Be = \frac{1}{4}(\lambda + 2 + \frac{1}{\lambda})e$$

Since $\frac{1}{4}(\lambda + 2 + \frac{1}{\lambda}) \geq 1$, we have that e is an eigenvector of eigenvalue at least 1 for B (this used that $\lambda \geq 0$, which is true since A is positive definite). Since the eigenvectors of A form an orthonormal basis of the whole space, all of B's eigenvectors are also eigenvectors of A. □

Fact 2 *For positive definite X_i and $\sum \lambda'_i = 1, \lambda'_i \geq 0$,*

$$Det(\sum_i \lambda'_i X_i) \geq \prod_i Det(X_i)^{\lambda'_i}$$

Proof: This is a straightforward generalization of fact 1.

Suppose first that for each i, λ'_i is exactly equal to $p_i/2^k$ for some integer p_i. In this case, we may apply fact 1 iteratively to find

$$Det(\sum_{i=1}^{2^k} X'_j) \geq \prod_{i=1}^{2^k} Det(X'_j)^{(1/2^k)}$$

Equating p_i of the $\{X'_j\}$ to X_i for each i, we recover fact 2 exactly. For general $\{\lambda'_i\}$, we have that the theorem must hold for any k-bit binary approximation to the λ'_i; fact 2 then follows from standard continuity arguments. □

Chapter 3

Perturbations

We begin by defining the notation that we will use in this chapter. We then define Renengar's condition number and the smoothed analysis model. We conclude by stating the main theorem to be proved in this chapter.

3.1 Notation, Definitions, Main Result

Throughout this chapter we use the notational convention that

- lower case letters such as a and α denote scalars,
- bold lower case letters such as $\boldsymbol{a}$ and $\boldsymbol{b}$ denote vectors,
- capital letters such as A denote matrices, and
- bold capital letters such as $\boldsymbol{C}$ denote convex sets.

If $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are column vectors, we let $[\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n]$ denote the matrix whose columns are the $\boldsymbol{a}_i$s.

For a vector $\boldsymbol{a}$, we let $\|\boldsymbol{a}\|$ denote the standard Euclidean norm of the vector. We will make frequent use of the Frobenius norm of a matrix, $\|A\|_F$, which is the square root of the sum of squares of the entries in the matrix. We extend this notation to let $\|A, \boldsymbol{a}\|_F$ denote the square root of the sum of squares of the entries in A and $\boldsymbol{a}$. Different choices of norm are possible; we use the Frobenius norm throughout this chapter. The following proposition relates several common choices of norm:

Proposition 1 (Choice of Norm) *For an n-by-d matrix A,*

$$\begin{aligned} \frac{\|A\|_F}{\sqrt{dn}} &\leq \|A\|_\infty \leq \|A\|_F\,, \text{ and} \\ \frac{\|A\|_F}{\sqrt{d}} &\leq \|A\|_{OP} \leq \|A\|_F\,, \end{aligned}$$

where $\|A\|_{OP}$ denotes the operator norm of A, $\max_{x \neq 0} \frac{\|Ax\|}{\|x\|}$.

We also make use of the following definitions:

Definition 4 (Ray) *For a vector $\boldsymbol{p}$, let* $\mathbf{Ray}\,(\boldsymbol{p})$ *denote* $\{\alpha \boldsymbol{p} : \alpha > 0\}$.

Definition 5 (Open Convex Cone) *An open convex cone is a convex set $\boldsymbol{C}$ such that for all $\boldsymbol{x} \in \boldsymbol{C}$ and all $\alpha > 0$, $\alpha\boldsymbol{x} \in \boldsymbol{C}$, and there exists a vector $\boldsymbol{t}$ such that $\boldsymbol{t}^T\boldsymbol{x} < 0$ for all $\boldsymbol{x} \in \boldsymbol{C}$.*

Warning 1 (Open Convex Cone?) *An open convex cone cannot contain the origin, and is not necessarily open in the topological sense.*

Definition 6 (Positive Half-Space) *For a vector $\boldsymbol{a}$ we let $\mathcal{H}(\boldsymbol{a})$ denote the half-space of points with non-negative inner product with $\boldsymbol{a}$.*

For example $\mathbb{R}^d$ and $\mathcal{H}(\boldsymbol{x})$ are not open convex cones, while $\{\boldsymbol{x} : \boldsymbol{x}_0 > 0\}$ and $\mathbf{Ray}\,(\boldsymbol{p})$ are open convex cones.

These definitions enable us to express the feasible $\boldsymbol{x}$ for the linear program

$$A\boldsymbol{x} \geq \mathbf{0} \text{ and } \boldsymbol{x} \in \boldsymbol{C}$$

as

$$\boldsymbol{C} \cap \bigcap_{i=1}^{n} \mathcal{H}(\boldsymbol{a}_i),$$

where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are the rows of A. Throughout this chapter, we will call a set *feasible* if it is non-empty, and *infeasible* if it is empty. Thus, we say that $\boldsymbol{C} \cap \bigcap_{i=1}^{n} \mathcal{H}(\boldsymbol{a}_i)$ is feasible if the corresponding linear program is feasible.

3.1.1 Definition of Condition Number for Linear Programming

For a feasible linear program of the form,

$$\max\ \boldsymbol{c}^T\boldsymbol{x} \quad s.t. \quad A\boldsymbol{x} \leq \boldsymbol{b} \tag{1}$$

we follow Renegar [Ren 94, Ren 95a, Ren 95b] in defining the primal condition number, C_P, of the program to be the normalized reciprocal of the distance to ill-posedness. A program is ill-posed if the program can be made both feasible and infeasible by arbitrarily small changes to the pair (A, b). The distance to ill-posedness of the pair (A, b) is the distance to the set of ill-posed programs under the Frobenius norm. We similarly define the dual condition number, C_D, to be the normalized reciprocal of the distance to ill-posedness of the dual program. The condition number, C_{PD}, is the maximum of C_P and C_D.

We can equivalently define the condition number without introducing the concept of ill-posedness. For programs of form (1), define $C_P^{(1)}(A, \boldsymbol{b})$ by

Definition 7 (Primal Distance to Ill-Posedness)

(a) if $A\boldsymbol{x} \leq \boldsymbol{b}$ is feasible,

$$C_P^{(1)}(A, \boldsymbol{b}) = \|A, \boldsymbol{b}\|_F \,/\, \sup\left\{\delta : \|\Delta A, \Delta\boldsymbol{b}\|_F \leq \delta \textit{ implies } (A + \Delta A)\boldsymbol{x} \leq (\boldsymbol{b} + \Delta\boldsymbol{b}) \textit{ is feasible}\right\},$$

(b) if $A\boldsymbol{x} \leq \boldsymbol{b}$ is infeasible,

$$C_P^{(1)}(A, \boldsymbol{b}) = \|A, \boldsymbol{b}\|_F \,/\, \sup\left\{\delta : \|\Delta A, \Delta\boldsymbol{b}\|_F \leq \delta \textit{ implies } (A + \Delta A)\boldsymbol{x} \leq (\boldsymbol{b} + \Delta\boldsymbol{b}) \textit{ is infeasible}\right\}$$

The dual of a program of form (1) is

$$\min\ \boldsymbol{b}^T\boldsymbol{y} \quad s.t. \quad A^T\boldsymbol{y} = \boldsymbol{c}, \quad \boldsymbol{y} \geq \boldsymbol{0},$$

and we define the dual condition number, $C_D^{(1)}(A, \boldsymbol{c})$, analogously.

Any linear program may be expressed in form (1); however, transformations among linear programming formulations do not in general (and commonly do not) preserve condition number [Ren 95a]. We will therefore have to define different condition numbers for each normal form we consider. For linear programs with canonical forms:

$$\begin{array}{rlr} \max\ \boldsymbol{c}^T\boldsymbol{x} \text{ s.t. } A\boldsymbol{x} \leq \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0} & \text{and its dual} \quad \min\ \boldsymbol{b}^T\boldsymbol{y} \text{ s.t. } A^T\boldsymbol{y} \leq \boldsymbol{c},\ \boldsymbol{y} \geq \boldsymbol{0} & (2) \\ \max\ \boldsymbol{c}^T\boldsymbol{x} \text{ s.t. } A\boldsymbol{x} = \boldsymbol{b},\ \boldsymbol{x} \geq \boldsymbol{0} & \text{and its dual} \quad \min\ \boldsymbol{b}^T\boldsymbol{y} \text{ s.t. } A^T\boldsymbol{y} \leq \boldsymbol{c} & (3) \\ \text{find } \boldsymbol{x} \neq \boldsymbol{0} \text{ s.t. } A\boldsymbol{x} \leq \boldsymbol{0} & \text{and its dual} \quad \text{find } \boldsymbol{y} \neq \boldsymbol{0} \text{ s.t. } A^T\boldsymbol{y} = \boldsymbol{0},\ \boldsymbol{y} \geq \boldsymbol{0} & (4) \end{array}$$

we define their condition numbers, $C_{PD}^{(2)}$, $C_{PD}^{(3)}$ and $C_{PD}^{(4)}$, analogously. We follow the convention that $\boldsymbol{0}$ is not considered a feasible solution to (4).

As we mentioned in the introduction, the condition numbers for numerous other problems (i.e., matrix inversion) are defined as the sensitivity of the output to perturbations in the input, and then shown to be equivalent to the distance to ill-posedness. Renegar inverts this scheme by defining the condition number for linear programming to be distance to ill-posedness, and then showing that the condition number does bound the sensitivity of the output to perturbations in the input [Ren 94, Ren 95a].

3.1.2 Smoothed Analysis of the Condition Number

Following [ST 01], we perform a smoothed analysis of these condition numbers. That is, we bound the distributions of these condition numbers for arbitrary programs under slight perturbations. We then derive bounds on the expectations of the logarithms of the condition numbers in terms of the size of the program and the magnitude of the perturbation.

For a linear program specified by $(\bar{A}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}})$, we consider the condition number of the program specified by $(A, \boldsymbol{b}, \boldsymbol{c})$, where A, $\boldsymbol{b}$, and $\boldsymbol{c}$ are a Gaussian random matrix and vectors of variance σ^2 centered at $\bar{A}$, $\bar{\boldsymbol{b}}$ and $\bar{\boldsymbol{c}}$ respectively. As the condition numbers are unchanged by multiplying all the data by a constant, we assume without loss of generality that in each input form the Frobenius norm of the data is at most 1. This also provides a scaling of the program so that σ measures the relative size of the random perturbation. For completeness, we recall needed facts about Gaussian random variables in section 3.8.1.

The following is the principal theorem of this chapter:

Theorem 5 (Smoothed Complexity of Renegar's Condition Number) *For every $\bar{A}$, $\bar{\boldsymbol{b}}$, and $\bar{\boldsymbol{c}}$ such that $\left\|\bar{A}, \bar{\boldsymbol{b}}, \bar{\boldsymbol{c}}\right\|_F \leq 1$, and for all $i \in \{1, 2, 3, 4\}$,*

$$\mathbf{Pr}_{A,\boldsymbol{b},\boldsymbol{c}}\left[C_{PD}^{(i)}(A, \boldsymbol{b}, \boldsymbol{c}) > \frac{2^{14}\ n^2 d^{3/2}}{\delta\sigma^2}\left(\log^2 \frac{2^{10}\ n^2 d^{3/2}}{\delta\sigma^2}\right)\right] \quad < \quad \delta$$

and hence

$$\mathbf{E}_{A,\boldsymbol{b},\boldsymbol{c}}\left[\log C_{PD}^{(i)}(A, \boldsymbol{b}, \boldsymbol{c})\right] \quad \leq \quad 21 + 3\log(nd/\sigma)$$

where A is a matrix and $\boldsymbol{b}$ and $\boldsymbol{c}$ are vectors of independent Gaussian random variables of variance σ^2, $\sigma^2 \leq 1/(nd)$, centered at $\bar{A}$, $\bar{\boldsymbol{b}}$, and $\bar{\boldsymbol{c}}$, respectively.

3.2 Primal Condition Number

In this section, we consider problems in conic form

$$\max \; \boldsymbol{c}^T \boldsymbol{x} \text{ such that } A\boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \boldsymbol{C},$$

where $\boldsymbol{C}$ is an open convex cone. Because $\boldsymbol{C}$ is an open convex cone, $\boldsymbol{0}$ cannot be a feasible solution of this program. The primal program of form (1) can be put into conic form with the introduction of the homogenizing variable x_0. Letting $\boldsymbol{C} = \{(\boldsymbol{x}, x_0) : x_0 > 0\}$, the primal program of form (1) is feasible if and only if

$$[-A, \boldsymbol{b}](\boldsymbol{x}, x_0) \geq \boldsymbol{0}, \; (\boldsymbol{x}, x_0) \in \boldsymbol{C}$$

is feasible. Similarly, the primal and dual programs of form (2) and the dual program of form (3) can also be put into conic form. In each case the transformation into conic form leaves the Frobenius norm unchanged. Also, a random Gaussian perturbation in the original form maps to a random Gaussian perturbation in the conic form.

The following is a generalization of the distance to ill-posedness that we will use throughout this section.

Definition 8 (Generalized Primal Distance to Ill-Posedness) *For an open convex cone, $\boldsymbol{C}$, and a matrix, A, we define $\rho(A, \boldsymbol{C})$ by*

a. if $A\boldsymbol{x} \geq 0$, $\boldsymbol{x} \in C$ is feasible, then

$$\rho(A, \boldsymbol{C}) = \sup\left\{\epsilon : \|\Delta A\|_F < \epsilon \text{ implies } (A + \Delta A)\boldsymbol{x} \geq \boldsymbol{0}, \; \boldsymbol{x} \in \boldsymbol{C} \text{ is feasible}\right\}$$

b. if $A\boldsymbol{x} \geq 0$, $\boldsymbol{x} \in C$ is infeasible, then

$$\rho(A, \boldsymbol{C}) = \sup\left\{\epsilon : \|\Delta A\|_F < \epsilon \text{ implies } (A + \Delta A)\boldsymbol{x} \geq \boldsymbol{0}, \; \boldsymbol{x} \in \boldsymbol{C} \text{ is infeasible}\right\}$$

We note that this definition makes sense even when A is a column vector. In this case, $\rho(\boldsymbol{a}, \boldsymbol{C})$ measures the distance to ill-posedness when we only allow perturbation to $\boldsymbol{a}$.

The primal program of form (4) is not quite in conic form; to handle it, we need

Definition 9 (Alternate Generalized Primal Distance to Ill-Posedness) *For a non-open convex cone, $\boldsymbol{C}$, and a matrix, A, we define $\rho(A, \boldsymbol{C})$ by*

a. if $A\boldsymbol{x} \geq 0$, $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{x} \in C$ is feasible, then

$$\rho(A, \boldsymbol{C}) = \sup\left\{\epsilon : \|\Delta A\|_F < \epsilon \text{ implies } (A + \Delta A)\boldsymbol{x} \geq \boldsymbol{0}, \; \boldsymbol{x} \neq \boldsymbol{0}, \; \boldsymbol{x} \in \boldsymbol{C} \text{ is feasible}\right\}$$

b. if $A\boldsymbol{x} \geq 0$, $\boldsymbol{x} \neq \boldsymbol{0}$, $\boldsymbol{x} \in C$ is infeasible, then

$$\rho(A, \boldsymbol{C}) = \sup\left\{\epsilon : \|\Delta A\|_F < \epsilon \text{ implies } (A + \Delta A)\boldsymbol{x} \geq \boldsymbol{0}, \; \boldsymbol{x} \neq \boldsymbol{0}, \; \boldsymbol{x} \in \boldsymbol{C} \text{ is infeasible}\right\}$$

This definition would allow us to prove the analog of lemma 12 for primal programs of form (4). We omit the details of this variation on the arguments in the interest of simplicity.

The main result of this section is:

Lemma 12 (Primal condition number is likely low) *For any open convex cone $\boldsymbol{C}$ and a Gaussian random matrix A of variance σ^2 centered at a matrix $\bar{A}$ satisfying $\left\|\bar{A}\right\|_F \leq 1$, for $\sigma \leq 1/\sqrt{nd}$, we have*

$$\mathbf{Pr}\left[\frac{\|A\|_F}{\rho(A, \boldsymbol{C})} \;\geq\; \frac{2^{12} n^2 d^{3/2}}{\delta \sigma^2} \log^2\left(\frac{2^9 n^2 d^{3/2}}{\delta \sigma^2}\right)\right] \;\leq\; \delta.$$

The analysis of C_P will proceed as follows: we consider the cases that the program is feasible and infeasible separately. In section 3.2.1, we show that it is unlikely that a program is feasible and yet can be made infeasible by a small change to its constraints (lemma 15). In section 3.2.2, we show that it is unlikely that a program is infeasible and yet can be made feasible by a small change to its constraints (lemma 20). In section 3.2.3, we combine these results to show that the primal condition number is low with high probability.

The thread of argument in these sections consists of a geometric characterization of those programs with poor condition number, and then a probabilistic argument demonstrating that this characterization is rarely satisfied. Throughout the proofs in this section, $\boldsymbol{C}$ will always refer to the original open cone, and a subscripted $\boldsymbol{C}$ (i.e., $\boldsymbol{C}_0$) will refer to a modification of this cone.

The key probabilistic tool used in the analysis is lemma 13, which was proved in [BD 02], and also in [Bal 93] and [BR 76]. We provide a proof in section 3.8.

Lemma 13 (ϵ-Boundaries are likely to be missed) *Let $\boldsymbol{K}$ be an arbitrary convex body, and let $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ denote the ϵ-boundary of $\boldsymbol{K}$; that is,*

$$\mathbf{bdry}(\boldsymbol{K}, \epsilon) = \left\{\boldsymbol{x} : \exists \boldsymbol{x}' \in \boldsymbol{K}, |\boldsymbol{x} - \boldsymbol{x}'| \leq \epsilon\right\} \setminus \boldsymbol{K}.$$

Let $\boldsymbol{x}$ be a d-dimensional Gaussian random vector with variance σ^2. Then,

$$\mathbf{Pr}\left[\boldsymbol{x} \in \mathbf{bdry}(\boldsymbol{K}, \epsilon)\right] \leq \frac{4\epsilon d^{1/4}}{\sigma}$$

In this section and the next, we use the following consequence of lemma 13 repeatedly.

Lemma 14 (Feasible likely quite feasible, single constraint) *Let $\boldsymbol{C}_0$ be any convex cone in $\mathbb{R}^d$ and let $\boldsymbol{a}$ be a Gaussian random vector of variance σ^2. Then,*

$$\mathbf{Pr}_{\boldsymbol{a}}\left[\rho(\boldsymbol{a}, \boldsymbol{C}_0) \leq \epsilon\right] \leq \left(\frac{4\epsilon d^{1/4}}{\sigma}\right).$$

Proof: Let $\boldsymbol{K}$ be the set of $\boldsymbol{a}$ for which $\boldsymbol{C}_0 \cap \mathcal{H}(\boldsymbol{a})$ is infeasible. Observe that $\rho(\boldsymbol{a}, \boldsymbol{C}_0)$ is exactly the distance from $\boldsymbol{a}$ to the boundary of $\boldsymbol{K}$. Since $\boldsymbol{K}$ is a convex cone, lemma 13 tells us that the probability that $\boldsymbol{a}$ has distance at most ϵ to the boundary of $\boldsymbol{K}$ is at most $\left(\frac{4\epsilon d^{1/4}}{\sigma}\right)$. □

3.2.1 Primal number, feasible case

In this subsection, we analyze the primal condition number in the feasible case, and prove:

Lemma 15 (Feasible is likely quite feasible, all constraints) *Let $\boldsymbol{C}$ be an open convex cone in $\mathbb{R}^d$ and let A be an n-by-d Gaussian random matrix of variance σ^2. Then,*

$$\mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C} \textit{ is feasible}) \textit{ and } (\rho(A, \boldsymbol{C}) \leq \epsilon)\right] \leq \left(\frac{4\epsilon n d^{5/4}}{\sigma}\right).$$

The remaining lemmas in this subsection establish a necessary geometric condition for ρ to be small. In the proof of lemma 15 at the end of this subsection, we use lemma 14 to show that this geometric condition is unlikely to be met.

Lemma 16 (Feasibility as a dot product) *For every vector $\boldsymbol{a}$ and every unit vector $\boldsymbol{p}$,*

$$\rho(\boldsymbol{a}, \mathbf{Ray}\,(\boldsymbol{p})) = \left|\boldsymbol{a}^T\boldsymbol{p}\right|$$

Proof: Since $\mathbf{Ray}\,(\boldsymbol{p}) \cap \mathcal{H}(\boldsymbol{a})$ is feasible if and only if $\mathbf{Ray}\,(-\boldsymbol{p}) \cap \mathcal{H}(\boldsymbol{a})$ is infeasible, it suffices to consider the case where $\mathbf{Ray}\,(\boldsymbol{p}) \cap \mathcal{H}(\boldsymbol{a})$ is feasible. In this case $\boldsymbol{a}^T\boldsymbol{p} \geq 0$. We first prove that $\rho(\boldsymbol{a}, \mathbf{Ray}\,(\boldsymbol{p})) \geq \boldsymbol{a}^T\boldsymbol{p}$. For every vector $\Delta\boldsymbol{a}$ of norm at most $\boldsymbol{a}^T\boldsymbol{p}$, we have

$$(\boldsymbol{a} + \Delta\boldsymbol{a})^T\boldsymbol{p} = \boldsymbol{a}^T\boldsymbol{p} + \Delta\boldsymbol{a}^T\boldsymbol{p} \geq \boldsymbol{a}^T\boldsymbol{p} - \|\Delta\boldsymbol{a}\| \geq 0.$$

Thus $\boldsymbol{p} \in \mathcal{H}(\boldsymbol{a} + \Delta\boldsymbol{a})$. As this holds for every $\Delta\boldsymbol{a}$ of norm at most $\boldsymbol{a}^T\boldsymbol{p}$, we have $\rho(\boldsymbol{a}, \mathbf{Ray}\,(\boldsymbol{p})) \geq \boldsymbol{a}^T\boldsymbol{p}$.

To show that $\rho(\boldsymbol{a}, \mathbf{Ray}\,(\boldsymbol{p})) \leq \boldsymbol{a}^T\boldsymbol{p}$, note that for any $\epsilon > 0$, setting $\Delta\boldsymbol{a} = -(\epsilon + \boldsymbol{a}^T\boldsymbol{p})\boldsymbol{p}$ yields

$$(\boldsymbol{a} + \Delta\boldsymbol{a})^T\boldsymbol{p} = \boldsymbol{a}^T\boldsymbol{p} + \Delta\boldsymbol{a}^T\boldsymbol{p} = \boldsymbol{a}^T\boldsymbol{p} - (\epsilon + \boldsymbol{a}^T\boldsymbol{p})\boldsymbol{p}^T\boldsymbol{p} = \boldsymbol{a}^T\boldsymbol{p} - (\epsilon + \boldsymbol{a}^T\boldsymbol{p}) = -\epsilon,$$

so $\mathbf{Ray}\,(\boldsymbol{p}) \cap \mathcal{H}(\boldsymbol{a} + \Delta\boldsymbol{a})$ is infeasible. As this holds for every $\epsilon > 0$, $\rho(\boldsymbol{a}, \mathbf{Ray}\,(\boldsymbol{p})) \leq \boldsymbol{a}^T\boldsymbol{p}$. □

Lemma 17 (Quite feasible region implies quite feasible point, single constraint) *For every $\boldsymbol{a}$ and every open convex cone $\boldsymbol{C}_0$ for which $\boldsymbol{C}_0 \cap \mathcal{H}(\boldsymbol{a})$ is feasible,*

$$\rho(\boldsymbol{a}, \boldsymbol{C}_0) = \max_{\boldsymbol{p} \in \boldsymbol{C}_0 : \|\boldsymbol{p}\| = 1} \boldsymbol{a}^T\boldsymbol{p}.$$

Proof: The "$\geq$" direction is obvious from lemma 16, so we concentrate on showing

$$\rho(\boldsymbol{a}, \boldsymbol{C}_0) \leq \max_{\boldsymbol{p} \in \boldsymbol{C}_0 : \|\boldsymbol{p}\| = 1} \boldsymbol{a}^T\boldsymbol{p}.$$

As $\boldsymbol{C}_0$ is open, there exists a vector $\boldsymbol{t}$ such that $\boldsymbol{t}^T\boldsymbol{x} < 0$ for each $\boldsymbol{x} \in \boldsymbol{C}_0$. If $\boldsymbol{a} \in \boldsymbol{C}_0$, then

$$\max_{\boldsymbol{p} \in \boldsymbol{C}_0 : \|\boldsymbol{p}\| = 1} \boldsymbol{a}^T\boldsymbol{p} = \|\boldsymbol{a}\|.$$

For every $\epsilon > 0$, $\boldsymbol{C}_0 \cap \mathcal{H}(\boldsymbol{a} - (\boldsymbol{a} + \epsilon\boldsymbol{t}))$ is infeasible; so $\rho(\boldsymbol{a}, \boldsymbol{C}_0) \leq \|\boldsymbol{a}\|$.

If $\boldsymbol{a} \notin \boldsymbol{C}_0$, let $\boldsymbol{q}$ be the point of $\boldsymbol{C}_0$ that is closest to $\boldsymbol{a}$. As $\boldsymbol{C}_0 \cap \mathcal{H}(\boldsymbol{a})$ is feasible, $\boldsymbol{q}$ is not the origin and we can define $\boldsymbol{p} = \boldsymbol{q}/\|\boldsymbol{q}\|$. As $\boldsymbol{C}_0$ is a cone, $\boldsymbol{q}$ is perpendicular to $\boldsymbol{a} - \boldsymbol{q}$.

Thus, the distance from $\boldsymbol{a}$ to $\boldsymbol{q}$ is the square root of $\|\boldsymbol{a}\|^2 - (\boldsymbol{a}^T\boldsymbol{p})^2$, and $\boldsymbol{p}$ must be the point of unit norm maximizing $\boldsymbol{a}^T\boldsymbol{p}$.

As $\boldsymbol{C}_0$ is convex, there is a plane through $\boldsymbol{q}$ separating $\boldsymbol{C}_0$ from $\boldsymbol{a}$ and perpendicular to the line segment $\boldsymbol{a} - \boldsymbol{q}$. Thus, every point of $\boldsymbol{C}_0$ has inner product at most zero with the vector $\boldsymbol{a} - \boldsymbol{q}$; and hence, for every $\epsilon > 0$, $\boldsymbol{C}_0 \cap \mathcal{H}(\boldsymbol{a} - \boldsymbol{q} + \epsilon\boldsymbol{t})$ is infeasible. To conclude the proof, we note that $\|\boldsymbol{q}\| = \boldsymbol{a}^T\boldsymbol{p}$. □

Lemma 18 (Quite feasible point for each constraint implies quite feasible point for all constraints) *If there exist vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ and unit vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \in \boldsymbol{C}_0$, $\boldsymbol{C}_0 \subset \mathbb{R}^d$, such that*

$$\begin{aligned} \boldsymbol{a}_i^T\boldsymbol{p}_i &\geq \epsilon, \text{ for all } i, \text{ and} \\ \boldsymbol{a}_i^T\boldsymbol{p}_j &\geq 0, \text{ for all } i \text{ and } j, \end{aligned}$$

then there exists a point $\boldsymbol{p}$ of unit norm, $\boldsymbol{p} \in \boldsymbol{C}_0$, such that

$$\boldsymbol{a}_i^T\boldsymbol{p} \geq \epsilon/d, \text{ for all } i.$$

Proof: We prove this using Helly's theorem [GDK 63]. Let $\boldsymbol{S}_i = \{\boldsymbol{x} \in \boldsymbol{C}_0 : \boldsymbol{a}_i^T\boldsymbol{x}/\|\boldsymbol{x}\| \geq \epsilon/d\}$. As $\boldsymbol{C}_0$ is open, there exists $\boldsymbol{t}$ such that $\boldsymbol{t}^T\boldsymbol{x} < 0$, $\forall \boldsymbol{x} \in \boldsymbol{C}_0$. Let $\boldsymbol{S}'_i = \boldsymbol{S}_i \bigcap \{\boldsymbol{x} : \boldsymbol{t}^T\boldsymbol{x} = -1\}$. The $\{\boldsymbol{S}'_i\}$ have similar intersection to the $\{\boldsymbol{S}_i\}$ in that $\boldsymbol{x} \in \boldsymbol{S}'_i \Rightarrow \boldsymbol{x} \in \boldsymbol{S}_i$ and $\boldsymbol{x} \in \boldsymbol{S}_i \Rightarrow \boldsymbol{x}/\boldsymbol{t}^T\boldsymbol{x} \in \boldsymbol{S}'_i$. However, the $\{\boldsymbol{S}'_i\}$ are convex sets in a $(d-1)$-dimensional subspace. By Helly's theorem, if every subcollection of d of the $\{\boldsymbol{S}'_i\}$ has a common point, then the entire collection has a common point. Because the $\{\boldsymbol{S}_i\}$ have similar intersection to the $\{\boldsymbol{S}'_i\}$, the same statement holds for the $\{\boldsymbol{S}_i\}$. So assume $n = d$.

Let $\boldsymbol{p} = \sum_{i=1}^d \boldsymbol{p}_i/d$. Then, for each $1 \leq j \leq d$,

$$\boldsymbol{a}_j^T\boldsymbol{p} = \boldsymbol{a}_j^T\left(\sum_{i=1}^d \boldsymbol{p}_i/d\right) \geq \boldsymbol{a}_j^T\left(\boldsymbol{p}_j/d\right) \geq \epsilon/d.$$

Moreover, $\boldsymbol{p}$ has norm at most one, so $\boldsymbol{p}/\|\boldsymbol{p}\|$ is a point that lies in each of $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_d$. □

In section 3.8 we discuss how lemma 18 can also be shown using the Brunn-Minkowski theory of convex bodies, as it was in [BD 02]. My thanks to co-author Shang-Hua Teng for this beautiful use of Helly's Theorem.

Lemma 19 (Quite feasible point for all constraints implies quite feasible program) *For every set of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ and $\boldsymbol{p}$ such that $\mathbf{Ray}(\boldsymbol{p}) \cap \bigcap_i \mathcal{H}(\boldsymbol{a}_i)$ is feasible,*

$$\rho([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \mathbf{Ray}(\boldsymbol{p})) = \min_i \rho(\boldsymbol{a}_i, \mathbf{Ray}(\boldsymbol{p})).$$

Proof: It suffices to observe that $\mathbf{Ray}(\boldsymbol{p}) \cap \bigcap_i \mathcal{H}(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i)$ is feasible if and only if $\boldsymbol{p}^T(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i) \geq 0$ for all i. □

We now prove the main result of this section.

Proof of Lemma 15: Let $\boldsymbol{C}_0 = \boldsymbol{C} \cap \bigcap_i \mathcal{H}(\boldsymbol{a}_i)$ and $\boldsymbol{C}_i = \boldsymbol{C} \cap \bigcap_{j \neq i} \mathcal{H}(\boldsymbol{a}_j)$. Note that

$$\{\boldsymbol{x} : A\boldsymbol{x} \geq 0, \ \boldsymbol{x} \in \boldsymbol{C}\} = \boldsymbol{C}_0.$$

Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be the columns of A. Our first step will be to show that if $\boldsymbol{C}_0$ is feasible, then

$$\rho([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \boldsymbol{C}) \leq \epsilon/d$$

implies that there exists an i for which

$$\rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon.$$

To show this, we prove the contrapositive. That is, we assume $\boldsymbol{C}_0$ is feasible and that $\rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \geq \epsilon$ for all i. Then, lemma 17 implies that there exist unit vectors $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n \in \boldsymbol{C}_0$ such that $\boldsymbol{a}_i^T \boldsymbol{p}_i \geq \epsilon$. Applying lemma 18, we find a unit vector $\boldsymbol{p} \in \boldsymbol{C}_0$ such that $\boldsymbol{a}_i^T \boldsymbol{p} \geq \epsilon/d$ for all i. From lemmas 16 and 19, we then compute

$$\rho([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \boldsymbol{C}) \geq \rho([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \mathbf{Ray}\,(\boldsymbol{p})) = \min_i \rho(\boldsymbol{a}_i, \mathbf{Ray}\,(\boldsymbol{p})) \geq \epsilon/d.$$

Thus, we now know

$$\begin{aligned} &\mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C} \text{ is feasible}) \text{ and } (\rho(A, \boldsymbol{C}) \leq \epsilon/d)\right] \\ \leq\ & \mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n}\left[\boldsymbol{C}_0 \text{ is feasible and } \exists i : \rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon\right]. \end{aligned}$$

To bound the latter probability, we use lemma 14, which tells us that

$$\mathbf{Pr}_{\boldsymbol{a}_i}\left[\rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon \text{ and } \boldsymbol{C}_i \text{ is feasible}\right] \leq \left(\frac{4\epsilon d^{1/4}}{\sigma}\right).$$

Applying a union bound and the fact that $\boldsymbol{C}_0$ feasible implies $\boldsymbol{C}_i$ is feasible, we compute

$$\begin{aligned} &\mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n}\left[\boldsymbol{C}_0 \text{ is feasible and } \exists i : \rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon\right] && \leq \\ &\sum_i \mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n}\left[\boldsymbol{C}_0 \text{ is feasible and } \rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon\right] && \leq \\ &\sum_i \mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_n}\left[\boldsymbol{C}_i \text{ is feasible and } \rho(\boldsymbol{a}_i, \boldsymbol{C}_i) \leq \epsilon\right] && \leq \left(\frac{4\epsilon n d^{1/4}}{\sigma}\right). \end{aligned} \tag{3.1}$$

Setting $\epsilon = d\epsilon'$ yields the lemma as stated. □

This concludes the analysis that it is unlikely that the primal program is both feasible and has small distance to ill-posedness. Next, we show that it is unlikely that the primal program is both infeasible and has small distance to ill-posedness.

3.2.2 Primal number, infeasible case

The main result of this subsection is:

Lemma 20 (Infeasible is likely quite infeasible) *Let $\boldsymbol{C}$ be an open convex cone in $\mathbb{R}^d$ and let A be a Gaussian random matrix of variance σ^2 centered at a matrix $\bar{A}$ satisfying $\|\bar{A}\|_F \leq 1$, where $\sigma \leq 1/\sqrt{d}$. Then,*

$$\mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C} \textit{ is infeasible}) \textit{ and } (\rho(A, \boldsymbol{C}) \leq \epsilon)\right] \leq \left(\frac{360\ \epsilon n^2 d^{3/2} \lceil \log^{1.5}(1/\epsilon) \rceil}{\sigma^2}\right).$$

To prove lemma 20, we consider adding the constraints one at a time. If the program is infeasible in the end, then there must be some next constraint that takes it from being feasible to being infeasible. Lemma 21 gives a sufficient geometric characterization for the program to be quite infeasible when the next constraint is added, and in the proof of lemma 20, we show that this characterization is met with good probability. The geometric characterization is that the program is quite feasible before the next constraint is added and every previously feasible point is far from being feasible for the next constraint.

Lemma 21 (The feasible to infeasible transition) *Let $\boldsymbol{C}$ be an open convex cone, $\boldsymbol{p}$ be a unit vector, $\boldsymbol{p} \in \boldsymbol{C}$, and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k+1}$ be vectors such that*

$$\begin{array}{rcl} \boldsymbol{a}_i^T \boldsymbol{p} & \geq & \alpha \text{ for } 1 \leq i \leq k, \text{ and} \\ \boldsymbol{a}_{k+1}^T \boldsymbol{x} & \leq & -\beta \text{ for all } \boldsymbol{x} \in \boldsymbol{C} \cap \bigcap_{i=1}^{k} \mathcal{H}(\boldsymbol{a}_i), \ \ \|\boldsymbol{x}\| = 1. \end{array}$$

Then,

$$\rho(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k+1}, \boldsymbol{C}) \geq \min \left\{ \frac{\alpha}{2}, \frac{\alpha\beta}{4\alpha + 2\|\boldsymbol{a}_{k+1}\|} \right\}.$$

Proof: We will prove this by showing that for all ϵ satisfying

$$\begin{array}{rcl} \epsilon & \leq & \alpha/2, \text{ and} \\ \epsilon & < & \dfrac{\beta}{4 + 2\|\boldsymbol{a}_{k+1}\|/\alpha}, \end{array} \qquad \begin{array}{r} (3.2) \\ (3.3) \end{array}$$

and $\{\Delta\boldsymbol{a}_1, \ldots, \Delta\boldsymbol{a}_{k+1}\}$ satisfying $\|\Delta\boldsymbol{a}_i\| < \epsilon$ for $1 \leq i \leq k+1$, we have

$$\boldsymbol{C} \cap \bigcap_{i=1}^{k+1} \mathcal{H}(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i) \text{ is infeasible.}$$

Assume by way of contradiction that

$$\boldsymbol{C} \cap \bigcap_{i=1}^{k+1} \mathcal{H}(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i) \text{ is feasible.}$$

Then, there exists a unit vector $\boldsymbol{x}' \in \boldsymbol{C} \cap \bigcap_{i=1}^{k+1} \mathcal{H}(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i)$. We first show that

$$\boldsymbol{x}' + \frac{\epsilon}{\alpha}\boldsymbol{p} \in \boldsymbol{C} \cap \bigcap_{i=1}^{k} \mathcal{H}(\boldsymbol{a}_i).$$

To see this consider any $i \leq k$ and note that

$$(\boldsymbol{a}_i + \Delta\boldsymbol{a}_i)^T \boldsymbol{x}' \geq 0 \implies \boldsymbol{a}_i^T \boldsymbol{x}' \geq -\Delta\boldsymbol{a}_i^T \boldsymbol{x}' \geq -\|\Delta\boldsymbol{a}_i\| \|\boldsymbol{x}'\| \geq -\epsilon.$$

Thus,

$$\boldsymbol{a}_i^T \left(\boldsymbol{x}' + \frac{\epsilon}{\alpha}\boldsymbol{p}\right) \geq \boldsymbol{a}_i^T \boldsymbol{x}' + \boldsymbol{a}_i^T \frac{\epsilon}{\alpha}\boldsymbol{p} \geq -\epsilon + \frac{\epsilon}{\alpha}\alpha \geq 0.$$

Also,

$$\boldsymbol{x}' \in \boldsymbol{C}, \quad \boldsymbol{p} \in \boldsymbol{C} \quad \Longrightarrow \quad \boldsymbol{x}' + \frac{\epsilon}{\alpha}\boldsymbol{p} \in \boldsymbol{C}$$

Let $\boldsymbol{x} = \boldsymbol{x}' + \frac{\epsilon}{\alpha}\boldsymbol{p}$. Then $\boldsymbol{x} \in \boldsymbol{C} \cap \cap_{i=1}^{k} \mathcal{H}(\boldsymbol{a}_i)$ and $\boldsymbol{x}$ has norm at most $1 + \epsilon/\alpha$ and at least $1 - \epsilon/\alpha$. To derive a contradiction, we now compute

$$\begin{aligned}
(\boldsymbol{a}_{k+1} + \Delta\boldsymbol{a}_{k+1})^T \boldsymbol{x}' &= (\boldsymbol{a}_{k+1} + \Delta\boldsymbol{a}_{k+1})^T (\boldsymbol{x} - (\epsilon/\alpha)\boldsymbol{p}) \\
&= \boldsymbol{a}_{k+1}^T \boldsymbol{x} + \Delta\boldsymbol{a}_{k+1}^T \boldsymbol{x} - (\epsilon/\alpha)\boldsymbol{a}_{k+1}^T \boldsymbol{p} - (\epsilon/\alpha)\Delta\boldsymbol{a}_{k+1}^T \boldsymbol{p} \\
&\leq -\beta \left\| \boldsymbol{x} \right\| + \left\| \Delta\boldsymbol{a}_{k+1} \right\| \left\| \boldsymbol{x} \right\| + (\epsilon/\alpha) \left\| \boldsymbol{a}_{k+1} \right\| + (\epsilon/\alpha) \left\| \Delta\boldsymbol{a}_{k+1} \right\| \\
&\leq -\beta(1 - \epsilon/\alpha) + \epsilon(1 + \epsilon/\alpha) + (\epsilon/\alpha) \left\| \boldsymbol{a}_{k+1} \right\| + (\epsilon^2/\alpha) \\
&= -\beta(1 - \epsilon/\alpha) + \epsilon \left((1 + \epsilon/\alpha) + \left\| \boldsymbol{a}_{k+1} \right\| /\alpha + \epsilon/\alpha\right) \\
&\leq -\beta/2 + \epsilon \left((2 + \left\| \boldsymbol{a}_{k+1} \right\| /\alpha\right), \text{ by (3.2)} \\
&< 0 \text{ by (3.3).}
\end{aligned}$$

□

The next two items are trivial.

Proposition 2 *For positive α, β and any vector $\boldsymbol{a}_{k+1}$,*

$$\frac{\alpha\beta}{2\alpha + \left\| \boldsymbol{a}_{k+1} \right\|} \geq \min \left\{ \frac{\alpha\beta}{2 + \left\| \boldsymbol{a}_{k+1} \right\|}, \frac{\beta}{2 + \left\| \boldsymbol{a}_{k+1} \right\|} \right\}.$$

Proof: For $\alpha \geq 1$, we have

$$\frac{\alpha\beta}{2\alpha + \left\| \boldsymbol{a}_{k+1} \right\|} = \frac{\beta}{2 + \left\| \boldsymbol{a}_{k+1} \right\| /\alpha} \geq \frac{\beta}{2 + \left\| \boldsymbol{a}_{k+1} \right\|},$$

while for $\alpha \leq 1$ we have

$$\frac{\alpha\beta}{2\alpha + \left\| \boldsymbol{a}_{k+1} \right\|} \geq \frac{\alpha\beta}{2 + \left\| \boldsymbol{a}_{k+1} \right\|}.$$

□

Proposition 3 *If $\boldsymbol{C} \cap \bigcap_{i=1}^{k} \mathcal{H}(\boldsymbol{a}_i)$ is infeasible, then*

$$\rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k], \boldsymbol{C}\right) \leq \rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \boldsymbol{C}\right).$$

Proof: Adding constraints cannot make it easier to change the program to make it feasible.

□

We now prove the main result of this section.

Proof of Lemma 20: Let $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be the columns of A, let

$$\boldsymbol{C}_k = \boldsymbol{C} \cap \bigcap_{i=1}^{k} \mathcal{H}(\boldsymbol{a}_k),$$

and let $\boldsymbol{C}_n$ be the final program. Let E_k denote the event that $\boldsymbol{C}_{k-1}$ is feasible and $\boldsymbol{C}_k$ is infeasible. Using the fact that $\boldsymbol{C}_n$ infeasible implies that E_k must hold for some k and proposition 3, we obtain

$$\begin{aligned}\mathbf{Pr}\left[\boldsymbol{C}_n \text{ is infeasible and } \rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \boldsymbol{C}\right) \leq \epsilon\right] &\leq \\ \sum_{k=1}^{n} \mathbf{Pr}\left[E_k \text{ and } \rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n], \boldsymbol{C}\right) \leq \epsilon\right] &\leq \\ \sum_{k=1}^{n} \mathbf{Pr}\left[E_k \text{ and } \rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k], \boldsymbol{C}\right) \leq \epsilon\right]. & \qquad (3.4)\end{aligned}$$

If $\boldsymbol{C}_k$ is feasible, define

$$\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) = \max_{\boldsymbol{p} \in \boldsymbol{C}_k : \|\boldsymbol{p}\|=1} \min_{1 \leq i \leq k} \boldsymbol{a}_i^T \boldsymbol{p}.$$

By equation 3.1 and lemma 16,

$$\mathbf{Pr}_{\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k}\left[\boldsymbol{C}_k \text{ is feasible and } \kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \leq \epsilon\right] \leq \frac{4\epsilon n d^{5/4}}{\sigma} \qquad (3.5)$$

By lemma 21 and proposition 2, E_{k+1} implies

$$\rho\left([\boldsymbol{a}_1, \ldots, \boldsymbol{a}_{k+1}], \boldsymbol{C}\right) \geq \min\left\{\frac{\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)}{2}, \frac{\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)\rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k)}{4 + 2\left\|\boldsymbol{a}_{k+1}\right\|}, \frac{\rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k)}{4 + 2\left\|\boldsymbol{a}_{k+1}\right\|}\right\}$$

$$\geq \frac{\min\left\{\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k),\ \kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)\rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k),\ \rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k)\right\}}{4 + 2\left\|\boldsymbol{a}_{k+1}\right\|} \qquad (3.6)$$

We now proceed to bound the probability that the numerator of this fraction is small. We first note that

$$\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)\rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k) \leq \delta$$

implies that either $\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \leq \delta$, $\rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k) \leq \delta$, or there exists an l between 1 and $\lceil \log(1/\delta) \rceil$ for which

$$\kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \leq 2^{-l+1} \text{ and } \rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k) \leq 2^l \delta.$$

We apply lemma 14 to bound

$$\mathbf{Pr}\left[E_{k+1} \text{ and } \rho(\boldsymbol{a}_{k+1}, \boldsymbol{C}_k) \leq \delta\right] \leq \frac{4\delta d^{1/4}}{\sigma},$$

and lemma 15 to bound

$$\mathbf{Pr}\left[E_{k+1} \text{ and } \kappa(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k) \leq \delta\right] \leq \frac{4\delta n d^{5/4}}{\sigma}.$$

For $1 \le l \le \lceil \log(1/\delta) \rceil$, we bound

$$\begin{aligned}
&\mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_{k+1}}\left[E_{k+1} \text{ and } \kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k) \le 2^{-l+1} \text{ and } \rho(\boldsymbol{a}_{k+1},\boldsymbol{C}_k) \le 2^l\delta\right] \\
&\quad= \quad \mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k}\left[\boldsymbol{C}_k \neq \emptyset \text{ and } \kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k) \le 2^{-l+1}\right] \cdot \\
&\qquad\quad \mathbf{Pr}_{\boldsymbol{a}_{k+1}}\left[\ \boldsymbol{C}_{k+1} = \emptyset \text{ and } \rho(\boldsymbol{a}_{k+1},\boldsymbol{C}_k) \le 2^l\delta \ \mid\ \boldsymbol{C}_k \neq \emptyset \text{ and } \kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k) \le 2^{-l+1}\right] \\
&\quad\le \quad \mathbf{Pr}_{\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k}\left[\boldsymbol{C}_k \neq \emptyset \text{ and } \kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k) \le 2^{-l+1}\right] \frac{2^l 4\delta d^{1/4}}{\sigma}\ , \text{ by lemma 14,} \\
&\quad\le \quad \frac{2^{-l+1}4nd^{5/4}}{\sigma}\frac{2^l 4\delta d^{1/4}}{\sigma}\ , \text{ by equation 3.5,} \\
&\quad= \quad \frac{32\delta n d^{3/2}}{\sigma^2}
\end{aligned}$$

Summing over the choice for l, we obtain

$$\begin{aligned}
&\mathbf{Pr}\left[E_{k+1} \text{ and } \min\left\{\kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k), \kappa(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k)\rho(\boldsymbol{a}_{k+1},\boldsymbol{C}_k), \rho(\boldsymbol{a}_{k+1},\boldsymbol{C}_k)\right\} < \delta\right] \\
&\qquad\le \quad \frac{4\delta n d^{5/4}}{\sigma} + \frac{4\delta d^{1/4}}{\sigma} + \lceil \log(1/\delta)\rceil \frac{32\delta n d^{3/2}}{\sigma^2} \\
&\qquad\le \quad \delta\left(\frac{4nd^{3/4} + 4 + 32\lceil\log(1/\delta)\rceil n d^{3/2}}{\sigma^2}\right), \text{ by } \sigma \le 1/\sqrt{d}, \\
&\qquad\le \quad \delta\left(\frac{40\lceil\log(1/\delta)\rceil n d^{3/2}}{\sigma^2}\right). \qquad (3.7)
\end{aligned}$$

On the other hand, we can bound the denominator of (3.6) by observing that $\boldsymbol{a}_{k+1}$ is a Gaussian centered at a point $\bar{\boldsymbol{a}}_{k+1}$ of norm at most 1; so, corollary 3 implies

$$\mathbf{Pr}\left[4 + 2\left\|\boldsymbol{a}_{k+1}\right\| \ge 6 + 2\sigma\sqrt{2d\ln(e/\epsilon)}\right] \le \epsilon.$$

By applying this bound and (3.7) with $\delta = \epsilon(6 + 2\sigma\sqrt{2d\ln(e/\epsilon)})$, we obtain

$$(\text{ via the schema: } \Pr[\frac{num}{den} \le \epsilon] \ \le \ \Pr[den \ge \frac{\delta}{\epsilon}] + \Pr[num \le \delta]\ )$$

$$\begin{aligned}
&\mathbf{Pr}\left[E_k \text{ and } \rho\left([\boldsymbol{a}_1,\ldots,\boldsymbol{a}_k],\boldsymbol{C}\right) \le \epsilon\right] \\
&\qquad\le \quad \epsilon + \epsilon\left(6 + 2\sigma\sqrt{2d\ln(e/\epsilon)}\right)\left(\frac{40\lceil\log(1/\epsilon(3+\sigma\sqrt{d\log(e/\epsilon)}))\rceil n d^{3/2}}{\sigma^2}\right) \\
&\qquad\le \quad \epsilon + \epsilon\left(6 + 3\sqrt{\ln(e/\epsilon)}\right)\left(\frac{40\lceil\log(1/\epsilon)\rceil n d^{3/2}}{\sigma^2}\right) \\
&\qquad\le \quad \epsilon\left(\frac{360\lceil\log^{1.5}(1/\epsilon)\rceil n d^{3/2}}{\sigma^2}\right)
\end{aligned}$$

Plugging this in to (3.4), we get

$$\mathbf{Pr}\left[\boldsymbol{C}_0 \text{ is infeasible and } \rho\left([\boldsymbol{a}_1, \dots, \boldsymbol{a}_n], \boldsymbol{C}\right) \leq \epsilon\right] \quad \leq \quad \frac{360\epsilon n^2 d^{3/2} \lceil \log^{1.5}(1/\epsilon) \rceil}{\sigma^2}.$$

□

3.2.3 Primal number, both cases

We combine the results of sections 3.2.1 and 3.2.2 to prove lemma 12, that the primal condition number is probably low.

Proof of Lemma 12: In lemma 15, we show that

$$\mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C}_0 \text{ is feasible}) \text{ and } (\rho(A, \boldsymbol{C}_0) \leq \epsilon)\right] \leq \left(\frac{4\epsilon n d^{5/4}}{\sigma}\right),$$

while in lemma 20, we show

$$\mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C}_0 \text{ is infeasible}) \text{ and } (\rho(A, \boldsymbol{C}_0) \leq \epsilon)\right] \quad \leq \quad \left(\frac{360\epsilon \lceil \log^{1.5}(1/\epsilon) \rceil n^2 d^{3/2}}{\sigma^2}\right).$$

Thus,

$$\begin{aligned} \mathbf{Pr}\left[\rho(A, \boldsymbol{C}) \leq \epsilon\right] &= \mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C}_0 \text{ is feasible}) \text{ and } (\rho(A, \boldsymbol{C}_0) \leq \epsilon)\right] \\ &\quad + \mathbf{Pr}\left[(A\boldsymbol{x} \geq 0,\ \boldsymbol{x} \in \boldsymbol{C}_0 \text{ is infeasible}) \text{ and } (\rho(A, \boldsymbol{C}_0) \leq \epsilon)\right] \\ &\leq \left(\frac{4\epsilon n d^{5/4}}{\sigma}\right) + \left(\frac{360\epsilon \lceil \log^{1.5}(1/\epsilon) \rceil n^2 d^{3/2}}{\sigma^2}\right) \\ &\leq \left(\frac{364\epsilon \lceil \log^{1.5}(1/\epsilon) \rceil n^2 d^{3/2}}{\sigma^2}\right) \end{aligned}$$

Letting $\alpha = 364 \frac{n^2 d^{3/2}}{\sigma^2}$ and $\epsilon = \delta / (3\alpha \log^{1.5}(\alpha/\delta))$ yields

$$\mathbf{Pr}\left[\frac{1}{\rho(A, \boldsymbol{C})} > \frac{1100\ n^2 d^{3/2}}{\delta \sigma^2} \log^{3/2}\left(\frac{370\ n^2 d^{3/2}}{\delta^2 \sigma^2}\right)\right] \leq \delta/2. \tag{3.8}$$

At the same time, corollary 3 tells us that

$$\mathbf{Pr}\left[\|A\|_F \geq 1 + \sigma\sqrt{nd\ 2\ln(2e/\delta)}\right] \leq \delta/2.$$

The lemma now follows by applying this bound, $\sigma \leq 1/\sqrt{nd}$, and (3.8), to get

$$\mathbf{Pr}\left[\frac{\|A\|_F}{\rho(A, \boldsymbol{C})} > \frac{(1 + \sqrt{2\ln(2e/\delta)})1100\ n^2 d^{3/2}}{\delta \sigma^2} \log^{3/2}\left(\frac{370\ n^2 d^{3/2}}{\delta^2 \sigma^2}\right)\right] < \delta$$

To derive the lemma as stated, we note

$$\frac{(1 + \sqrt{2\ln(2e/\delta)})1100\ n^2 d^{3/2}}{\delta \sigma^2} \log^{3/2}\left(\frac{370\ n^2 d^{3/2}}{\delta^2 \sigma^2}\right) \leq \frac{2^{12}\ n^2 d^{3/2}}{\delta^2 \sigma^2} \log^2\left(\frac{2^9\ n^2 d^{3/2}}{\delta^2 \sigma^2}\right)$$

. □

3.3 Dual Condition Number

In this section, we consider linear programs of the form

$$A^T \boldsymbol{y} = \boldsymbol{c}, \ \boldsymbol{y} \geq \boldsymbol{0}.$$

The dual program of form (1) and the primal program of form (3) are both of this type. The dual program of form (4) can be handled using a slightly different argument than the one we present. As in section 3.2, we omit the details of the modifications necessary for form (4). We begin by defining distance to ill-posedness appropriately for the form of linear program considered in this section:

Definition 10 (Dual Distance to Ill-Posedness) *For a matrix, A, and a vector $\boldsymbol{c}$, we define $\rho(A, \boldsymbol{c})$ by*

a. if $A^T \boldsymbol{y} = \boldsymbol{c}, \ \boldsymbol{y} \geq \boldsymbol{0}$ is feasible, then $\rho(A, \boldsymbol{c}) =$

$$\sup \left\{ \epsilon : \left\| \Delta A \right\|_F + \left\| \Delta \boldsymbol{c} \right\|_F < \epsilon \textit{ implies } (A + \Delta A)^T \boldsymbol{y} = \boldsymbol{c} + \Delta \boldsymbol{c}, \ \boldsymbol{y} \geq \boldsymbol{0} \textit{ is feasible} \right\}$$

b. if $A^T \boldsymbol{y} = \boldsymbol{c}, \ \boldsymbol{y} \geq \boldsymbol{0}$ is infeasible, then $\rho(A, \boldsymbol{c}) =$

$$\sup \left\{ \epsilon : \left\| \Delta A \right\|_F + \left\| \Delta \boldsymbol{c} \right\|_F < \epsilon \textit{ implies } (A + \Delta A)^T \boldsymbol{y} = \boldsymbol{c} + \Delta \boldsymbol{c}, \ \boldsymbol{y} \geq \boldsymbol{0} \textit{ is infeasible} \right\}$$

The main result of this section is:

Lemma 22 (Dual condition number is likely low.) *Let A and $\boldsymbol{c}$ be a Gaussian random matrix and vector of variance σ^2, $\sigma \leq 1/\sqrt{nd}$, centered at $\bar{A}$ and $\bar{\mathbf{c}}$, respectively. If $\left\| \bar{A} \right\|_F \leq 1$ and $\left\| \bar{\mathbf{c}} \right\| \leq 1$, then*

$$\mathbf{Pr} \left[\frac{\left\| A, \boldsymbol{c} \right\|_F}{\rho(A, \boldsymbol{c})} > \frac{1000 \ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \log^{1.5} \left(\frac{200 \ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \right) \right] \quad \leq \quad \epsilon.$$

We begin by giving several common definitions that will be useful in our analysis of the dual condition number. We define a change of variables, and we then develop a sufficient geometric condition for the dual condition number to be low. In lemma 25 and in the proof of lemma 22, we show that this geometric condition is met with good probability.

Definition 11 (Cone) *For a set of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, let* **Cone** $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ *denote* $\{ \boldsymbol{x} : \boldsymbol{x} = \sum_i \lambda_i \boldsymbol{a}_i, \quad \lambda_i \geq 0 \}$.

Definition 12 (Hull) *For a set of vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, let* **Hull** $(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ *denote* $\{ \boldsymbol{x} : \boldsymbol{x} = \sum_i \lambda_i \boldsymbol{a}_i, \quad \lambda_i \geq 0, \quad \sum_i \lambda_i = 1 \}$.

Definition 13 (Boundary of a Set) *For a convex set $\boldsymbol{S}$, let* **bdry**$(\boldsymbol{S})$ *denote the boundary of $\boldsymbol{S}$, i.e.,* $\{ \boldsymbol{x} : \forall \epsilon > 0, \ \exists \boldsymbol{e}, \ \| \boldsymbol{e} \| \leq \epsilon, \ \textit{s.t.} \ \boldsymbol{x} + \boldsymbol{e} \in \boldsymbol{S}, \ \boldsymbol{x} - \boldsymbol{e} \notin \boldsymbol{S} \}$.

Definition 14 (Point-To-Set Distance) *Let* **dist** $(\boldsymbol{x}, \boldsymbol{S})$ *denote the distance of $\boldsymbol{x}$ to $\boldsymbol{S}$, i.e.,* $\min \{ \epsilon : \exists \boldsymbol{e}, \ \| \boldsymbol{e} \| \leq \epsilon, \ \textit{s.t.} \ \boldsymbol{x} + \boldsymbol{e} \in \boldsymbol{S} \}$.

Note that $\mathbf{Cone}\,(\boldsymbol{a}_1,\dots,\boldsymbol{a}_n)$ is not an open cone, while $\mathbf{Hull}\,(\boldsymbol{a}_1,\dots,\boldsymbol{a}_n)$ is the standard convex hull of $\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_n\}$.

We observe that there exists a solution to the system $A^T\boldsymbol{y}=\boldsymbol{c},\ \boldsymbol{y}\geq\boldsymbol{0}$ if and only if

$$\boldsymbol{c}\in\mathbf{Cone}\,(\boldsymbol{a}_1,\dots,\boldsymbol{a}_n)\,,$$

and that for $\boldsymbol{c}\neq\boldsymbol{0}$, this holds if and only if

$$\mathbf{Ray}\,(\boldsymbol{c})\ \text{ intersects }\mathbf{Hull}\,(\boldsymbol{a}_1,\dots,\boldsymbol{a}_n)\,.$$

The main idea we need beyond the ideas of section 3.2 is to perform an illuminating change of variables. We set

$$\begin{aligned}\boldsymbol{z} &= (1/n)\sum_{i=1}^{n}\boldsymbol{a}_i,\ \text{and}\\ \boldsymbol{x}_i &= \boldsymbol{a}_i-\boldsymbol{z},\text{for } i=1 \text{ to } n-1.\end{aligned}$$

For notational convenience, we let $\boldsymbol{x}_n=\boldsymbol{a}_n-\boldsymbol{z}$, although $\boldsymbol{x}_n$ is not independent of $\{\boldsymbol{z},\boldsymbol{x}_1,\dots,\boldsymbol{x}_{n-1}\}$.

We can restate the condition for the linear progam to be ill-posed in these new variables:

Lemma 23 (Ill-posedness in new variables.)

$A^T\boldsymbol{y}=\boldsymbol{c},\ \boldsymbol{y}\geq\boldsymbol{0},\ \boldsymbol{c}\neq\boldsymbol{0}$ *is ill-posed if and only if* $\boldsymbol{z}\in\mathbf{bdry}(\mathbf{Ray}\,(\boldsymbol{c})-\mathbf{Hull}\,(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n))$.

Proof: We observe

$$\begin{aligned} A^T\boldsymbol{y}=\boldsymbol{c},\ \boldsymbol{y}\geq\boldsymbol{0}\text{ is feasible} &\iff \mathbf{Ray}\,(\boldsymbol{c})\ \text{ intersects }\mathbf{Hull}\,(\boldsymbol{a}_1,\dots,\boldsymbol{a}_n)\\ &\iff \mathbf{Ray}\,(\boldsymbol{c})\ \text{ intersects }\boldsymbol{z}+\mathbf{Hull}\,(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)\\ &\iff \boldsymbol{z}\in\mathbf{Ray}\,(\boldsymbol{c})-\mathbf{Hull}\,(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)\,.\end{aligned}$$

For $\boldsymbol{c}\neq\boldsymbol{0}$, $\mathbf{Ray}\,(\boldsymbol{c})-\mathbf{Hull}\,(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$ is a continuous mapping from $\boldsymbol{c},\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ to subsets of Euclidean space, and so for $\boldsymbol{z}$ in the set and not on the boundary, a sufficiently small change to all the variables simultaneously will always leave $\boldsymbol{z}$ in the set, and similarly for $\boldsymbol{z}$ not in the set and not on the boundary.

To establish the other direction, if $\boldsymbol{z}$ is on the boundary, we can just perturb $\boldsymbol{z}$ to bring it in or out of the set. Although $\boldsymbol{z},\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ are determined by the $\boldsymbol{a}_1,\dots,\boldsymbol{a}_n$, we can perturb the $\boldsymbol{a}_1,\dots,\boldsymbol{a}_n$ so as to change the value of $\boldsymbol{z}$ without changing the values of any of the $\boldsymbol{x}_1,\dots,\boldsymbol{x}_n$ (see the proof of lemma 24 below for more detail on why the change of variables permits this).

The lemma is also true for $\boldsymbol{c}=\boldsymbol{0}$, but we will not need this. □

Note that $\mathbf{Ray}\,(\boldsymbol{c})-\mathbf{Hull}\,(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$ is a convex set. The following lemma will allow us to apply lemma 13 to determine the probability that $\boldsymbol{z}$ is near the boundary of this convex set.

Lemma 24 (Independence of mean among new variables.) *Let* $\boldsymbol{a}_1,\dots,\boldsymbol{a}_n$ *be Gaussian random vectors of variance* σ^2 *lying in* $\mathbb{R}^d$*. Let*

$$\boldsymbol{z}=\frac{1}{n}\sum_i\boldsymbol{a}_i\ \textit{ and }\ \boldsymbol{x}_i=\boldsymbol{a}_i-\boldsymbol{z},\ \textit{ for } 1\leq i\leq n.$$

Then, $\boldsymbol{z}$ is a Gaussian random vector of variance σ^2/n and is independent of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

Proof: As $\boldsymbol{z}$ is the average of n Gaussian random vectors of variance σ^2, it is a Gaussian random vector of variance σ^2/n. We have that $\boldsymbol{z}$ is independent of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ because the linear combination of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ used to obtain $\boldsymbol{z}$ is orthogonal to the linear combinations of $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ used to obtain the $\boldsymbol{x}_i$s. □

We proceed to apply lemma 13.

Lemma 25 (Mean is likely far from ill-posedness.) *Let $\boldsymbol{c}$ and $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ be Gaussian random vectors of variance σ^2 lying in $\mathbb{R}^d$. Let*

$$\boldsymbol{z} = \frac{1}{n}\sum_i \boldsymbol{a}_i \text{ and } \boldsymbol{x}_i = \boldsymbol{a}_i - \boldsymbol{z}, \text{ for } 1 \leq i \leq n.$$

Then,

$$\mathbf{Pr}\left[\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right))\right) \leq \epsilon\right] \leq \frac{8\epsilon d^{1/4} n^{1/2}}{\sigma}.$$

Proof: Let $\boldsymbol{c}$ be arbitrary. By lemma 24, we can choose $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ and then choose $\boldsymbol{z}$ independently. Having chosen $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, we fix the convex body $\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)$ and apply lemma 13. The factor of 2 arises because $\boldsymbol{z}$ must miss an ϵ boundary on either side of the convex body. □

Lemma 26 (Geometric condition to be far from ill-posedness in new variables.) *If*

$$\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right))\right) > \alpha \tag{3.9}$$

and

$$\begin{aligned} \|\Delta \boldsymbol{x}_i\| &\leq \alpha/4, \\ \|\Delta \boldsymbol{z}\| &\leq \alpha/4, \\ \|\Delta \boldsymbol{c}\| &\leq \frac{\alpha\left\|\boldsymbol{c}\right\|}{2\alpha + 4(\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\|)}, \end{aligned}$$

then

$$\boldsymbol{z} + \Delta\boldsymbol{z} \notin \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c} + \Delta\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n\right))$$

Proof: Assume for the purpose of showing a contradiction that

$$\boldsymbol{z} + \Delta\boldsymbol{z} \in \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c} + \Delta\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n\right))$$

Consider the case that $\boldsymbol{z} \notin \mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right)$. We will show that $\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right))\right) \leq \alpha$, contradicting our lemma assumption (3.9). Since $\boldsymbol{z} + \Delta\boldsymbol{z} \in \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c} + \Delta\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n\right))$,

$$\boldsymbol{z} + \Delta\boldsymbol{z} = \lambda(\boldsymbol{c} + \Delta\boldsymbol{c}) - \sum_i \gamma_i(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i),$$

for some $\lambda \geq 0$ and $\gamma_1, \ldots, \gamma_n \geq 0, \sum_i \gamma_i = 1$. We establish an upper bound on λ by noting that

$$\lambda = \frac{\|\boldsymbol{z} + \Delta\boldsymbol{z} + \sum_i \gamma_i(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)\|}{\|\boldsymbol{c} + \Delta\boldsymbol{c}\|}.$$

We lower bound the denominator by $\|\boldsymbol{c}\|/2$ by observing that

$$\|\Delta\boldsymbol{c}\| \leq \frac{\alpha\|\boldsymbol{c}\|}{2\alpha + 4(\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\|)} \leq \|\boldsymbol{c}\|/2.$$

We upper bound the numerator by

$$\begin{aligned} \left\|\boldsymbol{z} + \Delta\boldsymbol{z} + \sum_i \gamma_i(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)\right\| &\leq \|\boldsymbol{z}\| + \alpha/4 + \sum_i \gamma_i(\|\boldsymbol{x}_i\| + \|\Delta\boldsymbol{x}_i\|) \\ &\leq \|\boldsymbol{z}\| + \alpha/4 + \max_i \|\boldsymbol{x}_i\| + \alpha/4 \\ &= \|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\| + \alpha/2. \end{aligned}$$

Thus,

$$\lambda \leq \frac{\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\| + \alpha/2}{\|\boldsymbol{c}\|/2}$$

Since

$$\boldsymbol{z} + \Delta\boldsymbol{z} - \lambda\Delta\boldsymbol{c} + \sum_i \gamma_i \Delta\boldsymbol{x}_i = \lambda\boldsymbol{c} - \sum_i \gamma_i \boldsymbol{x}_i \in \mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

We find that

$$\begin{aligned} \mathbf{dist}(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))) &\leq \left\|\Delta\boldsymbol{z} - \lambda\Delta\boldsymbol{c} + \sum_i \gamma_i \Delta\boldsymbol{x}_i\right\| \\ &\leq \|\Delta\boldsymbol{z}\| + \lambda\|\Delta\boldsymbol{c}\| + \sum_i \gamma_i \|\Delta\boldsymbol{x}_i\| \end{aligned}$$

$$\begin{aligned} &\leq \frac{\alpha}{4} + \left(\frac{\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\| + \alpha/2}{\|\boldsymbol{c}\|/2}\right)\left(\frac{\alpha\|\boldsymbol{c}\|}{2\alpha + 4(\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\|)}\right) + \frac{\alpha}{4} \\ &= \alpha. \end{aligned}$$

This establishes a contradiction in the case that $\boldsymbol{z} \notin \mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Now consider the case that $\boldsymbol{z} \in \mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Since

$$\boldsymbol{z} + \Delta\boldsymbol{z} \in \mathbf{bdry}(\mathbf{Ray}(\boldsymbol{c} + \Delta\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n))$$

there exists a hyperplane H passing through $\boldsymbol{z} + \Delta\boldsymbol{z}$ and tangent to the convex set $\mathbf{Ray}(\boldsymbol{c} + \Delta\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n)$. By the assumption that $\mathbf{dist}(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n))) > \alpha$, there is some $\delta_0 > 0$ such that, for every $\delta \in (0, \delta_0)$, every point within $\alpha + \delta$ of $\boldsymbol{z}$ lies within $\mathbf{Ray}(\boldsymbol{c}) - \mathbf{Hull}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$. Choose

$\delta \in (0, \delta_0)$ that also satifises $\delta \leq \|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\|$. Let $\boldsymbol{q}$ be a point at distance $\frac{3\alpha}{4} + \delta$ from $\boldsymbol{z} + \Delta\boldsymbol{z}$ in the direction perpendicular to H. Since $\textbf{dist}\,(\boldsymbol{z}, \boldsymbol{z} + \Delta\boldsymbol{z}) \leq \frac{\alpha}{4}$, and $\textbf{dist}\,(\boldsymbol{z} + \Delta\boldsymbol{z}, \boldsymbol{q}) \leq \frac{3\alpha}{4} + \delta$,

$$\boldsymbol{q} \in \textbf{Ray}\,(\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$$

At the same time,

$$\textbf{dist}\,(\boldsymbol{q}, \textbf{Ray}\,(\boldsymbol{c} + \Delta\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n)) > \frac{3\alpha}{4}$$

Because $\boldsymbol{q} \in \textbf{Ray}\,(\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, there exist $\lambda \geq 0$ and $\gamma_1, \ldots, \gamma_n \geq 0, \sum_i \gamma_i = 1$ such that

$$\boldsymbol{q} = \lambda\boldsymbol{c} - \sum_i \gamma_i \boldsymbol{x}_i.$$

We upper bound λ as before,

$$\lambda = \frac{\|\boldsymbol{q} + \sum_i \gamma_i \boldsymbol{x}_i\|}{\|\boldsymbol{c}\|} \leq \frac{\|\boldsymbol{z}\| + \alpha + \delta + \max_i \|\boldsymbol{x}_i\|}{\|\boldsymbol{c}\|} \leq \frac{\|\boldsymbol{z}\| + \max_i \|\boldsymbol{x}_i\| + \alpha/2}{\|\boldsymbol{c}\|/2}$$

Hence

$$\boldsymbol{q} + \lambda\Delta\boldsymbol{c} - \sum_i \gamma_i \Delta\boldsymbol{x}_i = \lambda(\boldsymbol{c} + \Delta\boldsymbol{c}) - \sum_i \gamma_i(\boldsymbol{x}_i + \Delta\boldsymbol{x}_i)$$

$$\in \quad \textbf{Ray}\,(\boldsymbol{c} + \Delta\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n)$$

and thus

$$\begin{aligned} \textbf{dist}\,(\boldsymbol{q}, \textbf{Ray}\,(\boldsymbol{c} + \Delta\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1 + \Delta\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n + \Delta\boldsymbol{x}_n)) &\leq \left\|\lambda\Delta\boldsymbol{c} - \sum_i \gamma_i \Delta\boldsymbol{x}_i\right\| \\ &\leq \lambda\|\Delta\boldsymbol{c}\| + \max_i \|\Delta\boldsymbol{x}_i\| \\ &\leq \alpha/2 + \alpha/4 \\ &\leq 3\alpha/4 \end{aligned}$$

which is a contradiction. This concludes the proof of the lemma. □

We now derive a consequence of lemma 26 that uses both the original and the new variables.

Lemma 27 (Reciprocal of distance to ill-posedness.) *Let* $\boldsymbol{c}$ *and* $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ *be vectors in* $\mathbb{R}^d$. *Let*

$$\boldsymbol{z} = \frac{1}{n}\sum_i \boldsymbol{a}_i \text{ and } \boldsymbol{x}_i = \boldsymbol{a}_i - \boldsymbol{z}, \text{ for } 1 \leq i \leq n.$$

$$k_1 = \textbf{dist}\,(\boldsymbol{z}, \textbf{bdry}(\textbf{Ray}\,(\boldsymbol{c}) - \textbf{Hull}\,(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)))$$

$$k_2 = \|\boldsymbol{c}\|$$

Then

$$\frac{1}{\rho(A, \boldsymbol{c})} \leq \max\left\{\frac{8}{k_1}, \frac{4}{k_2}, \frac{24 \max_i \|\boldsymbol{a}_i\|}{k_1 k_2}\right\}.$$

Proof: By the definition of k_1 and k_2 and lemma 26, we can tolerate any change of magnitude up to $k_1/4$ in $\boldsymbol{z}, \{\boldsymbol{x}_i\}$ and any change of up to $\frac{k_1 k_2}{2k_1+4(\|\boldsymbol{z}\|+\max\|\boldsymbol{x}_i\|)}$ in $\boldsymbol{c}$ without the program becoming ill-posed. We show that this means we can tolerate any change of up to $k_1/8$ in $\boldsymbol{a}_i$ without the program becoming ill-posed. Formally, we need to show that if $\|\Delta \boldsymbol{a}_i\| \leq k_1/8$ for all i, then $\|\Delta \boldsymbol{z}\| \leq k_1/4$ and $\|\Delta \boldsymbol{x}_i\| \leq k_1/4$. Since $\Delta \boldsymbol{z} = (1/n)\sum \Delta \boldsymbol{a}_i$, $\|\Delta \boldsymbol{z}\| \leq k_1/8$. Since $\Delta \boldsymbol{x}_i = \Delta \boldsymbol{a}_i - \Delta \boldsymbol{z}$, $\|\Delta \boldsymbol{x}_i\| \leq k_1/8 + k_1/8 = k_1/4$. Thus

$$\rho(A, \boldsymbol{c}) \geq \min\left\{\frac{k_1}{8}, \frac{k_1 k_2}{2k_1 + 4(\|\boldsymbol{z}\| + \max\|\boldsymbol{x}_i\|)}\right\}$$

which implies

$$\frac{1}{\rho(A, \boldsymbol{c})} \leq \max\left\{\frac{8}{k_1}, \frac{4}{k_2}, \frac{8(\|\boldsymbol{z}\| + \max\|\boldsymbol{x}_i\|)}{k_1 k_2}\right\}$$

Since $\boldsymbol{z} = (1/n)\sum \boldsymbol{a}_i \Rightarrow \|\boldsymbol{z}\| \leq \max\|\boldsymbol{a}_i\|$, and $\boldsymbol{x}_i = \boldsymbol{a}_i - \boldsymbol{z} \Rightarrow \|\boldsymbol{x}_i\| \leq \|\boldsymbol{a}_i\| + \|\boldsymbol{z}\| \leq 2\max\|\boldsymbol{a}_i\|$, we have

$$\frac{1}{\rho(A, \boldsymbol{c})} \leq \max\left\{\frac{8}{k_1}, \frac{4}{k_2}, \frac{24 \max\|\boldsymbol{a}_i\|}{k_1 k_2}\right\}$$

This concludes the proof. □

Proof of Lemma 22: Let

$$\boldsymbol{z} = \frac{1}{n}\sum_i \boldsymbol{a}_i \text{ and } \boldsymbol{x}_i = \boldsymbol{a}_i - \boldsymbol{z}, \text{ for } 1 \leq i \leq n,$$

$$k_1 = \mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\right))\right) \text{ and } k_2 = \|\boldsymbol{c}\|.$$

We will apply the bound of lemma 27. We first lower bound $\min\{k_1, k_2, k_1k_2\}$. We begin by noting that if

$$\min\{k_1, k_2, k_1 k_2\} < \epsilon,$$

then either

$$\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\right))\right) < \epsilon, \tag{3.10}$$

or

$$\|\boldsymbol{c}\| < \epsilon, \tag{3.11}$$

or there exists some integer l, $1 \leq l \leq \lceil \log \frac{1}{\epsilon} \rceil$, for which

$$\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\right))\right) < 2^l \epsilon \quad \text{and} \quad \|\boldsymbol{c}\| \leq 2^{-l+1}. \tag{3.12}$$

The probabilities of events 3.10 and 3.11 will also be bounded in our analysis of event 3.12.

By corollary 4, for $d \geq 2$, we have

$$\mathbf{Pr}\left[\|\boldsymbol{c}\| \leq \epsilon\right] \leq \frac{e\epsilon}{\sigma},$$

which translates to

$$\mathbf{Pr}\left[\|\boldsymbol{c}\| \leq 2^{-l+1}\right] \leq \frac{e2^{-l+1}}{\sigma},$$

while lemma 25 implies

$$\mathbf{Pr}\left[\mathbf{dist}\left(\boldsymbol{z}, \mathbf{bdry}(\mathbf{Ray}\left(\boldsymbol{c}\right) - \mathbf{Hull}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\right))\right) < 2^l \epsilon\right] \leq \frac{8 \cdot 2^l \epsilon d^{1/4} n^{1/2}}{\sigma}.$$

Thus, we compute

$$\begin{aligned}
\mathbf{Pr}\left[\min\left\{k_1, k_2, k_1 k_2\right\} < \epsilon\right] &\leq \frac{8\ \epsilon d^{1/4} n^{1/2}}{\sigma} + \frac{e\epsilon}{\sigma} + \sum_{l=1}^{\lceil \log \frac{1}{\epsilon} \rceil} \frac{e2^{-l+1}}{\sigma} \frac{8 \cdot 2^l \epsilon d^{1/4} n^{1/2}}{\sigma} \\
&= \frac{8\ \epsilon d^{1/4} n^{1/2}}{\sigma} + \frac{e\epsilon}{\sigma} + \frac{16 e \epsilon d^{1/4} n^{1/2}}{\sigma^2} \log(\frac{1}{\epsilon}) \\
&\leq \frac{55\ \epsilon d^{1/4} n^{1/2}}{\sigma^2} \log(\frac{1}{\epsilon}).
\end{aligned}$$

We re-write this as

$$\mathbf{Pr}\left[\max\left\{1/k_1, 1/k_2, 1/k_1 k_2\right\} > \frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \log(\frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2})\right] < \frac{\epsilon}{2}.$$

From corollary 3, we know that

$$\mathbf{Pr}\left[\|A, \boldsymbol{c}\|_F > 3 + \sigma\sqrt{(d+1)n\ 2\ln(2e/\epsilon)}\right] < \frac{\epsilon}{2}.$$

Thus,

$$\mathbf{Pr}\left[\frac{\|A, \boldsymbol{c}\|_F}{\rho(A, \boldsymbol{c})} > \frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \log(\frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2})(3 + \sigma\sqrt{(d+1)n 2\ln(2e/\epsilon)})\right] \quad \leq \quad \epsilon.$$

To derive the lemma as stated, we conclude with

$$\frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \log(\frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2})(3 + \sigma\sqrt{(d+1)n\ 2\ln(2e/\epsilon)}) \leq$$

$$\frac{1000\ d^{1/4} n^{1/2}}{\epsilon \sigma^2} \log^{1.5}\left(\frac{200\ d^{1/4} n^{1/2}}{\epsilon \sigma^2}\right)$$

□

3.4 Combining the Primal and Dual Analyses

Our main theorem is now very easy to prove.

Proof of Theorem 5: Apply lemmas 22 and 12:

$$\frac{2^{13}n^2d^{3/2}}{\epsilon\sigma^2}\log^2\left(\frac{2^9n^2d^{3/2}}{\epsilon\sigma^2}\right)+\frac{2^{11}d^{1/4}n^{1/2}}{\epsilon\sigma^2}\log^{1.5}\left(\frac{2^8d^{1/4}n^{1/2}}{\epsilon\sigma^2}\right)$$

$$\leq\frac{2^{14}n^2d^{3/2}}{\epsilon\sigma^2}\log^2\left(\frac{2^{10}n^2d^{3/2}}{\epsilon\sigma^2}\right)$$

□

We recall the four canonical forms for linear programs that we have discussed.

$$\begin{array}{rll}
\max\ \boldsymbol{c}^T\boldsymbol{x}\quad \textit{s.t.}\quad A\boldsymbol{x}\leq\boldsymbol{b} & \text{and its dual} & \min\ \boldsymbol{b}^T\boldsymbol{y}\quad \textit{s.t.}\quad A^T\boldsymbol{y}=\boldsymbol{c},\quad \boldsymbol{y}\geq\boldsymbol{0}, \quad (1)\\
\max\ \boldsymbol{c}^T\boldsymbol{x}\ \text{ s.t. } A\boldsymbol{x}\leq\boldsymbol{b},\ \boldsymbol{x}\geq\boldsymbol{0} & \text{and its dual} & \min\ \boldsymbol{b}^T\boldsymbol{y}\ \text{ s.t. } A^T\boldsymbol{y}\leq\boldsymbol{c},\ \boldsymbol{y}\geq\boldsymbol{0} \quad (2)\\
\max\ \boldsymbol{c}^T\boldsymbol{x}\ \text{ s.t. } A\boldsymbol{x}=\boldsymbol{b},\ \boldsymbol{x}\geq\boldsymbol{0} & \text{and its dual} & \min\ \boldsymbol{b}^T\boldsymbol{y}\ \text{ s.t. } A^T\boldsymbol{y}\leq\boldsymbol{c} \quad (3)\\
\text{find } \boldsymbol{x}\neq\boldsymbol{0} \text{ s.t. } A\boldsymbol{x}\leq\boldsymbol{0} & \text{and its dual} & \text{find } \boldsymbol{y}\neq\boldsymbol{0} \text{ s.t. } A^T\boldsymbol{y}=\boldsymbol{0},\ \boldsymbol{y}\geq\boldsymbol{0}, \quad (4)
\end{array}$$

Renegar developed efficient algorithms for both solving and estimating the condition number of programs in form (2) in [Ren 94]. Vera [Ver 96] developed efficient algorithms for forms (1) and (3). Cucker and Peña developed algorithms for form (4) in [CP 01]. In [FV 00], Freund and Vera give a unified approach which both efficiently estimates the condition number and solves the linear programs in any of these forms. A bound on the smoothed complexity of all of these algorithms follows from theorem 5.

3.5 Future Avenues of Investigation

We hope that smoothed analysis of algorithms provides an attractive avenue for other researchers to explore the discrepancy that is sometimes observed between the worst-case complexity and the typical performance of algorithms. We also hope that this work illuminates some of the potential shared interests of the numerical analysis and theoretical computer science communities. One potential direction for future research is the application of smoothed analysis to other problem domains, but there are several others we would also like to highlight.

We do not address in this thesis the question of the actual distribution of condition numbers. We would be particularly interested to hear the results of computational experiments, like those of Freund and Ordoñez[FO 02], that explore the distribution of condition numbers occurring in real-world problems.

In section 3.7, we discuss several alternative models of perturbation. We present several negative results for the condition number under these models of perturbation, but we are not aware of any such negative result for the simplex algorithm. It would be very intriguing if under one of these models of relative perturbation the simplex algorithm ran in polynomial time.

The most significant open challenge of smoothed analysis is the leap from *analysis* to *synthesis*. At the time of this writing, I am not aware of a new algorithm that has been proposed based on its superior performance in the smoothed complexity setting. The discovery of a new and practically useful algorithm suggested by smoothed complexity would be very exciting.

3.6 The Perceptron Algorithm

In this section we describe the *perceptron* algorithm, a classic algorithm from maching learning that also solves linear programming problems. The perceptron algorithm is one of a host of "elementary" algorithms for linear programming that has a low cost per iteration, but which requires a number of iterations polynomial in the condition number. In contrast, the ellipsoid method and all the interior point methods have a much higher cost per iteration, but require a number of iterations proportional to the log of the condition number. Algorithms for linear programming with convergence rates polynomial in the condition number have recently attracted some attention because of their ability to quickly find an approximate solution to a large system[Bie 01].

We begin by describing the perceptron algorithm in its classic machine learning setting and the running time analysis. We then describe how linear programs may be mapped to inputs to the perceptron algorithm. Lastly, we point out that the analysis of the running time shows that the perceptron algorithm requires a number of iterations polynomial in the primal condition number.

3.6.1 Algorithm Definition and Analysis

The perceptron algorithm is a classic algorithm for solving the following machine learning problem, finding a separating hyperplane:

> Given a set of points in d-dimensional space, each labeled as "positive" or "negative", find a separating hyperplane (a hyperplane with all positives on one side and all negatives on the other) if one exists. That is, we want to find $(\boldsymbol{w}, w_0)$ such that for all positive points $\boldsymbol{a}_i$, we have $\boldsymbol{a}_i^T\boldsymbol{w} > w_0$ and for all negative points $\boldsymbol{a}_i$, we have $\boldsymbol{a}_i^T\boldsymbol{w} < w_0$.

To run the perceptron algorithm, one first performs two standard simplifying transformations to the data. The first is to give each point a $(d+1)^{st}$ coordinate with value 1, which allows us to assume that the separating hyperplane passes through the origin. That is, for positive points $\boldsymbol{a}_i$ we rewrite the requirement $\boldsymbol{a}_i^T\boldsymbol{w} > w_0$ as $(\boldsymbol{a}_i, 1)^T(\boldsymbol{w}, -w_0) > 0$, and similarly for negatives. The second transformation is to flip all negative points through the origin and view them as positives. (I.e., replace the constraint $(\boldsymbol{a}_i, 1)^T(\boldsymbol{w}, -w_0) < 0$ with $(-\boldsymbol{a}_i, -1)^T(\boldsymbol{w}, -w_0) > 0$.) The problem is now reduced to (in the new variables):

> Given a set of points $\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_m$, find a vector $\boldsymbol{w}$ such that $\boldsymbol{a}_i^T\boldsymbol{w} > 0$ for all i, if one exists.

The perceptron algorithm now works as follows:

Perceptron Algorithm

1. Initialize $\boldsymbol{w} = \boldsymbol{0}$ (the all-zero vector).
2. Pick some $\boldsymbol{a}_i$ such that $\boldsymbol{a}_i^T\boldsymbol{w} \le 0$ and update $\boldsymbol{w}$ by

$$\boldsymbol{w} \;\leftarrow\; \boldsymbol{w} + \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|}$$

3. If we do not have $\boldsymbol{a}_i^T\boldsymbol{w} > 0$ for all i, go back to step 2.

While it is not hard to construct instances where the running time of this algorithm is exponential in the dimension d, a beautiful theorem of Block and Novikoff (see Minsky and Papert [MP 69]) upper-bounds the running time in terms of the "wiggle room" available for a solution. Specifically, let $\boldsymbol{w}^*$ denote the solution that maximizes the wiggle room ν defined as $\nu = \min_i \frac{|\boldsymbol{a}_i^T \boldsymbol{w}^*|}{\|\boldsymbol{a}_i\| \|\boldsymbol{w}^*\|}$. In other words, not only is $\boldsymbol{w}^*$ feasible ($\boldsymbol{a}_i^T \boldsymbol{w}^* > 0 \quad \forall i$), but every $\boldsymbol{w}$ within angle $\arcsin(\nu)$ of $\boldsymbol{w}^*$ is also feasible.

Theorem 6 (Block-Novikoff) *The perceptron algorithm terminates in at most $1/\nu^2$ iterations.*

Note that this implies the perceptron algorithm eventually converges to a feasible solution if one exists with non-zero wiggle room. We provide a proof of the theorem here.

Proof: ([MP 69]) Let $\boldsymbol{w}^*$ be a solution of wiggle room ν, and for convenience scale $\boldsymbol{w}^*$ so that it is a unit vector. (Scaling does not change the value of ν.) To show convergence within the specified number of iterations, we consider the quantities $\boldsymbol{w}^T\boldsymbol{w}^*$ and $\|\boldsymbol{w}\|$. Notice that $\boldsymbol{w}^T\boldsymbol{w}^* \leq \|\boldsymbol{w}\|$ since $\boldsymbol{w}^*$ is a unit vector. In each step, $\boldsymbol{w}^T\boldsymbol{w}^*$ increases by at least ν since $(\boldsymbol{w} + \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|})^T\boldsymbol{w}^* = \boldsymbol{w}^T\boldsymbol{w}^* + \frac{\boldsymbol{a}_i^T\boldsymbol{w}^*}{\|\boldsymbol{a}_i\|} \geq \boldsymbol{w}^T\boldsymbol{w}^* + \nu$. However, $\|\boldsymbol{w}\|^2$ never increases by more than 1 in a given step since $(\boldsymbol{w} + \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|})^2 = \boldsymbol{w}^2 + 2\frac{\boldsymbol{a}_i^T}{\|\boldsymbol{a}_i\|}\boldsymbol{w} + (\frac{\boldsymbol{a}_i}{\|\boldsymbol{a}_i\|})^2 \leq (\boldsymbol{w}^2 + 1)$, where we observed that $\frac{\boldsymbol{a}_i^T}{\|\boldsymbol{a}_i\|}\boldsymbol{w} < 0$ for any i we would use in an update step. Therefore, after t steps we have $\boldsymbol{w}^T\boldsymbol{w}^* \geq \nu t$ and $\|\boldsymbol{w}\| \leq \sqrt{t}$. The observation that $\boldsymbol{w}^T\boldsymbol{w}^* \leq \|\boldsymbol{w}\|$ implies $t\nu \leq \sqrt{t}$, and therefore $t \leq 1/\nu^2$. □

In the learning setting, the problem instance consists of points (examples), and the solution is a hyperplane. In the linear programming setting, the problem instance consists of hyperplanes (constraints) and the solution is a point. To transform a linear programming feasibility problem into a separating hyperplane problem, we define the *polar* of a conic linear program.

Definition 15 (Polar) *For any d-dimensional space S filled with points and d-dimensional hyperplanes through the origin, we define the polar of S to be the d-dimensional space $P(S)$, where, for every point $\mathbf{p}$ in S, we define a hyperplane $\mathbf{p}^T\boldsymbol{x} = 0$ in $P(S)$, and for every hyperplane $\mathbf{h}^T\boldsymbol{x} = 0$ in S, we define a point $\mathbf{h}$ in $P(S)$.*

Because the feasibility problem $\boldsymbol{h}^T\boldsymbol{x} > 0$ is a cone, any feasible point $\boldsymbol{x}$ defines a feasible ray from the origin. Thus it is fair to say $P(P(S)) = S$, because two distinct points in S may map to the same hyperplane in $P(S)$, but in this case they belonged to the same ray in S, which makes them equivalent for our purposes. Because $P(P(S)) = S$, the polar is sometimes called the geometric dual.

3.6.2 The Input Mapping

For the linear programming feasibility problem

$$\boldsymbol{a}_i^T\boldsymbol{x} \quad \leq \quad b_i \quad \forall i$$

we create the conic linear program

$$\begin{aligned} (-\boldsymbol{a}_i, b_i)^T(\boldsymbol{x}, x_0) \quad &\geq \quad 0 \quad \forall i \\ x_0 \quad &> \quad 0 \end{aligned}$$

which is also a separating hyperplane problem when viewed in the polar space. We apply the perceptron algorithm with the modification that it tests for inequality or strict inequality as appropriate in step 2. We now relate the wiggle room to the primal condition number.

3.6.3 Wiggle Room

Let M denote the matrix whose i^{th} row is $\boldsymbol{m}_i = \frac{(-\boldsymbol{a}_i, b_i)}{\sqrt{n}\|(\boldsymbol{a}_i, b_i)\|}$. Then the system we are considering is

$$M(\boldsymbol{x}, x_0) \geq \boldsymbol{0}, \quad (\boldsymbol{x}, x_0) \in \boldsymbol{C} = \{(\boldsymbol{x}, x_0) : x_0 > 0\},$$

where $\boldsymbol{C}$ is an open convex cone and $\|M\|_F = 1$. For this system,

$$C_P = \frac{\|M\|_F}{\rho(M, \boldsymbol{C})} = \frac{1}{\rho(M, \boldsymbol{C})}$$

Since ν is implicitly defined with respect to a set of linear inequalities, let $\nu(M, \boldsymbol{C})$ denote the wiggle room for the set of constraints defining $M(\boldsymbol{x}, x_0) \geq \boldsymbol{0}, (\boldsymbol{x}, x_0) \in \boldsymbol{C}$. Since $\boldsymbol{C}$ is an open convex cone, there is some such set of constraints.

Claim 1 (Wiggle Room and Primal Condition Number) *For* $\boldsymbol{C} = \{(\boldsymbol{x}, x_0) : x_0 > 0\}$,

$$\nu(M, \boldsymbol{C}) \leq \rho(M, \boldsymbol{C}) \leq (3d)\nu(M, \boldsymbol{C})$$

Proof: Suppose there exists a unit vector $\boldsymbol{w}^*$ with wiggle room ν. Then $\boldsymbol{w}^* \in \boldsymbol{C}$ and for every i, $\rho(\boldsymbol{m}_i, \textbf{Ray}\,(\boldsymbol{w}^*)) \geq \nu$. This implies $\rho(M, \textbf{Ray}\,(\boldsymbol{w}^*)) \geq \nu$ and hence $\rho(M, \boldsymbol{C}) \geq \nu$.

Now suppose $\rho(M, \boldsymbol{C}) = \eta$. Then there exists a unit vector $\boldsymbol{w} \in \boldsymbol{C}$ such that $\rho(\boldsymbol{m}_i, \textbf{Ray}\,(\boldsymbol{w})) \geq \eta/d$ for every i. This vector $\boldsymbol{w}$ is not too close to any constraint $\boldsymbol{m}_i$, but it might be very close to $\boldsymbol{C}$. To fix this, we construct $\boldsymbol{w}' = \boldsymbol{w} + (\boldsymbol{0}, \eta/2d)$. Since $\|\boldsymbol{w}'\| \leq 1 + \frac{\eta}{2d} \leq 3/2$, we have $\boldsymbol{w}'^T \boldsymbol{m}_i / \|\boldsymbol{w}'\| \geq \frac{1}{\|\boldsymbol{w}'\|}(\frac{\eta}{d} - \frac{\eta}{2d}) \geq \frac{\eta}{3d}$ and $\boldsymbol{w}'^T(\boldsymbol{0}, 1)/\|\boldsymbol{w}'\| \geq \frac{1}{\|\boldsymbol{w}'\|}(\frac{\eta}{2d}) \geq \frac{\eta}{3d}$ as well. Therefore $\frac{\boldsymbol{w}'}{\|\boldsymbol{w}'\|}$ has wiggle room at least $\frac{\eta}{3d}$. □

3.6.4 Smoothed Analysis of the Perceptron Algorithm

Under the hypothesis of theorem 5 on $A, \boldsymbol{b}$ defining a linear system $A\boldsymbol{x} \leq \boldsymbol{b}$, $\sigma^2 \leq 1/nd$, we have that $\|A, \boldsymbol{b}\|_F = O(1)$ with high probability. From lemma 15 we have that

$$\rho((A, -\boldsymbol{b}), \boldsymbol{C}) \geq \frac{\epsilon\sigma}{4nd^{5/4}}$$

with probability at least $1 - \epsilon$, and hence

$$\rho(M, \boldsymbol{C}) \geq \frac{\epsilon\sigma}{4nd^{5/4}} \implies \nu \geq \frac{\epsilon\sigma}{12nd^{9/4}}$$

An argument that does not use lemma 15 as a black box gives $\nu \geq \frac{\epsilon\sigma}{12nd^{5/4}}$ (since in lemma 15 we constucted a point that was this far awar from every constraint simultaneously), and therefore implies the following theorem:

Theorem 7 (Smoothed Complexity of the Perceptron Algorithm for Feasibility) *Under the hypothesis of theorem 5 for the linear progamming feasibility problem* $A\boldsymbol{x} \leq \boldsymbol{b}$,

with probability at least $1-\epsilon$, one of the following holds
(i) the perceptron algorithm returns a feasible solution in at most $\frac{144n^2d^{5/2}}{\epsilon^2\sigma^2}$ iterations
(ii) the problem is infeasible

Suppose now that we seek to bound the time to optimize, not just to solve the feasibility problem. Consider the system $\boldsymbol{c}^T\boldsymbol{x} \geq c_0,\ A\boldsymbol{x} \leq \boldsymbol{b}$, where $\{\boldsymbol{c}, c_0\}$ are fixed, but $\{A, \boldsymbol{b}\}$ are random variables as before. Rather than considering $\boldsymbol{C} = \{(\boldsymbol{x}, x_0) : x_0 > 0\}$, consider $\boldsymbol{C'} = \{(\boldsymbol{x}, x_0) : \boldsymbol{c}^T\boldsymbol{x} > c_0^T x_0\}$. Then we have that $Mx \geq 0, \boldsymbol{x} \in \boldsymbol{C'}$ has wiggle room at least $\frac{\epsilon\sigma}{12nd^{5/4}}$ if it is feasible at all. We assume that c_0 is known in theorem 8 – not assuming this would require multiplying by the time to do binary search on c_0. The theorem on optimization is:

Theorem 8 (Smoothed Complexity of the Perceptron Algorithm for Optimization) *Consider the linear progamming problem*

$$\max \boldsymbol{c}^T\boldsymbol{x} \quad s.t. \quad A\boldsymbol{x} \leq \boldsymbol{b}$$

under the hypothesis of theorem 5 but with $\boldsymbol{c}, c_0$ fixed. Define $p(c_o)$ to be the probability that the objective value c_0 is strictly achievable and the maximum is well-defined, i.e.,

$$p(c_0) = \Pr[\boldsymbol{c}^T\boldsymbol{x} > c_0 \textit{ for some } \boldsymbol{x} \textit{ s.t. } A\boldsymbol{x} \leq \boldsymbol{b} \textit{ and the linear program is bounded}]$$

Then with probability at least $p(c_0) - \epsilon$, the perceptron algorithm run on the system

$$\boldsymbol{c}^T\boldsymbol{x} > c_0, \quad A\boldsymbol{x} \leq \boldsymbol{b}$$

returns a feasible solution in at most $\frac{144n^2d^{5/2}}{\epsilon^2\sigma^2}$ iterations

To see the strength of the guarantee provided by the theorem, consider that if we ignore the case that the objective value is unbounded, the objective value c_0 is only strictly achievable with probability $p(c_0)$. Most of the theorem is straightforward from our previous discussion: with probability at least $1-\epsilon$, the linear program is either infeasible or has good wiggle room, and the program is either infeasible or unboudned with probabiltiy $1-p(c_0)$. From this we can lower bound the probability that the program is feasible and bounded with good wiggle room

$$\begin{aligned} \Pr[\text{good wiggle room and bounded}] &\geq \\ \Pr[\text{good wiggle room or infeasible}] - \Pr[\text{infeasible or unbounded}] &\geq \\ (1-\epsilon) - (1-p(c_0)) &= p(c_0) - \epsilon \end{aligned}$$

To see why we add the curious caveat about the linear program being bounded, consider the possibility that the perceptron algorithm returns a solution $(\boldsymbol{x}, x_0)$ with $x_0 < 0$. Then we do not know whether the system $\boldsymbol{c}^T\boldsymbol{x} \geq c_0,\ A\boldsymbol{x} \leq \boldsymbol{b}$ is feasible, but if it is, we can construct a solution of arbitrarily large objective value as follows: let $\boldsymbol{x'}$ be a solution to $\boldsymbol{c}^T\boldsymbol{x} \geq c_0,\ A\boldsymbol{x} \leq \boldsymbol{b}$. We find that $\boldsymbol{x'} + \lambda(\boldsymbol{x'} - \frac{\boldsymbol{x}}{x_0})$ satisfies $A\boldsymbol{x} \leq \boldsymbol{b}$ for any positive λ, and the objective value grows without bound as λ increases.

To illuminate why we do not simply consider the open convex cone $\boldsymbol{C''} = \boldsymbol{C'} \cap \boldsymbol{C}$, note that $\boldsymbol{C''}$ might by itself have very small wiggle room. The wiggle room is approximately half the angle between the vectors $(\boldsymbol{c}, -c_0)$ and $(\boldsymbol{0}, 1)$, which is $O(\frac{\|\boldsymbol{c}\|}{c_0})$. This value can be arbitrarily small independent of the other parameters we have specified.

3.7 Alternative Models of Perturbation

3.7.1 The Original Spielman-Teng Model

The model in [ST 01] is to start with a linear program L given by

$$\begin{array}{rrcl} & \max & \boldsymbol{c}^T\boldsymbol{x} & \\ \text{s.t.} & \boldsymbol{a}_i^T x & \leq & b_i \quad \forall i \in \{1, \dots, n\} \\ & \|\boldsymbol{a}_i\| & \leq & 1 \quad \forall i \\ & b_i & \in & \{\pm 1\} \quad \forall i \end{array}$$

As remarked in [ST 01], any linear program can be transformed in an elementary way into this formulation. Now let $\tilde{\boldsymbol{a}}_i = \boldsymbol{a}_i + \sigma \boldsymbol{g}_i$, where each $\boldsymbol{g}_i$ is chosen independently according to a d-dimensional Gaussian distribution of unit variance and zero mean. Then the perturbed linear program is given by

$$\begin{array}{rrcl} & \max & \boldsymbol{c}^T\boldsymbol{x} & \\ \text{s.t.} & \tilde{\boldsymbol{a}}_i^T x & \leq & b_i \quad \forall i \end{array}$$

The details necessary to extend our analysis of the primal condition number to this model are covered in [BD 02]. Let it suffice to say that it is easier to carry out a smoothed analysis of the condition number in the model considered throughout the bulk of this chapter. Note also that in the original Spielman-Teng model $\|A\|_F$ is about $\sqrt{n}$, which explains the difference by a factor of n between the running time for the perceptron algorithm given in [BD 02] and the running time given here.

3.7.2 Zero-Preserving Perturbations

Many linear programs that are encountered in practice are sparse. We consider here the possibility of modelling this phenomenon using *zero-preserving additive perturbations*. We show that in this model it is not possible to bound the condition number by $poly(n, d, \frac{1}{\sigma})$ with probability at least 1/2.

Consider $A\boldsymbol{x} \geq \boldsymbol{0}$, where $\|\bar{A}\|_F \leq 1$ and A is centered at $\bar{A}$, but only the nonzero entries of A are Gaussian random variables of variance σ^2; the rest are fixed to be equal to zero. For ease of exposition, we will normalize $\|\bar{A}\|_F$ to be 1 at the end of this subsection. Define the matrix

$$A = \begin{bmatrix} 1 & -\epsilon & & \\ & 1 & -\epsilon & \\ \dots & & & \end{bmatrix}$$

where ϵ is a parameter (assumed to be small), and consider the linear program $A\boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \boldsymbol{C}$ where $\boldsymbol{C} = \{\boldsymbol{x} : \boldsymbol{x} \geq \boldsymbol{0}\}$. We have $\|A\|_F \approx \sqrt{n}$, while the i^{th} constraint is exactly

$$x_i \geq \epsilon x_{i+1}$$

Adding a zero-preserving additive perturbation of magnitude σ^2, we find that

$$\Pr[|a_{i,i} - 1| \geq \delta] \leq e^{-\frac{\delta^2}{4\sigma^2}} \qquad \frac{\delta^2}{\sigma^2} > 2 \tag{3.13}$$

$$\Pr[|a_{i,i+1} - \epsilon| \geq \delta] \leq e^{-\frac{\delta^2}{4\sigma^2}} \qquad \frac{\delta^2}{\sigma^2} > 2 \tag{3.14}$$

by applying corollary 4. Setting $\delta = \sigma\sqrt{8 \log n}$ yields that none of the events (3.13), (3.14) happen for any i with probability at least 1/2 (by a union bound). Assuming that none of the events occur, we have that $A\boldsymbol{x} \geq \boldsymbol{0}, \boldsymbol{x} \in \boldsymbol{C}$ is still feasible, but $\rho(A, \boldsymbol{C})$ is at most

$$\left(\frac{\epsilon + \delta}{1 - \delta}\right)^n = \left(\frac{\epsilon + \sigma\sqrt{8\log n}}{1 - \sigma\sqrt{8\log n}}\right)^n$$

which is exponentially small ($(\frac{O(1)}{n})^n$), for $\epsilon = \frac{1}{n}$ and $\sigma = \frac{1}{n^2}$. Normalizing so that $\|A\|_F = 1$ is equivalent to using $\sigma \approx \frac{1}{n\sqrt{n}}$, which still shows the negative result.

This analysis may easily be extended to show that *zero-preserving multiplicative perturbations* can also lead to very small wiggle room.

3.7.3 Non-Gaussian Perturbations

Suppose we wanted to model A as a uniformly distributed random variable within the ball of radius σ' centered at $\bar{A}$. In order to extend all the results of this chapter to this model, we just need the following lemma and easy fact:

Lemma 28 (Small Boundaries are Easily Missed, Uniform Perturbation Case) *Let $\boldsymbol{g}$ be a random variable that is uniformly distributed within the ball of radius σ' centered at $\boldsymbol{g}_0$, and let $\boldsymbol{K}$ be an arbitrary convex body. Then,*

$$\Pr[\boldsymbol{g} \in \mathbf{bdry}(\boldsymbol{K}, \epsilon)] \leq \left(\frac{\epsilon d}{\sigma'}\right)$$

Proof: As in the proof of lemma 30, the volume of $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ is at most ϵ times the surface area of a ball of radius σ', which is $\frac{2(\sigma')^{d-1}\pi^{d/2}}{\Gamma(d/2)}$. Since the volume of a ball of radius σ' is $\frac{2(\sigma')^d\pi^{d/2}}{d\Gamma(d/2)}$, the ratio between these two quantities is $\frac{\epsilon d}{\sigma'}$. □

As σ' is roughly analogous to $\sigma\sqrt{d}$, this bound matches lemma 30. While the bound in lemma 30 turns out not to be tight, lemma 28 clearly is.

Fact 3 (Bounds on the Magnitude of a Uniform Random Variable) *Let $\boldsymbol{g}$ be a random variable that is uniformly distributed within the d-dimensional ball of radius σ' centered at the origin. Then*

$$\Pr[\|\boldsymbol{g}\| > \sigma'] = 0$$

$$\Pr[\|\boldsymbol{g}\| \leq c\sigma'] = c^d$$

The proof of this fact is straightforward from the formula for the volume of a d-dimensional ball. The bounds are very similar to those derived for Gaussian random variables in section 3.8.1.

Results for this alternative model of perturbation analogous to everything developed earlier in this chapter follow straightforwardly. Extending this theory to other models of perturbation may be done in a similar fashion.

3.8 Technical Matters

The statements in this section are used in a black-box manner by the rest of the chapter.

3.8.1 A Bound on the Sum of Gaussian Random Variables

We recall that the probability density function of a Gaussian random variable is given by

$$\mu(x) = (1/\sqrt{2\pi})e^{-x^2/2}.$$

A Gaussian random vector of variance σ^2 is a vector where each element is a Gaussian random variable of variance σ^2. A Gaussian random matrix is defined similarly. The probability density function of a d-dimensional Gaussian random vector of variance σ^2 centered at $\bar{\boldsymbol{x}}$ is given by

$$\mu(\boldsymbol{x}) = (1/(\sigma\sqrt{2\pi})^d)e^{-\|\boldsymbol{x}-\bar{\boldsymbol{x}}\|/(2\sigma^2)}.$$

The distribution we are analyzing is the Chi-Squared distribution, and bounds of this form are well-known. We thank DasGupta and Gupta [DG 99] for this particular derivation.

Fact 4 (Sum of Gaussians) *Let $X_1, \ldots, X_d$ be independent $N(0,\sigma)$ random variables. Then*

$$\Pr[\sum_{i=1}^{d} X_i^2 \geq \kappa^2] \leq e^{\frac{d}{2}(1-\frac{\kappa^2}{d\sigma^2}+\ln\frac{\kappa^2}{d\sigma^2})}$$

Proof: For simplicity, we begin with $Y_i \sim N(0,1)$. A simple integration shows that if $Y \sim N(0,1)$ then $E[e^{tY^2}] = \frac{1}{\sqrt{1-2t}} \quad (t < \frac{1}{2})$. We proceed with

$$\begin{array}{rcl}
\Pr[\sum_{i=1}^{d} Y_i^2 \geq k] & = & \\
\Pr[\sum_{i=1}^{d} Y_i^2 - k \geq 0] & = & (\text{for } t > 0) \\
\Pr[e^{t(\sum_{i=1}^{d} Y_i^2 - k)} \geq 1] & \leq & (\text{by Markov's Ineq.}) \\
\mathbf{E}[e^{t(\sum_{i=1}^{d} Y_i^2 - k)}] & = & \\
\left(\frac{1}{1-2t}\right)^{d/2} e^{-kt} & \leq & (\text{letting } t = \frac{1}{2} - \frac{d}{2k}) \\
\left(\frac{k}{d}\right)^{d/2} e^{-\frac{k}{2}+\frac{d}{2}} & = & e^{\frac{d}{2}(1-\frac{k}{d}+\ln\frac{k}{d})}
\end{array}$$

Since

$$\Pr[\sum_{i=1}^{d} Y_i^2 \geq k] = \Pr[\sum_{i=1}^{d} X_i^2 \geq \sigma^2 k]$$

we set $k = \frac{\kappa^2}{\sigma^2}$ and obtain $e^{\frac{d}{2}(1-\frac{k}{d}+\ln\frac{k}{d})} = e^{\frac{d}{2}(1-\frac{\kappa^2}{d\sigma^2}+\ln\frac{\kappa^2}{d\sigma^2})}$ which was our desired bound. □

Fact 5 (Alternative Sum of Gaussians) *Let $X_1, \ldots, X_d$ be independent $N(0, \sigma)$ random variables. Then*

$$\Pr[\sum_{i=1}^{d} X_i^2 \geq cd\sigma^2] \leq e^{\frac{d}{2}(1-c+\ln c)} \qquad c \geq 1$$

$$\Pr[\sum_{i=1}^{d} X_i^2 \leq cd\sigma^2] \leq e^{\frac{d}{2}(1-c+\ln c)} \qquad c \leq 1$$

Proof: The first inequality is proved by setting $k = cd$ in the last line of the proof of fact 4. To prove the second inequality, begin the proof of fact 4 with $\Pr[\sum_{i=1}^{d} Y_i^2 \leq k]$ and continue in the obvious manner. □

Corollary 3 *Let $\boldsymbol{x}$ be a d-dimensional Gaussian random vector of variance σ^2 centered at the origin. Then, for $d \geq 2$ and $\epsilon \leq 1/e^2$,*

$$\mathbf{Pr}\left[\|\boldsymbol{x}\| \geq \sigma\sqrt{d(1 + 2\ln(1/\epsilon)}\right] \leq \epsilon$$

Proof: Set $c = 1 + 2\ln(1/\epsilon)$ in fact 5. We then compute

$$e^{\frac{d}{2}(1-c+\ln c)} \leq e^{1-c+\ln c} \leq e^{-2\ln\frac{1}{\epsilon}+\ln(1+2\ln\frac{1}{\epsilon})} = \epsilon e^{-\ln\frac{1}{\epsilon}+\ln(1+2\ln\frac{1}{\epsilon})}$$

We now seek to show

$$\begin{aligned} e^{-\ln\frac{1}{\epsilon}+\ln(1+2\ln\frac{1}{\epsilon})} &\leq 1 \\ \Leftrightarrow \quad -\ln\frac{1}{\epsilon} + \ln(1 + 2\ln\frac{1}{\epsilon}) &\leq 0 \\ \Leftrightarrow \quad 1 + 2\ln\frac{1}{\epsilon} &\leq \frac{1}{\epsilon} \end{aligned}$$

For $\epsilon = 1/e^2$, the left-hand side of the last inequality is 5, while the right-hand side is greater than 7. Taking derivatives with respect to $1/\epsilon$, we see that the right-hand side grows faster as we increase $1/\epsilon$ (decrease ϵ), and therefore will always be greater. □

Corollary 4 *Let $\boldsymbol{x}$ be a d-dimensional Gaussian random vector of variance σ^2 centered at the origin. Then, for $d \geq 2$,*

$$\mathbf{Pr}\left[\|\boldsymbol{x}\| \leq \epsilon\right] \leq \frac{e\epsilon}{\sigma}$$

Proof: If $\epsilon \leq \sigma$, set $c = \frac{\epsilon^2}{d\sigma^2}$ in fact 5.

$$e^{\frac{d}{2}(1-c+\ln c)} \leq e^{1-c+\ln c} \leq e^{1+\ln c} = \frac{e\epsilon^2}{d\sigma^2} \leq \frac{e\epsilon}{\sigma}$$

If $\epsilon > \sigma$, the statement is vacuously true. □

3.8.2 An Application of the Brunn-Minkowski Theory

We state the following analogue of lemma 18 and show how it can be derived from the Brunn-Minkowski theory of convex bodies.

Lemma 29 (Brunn-Minkowski) *Let $\boldsymbol{K}$ be a d-dimensional convex body, and let $\bar{\boldsymbol{x}}$ denote the center of mass of $\boldsymbol{K}$, $\bar{\boldsymbol{x}} = \mathbf{E}_{\boldsymbol{x}\in\boldsymbol{K}}[\boldsymbol{x}]$. Then for every $\boldsymbol{w}$,*

$$\frac{\max_{\boldsymbol{x}\in\boldsymbol{K}} \boldsymbol{w}^T(\boldsymbol{x}-\bar{\boldsymbol{x}})}{\max_{\boldsymbol{x}\in\boldsymbol{K}} \boldsymbol{w}^T(\bar{\boldsymbol{x}}-\boldsymbol{x})} \leq d$$

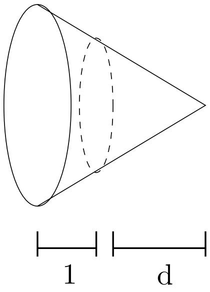

Figure 3-1: Worst case $\boldsymbol{K}$ for lemma 29.

Proof: The entire proof consists of showing that figure 3-1 is the worst case for the bound we want. Without loss of generality, let $\bar{\boldsymbol{x}}$ be the origin. Let $\boldsymbol{K}$ and $\boldsymbol{w}$ be fixed, and let $\boldsymbol{w}$ be a unit vector. Consider the body $\boldsymbol{K'}$ that is rotationally symmetric about $\boldsymbol{w}$ and has the same $(d-1)$-dimensional volume for every cross section $\boldsymbol{K}_r = \{\boldsymbol{x} : \boldsymbol{x} \in \boldsymbol{K}, \boldsymbol{w}^T\boldsymbol{x} = r\}$, i.e., $vol_{d-1}(\boldsymbol{K}_r) = vol_{d-1}(\boldsymbol{K}'_r)$. $\boldsymbol{K'}$ is referred to as the *Schwarz rounding* of $\boldsymbol{K}$ in [GW 93]. $\boldsymbol{K'}$ has the same mean as $\boldsymbol{K}$, and also the same min and max as $\boldsymbol{K}$ when we consider the projection along $\boldsymbol{w}$, but $\boldsymbol{K'}$ will be easier to analyze. Denote the radius of the $(d-1)$-dimensional ball $\boldsymbol{K}'_r$ by $radius(\boldsymbol{K}'_r)$. That $\boldsymbol{K'}$ is convex follows from the Brunn-Minkowski inequality

$$vol_n((1-\lambda)\boldsymbol{A} + \lambda\boldsymbol{B})^{1/n} \geq (1-\lambda)vol_n(\boldsymbol{A})^{1/n} + (\lambda)vol_n(\boldsymbol{B})^{1/n}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are convex bodies in $\mathcal{R}^n$, $0 \leq \lambda \leq 1$, and $+$ denotes the Minkoski sum. Proofs of this inequality can be found in both [Gar 02] and [GW 93]. To see the implication of the theorem from the inequality, let $\boldsymbol{A}, \boldsymbol{B}$ be two cross sections of $\boldsymbol{K}$, $\boldsymbol{A} = \boldsymbol{K}_{r_1}$, $\boldsymbol{B} = \boldsymbol{K}_{r_2}$ and consider the cross-section $\boldsymbol{K}_{(r_1+r_2)/2}$. By convexity of $\boldsymbol{K}$, $\frac{1}{2}\boldsymbol{A} + \frac{1}{2}\boldsymbol{B} \subset \boldsymbol{K}_{(r_1+r_2)/2}$, and therefore

$$vol_{d-1}(\boldsymbol{K}_{(r_1+r_2)/2})^{1/(d-1)} \geq \frac{1}{2}vol_{d-1}(\boldsymbol{K}_{r_1})^{1/(d-1)} + \frac{1}{2}vol_{d-1}(\boldsymbol{K}_{r_2})^{1/(d-1)}$$

This implies that $radius(\boldsymbol{K}'_{(r_1+r_2)/2}) \geq \frac{1}{2}radius(\boldsymbol{K}'_{r_1}) + \frac{1}{2}radius(\boldsymbol{K}'_{r_2})$, which yields that $\boldsymbol{K'}$ is convex.

Let $radius(\boldsymbol{K}'_0) = R$, and let $[\max \boldsymbol{w}^T(\boldsymbol{x}-\bar{\boldsymbol{x}})] = r_0$. Then $radius(\boldsymbol{K}'_r) \geq R(1-\frac{r}{r_0})$ for $r \in [0, r_0]$ by convexity. Similarly, $radius(\boldsymbol{K}'_r) \leq R(1-\frac{r}{r_0})$ for $r < 0$ by convexity. Using our assumption that the center of mass coincides with the origin, we can derive that the least possible value for $r_1 = [\max \boldsymbol{w}^T(\bar{\boldsymbol{x}}-\boldsymbol{x})]$ is given by $\int_{r=0}^{r_1} r(1+\frac{r}{r_0})^{d-1}dr = \int_{r=0}^{r_0} r(1-\frac{r}{r_0})^{d-1}dr$ which yields $r_1 = \frac{r_0}{d}$. □

3.8.3 Small Boundaries are Easily Missed

Throughout this subsection, let $\boldsymbol{K}$ be an arbitrary convex body, and let $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ denote the ϵ-boundary of $\boldsymbol{K}$, i.e.,

$$\mathbf{bdry}(\boldsymbol{K}, \epsilon) = \{\boldsymbol{x} : \exists \boldsymbol{x}' \in \boldsymbol{K}, \left\|\boldsymbol{x} - \boldsymbol{x}'\right\| \leq \epsilon\} \setminus \boldsymbol{K}$$

Let $\boldsymbol{g}$ be chosen according to a d-dimensional Gaussian distribution with mean $\bar{\boldsymbol{g}}$ and variance σ^2, $\boldsymbol{g} \sim N(\bar{\boldsymbol{g}}, \sigma)$.

We thank Ryan O'Donnell for directing us to this theorem by Keith Ball[Bal 93].

Theorem 9 (K. Ball) *Let g denote the probability density function of $\boldsymbol{g}$. For any convex body $\boldsymbol{K}$,*

$$\int_{\mathbf{bdry}(\boldsymbol{K})} g \leq \frac{4d^{1/4}}{\sigma}$$

Lemma 13 (restated here for convenience) is an easy corollary of theorem 9.

Lemma 13 (Small Boundaries are Easily Missed)

$$\Pr[\boldsymbol{g} \in \mathbf{bdry}(\boldsymbol{K}, \epsilon)] \leq \left(\frac{4\epsilon d^{1/4}}{\sigma}\right)$$

Proof:

$$\Pr[\boldsymbol{g} \in \mathbf{bdry}(\boldsymbol{K}, \epsilon)] = \int_{\epsilon'=0}^{\epsilon} \int_{\mathbf{bdry}(\boldsymbol{C}_{\epsilon'})} g \leq \frac{4\epsilon d^{1/4}}{\sigma}$$

where $\boldsymbol{C}_{\epsilon'}$ is the convex body consisting of points within distance ϵ' of $\boldsymbol{K}$. □

Ryan O'Donnell also communicated to us that F. Nazarov has proved a matching lower bound for theorem 9.

Ball's proof uses integral geometry tools that may lie outside the repertoire of most theoretical computer scientists. For this reason, we provide an entirely self-contained proof below of the slightly weaker statement.

Lemma 30 (Small Boundaries are Easily Missed – Weaker Version)

$$\Pr[\boldsymbol{g} \in \mathbf{bdry}(\boldsymbol{K}, \epsilon)] = O\left(\frac{\epsilon d^{1/2}}{\sigma}\right)$$

This weaker statement also appeared in [Bal 93] and [BR 76] (again, thanks to Ryan O'Donnell for these pointers). On a humble note, Ball derived lemma 30 in 6 lines, while we take two pages. The train of thought is the same, but Ball uses some properties of Gaussians that we were unaware of when coming up with the following proof.

Before proving lemma 30, we prove fact 6, which will be useful in proving lemma 30.

Fact 6 (Surface Area of a Convex Body) *Let $\boldsymbol{A}$ be a convex body in $\mathcal{R}^d$, $\boldsymbol{A} \subset B$.*

$$vol_{d-1}(\mathbf{bdry}(\boldsymbol{A})) \leq vol_{d-1}(\mathbf{bdry}(B))$$

Proof: Because $\boldsymbol{A}$ is convex, we can imagine transforming B into $\boldsymbol{A}$ by a series of hyperplane cuts, where on each such cut we throw away everything from B on one side of the hyperplane. The surface area of B strictly decreases after each cut, until finally B equals $\boldsymbol{A}$. □

Proof of Lemma 30: This bound is easily seen to be tight to within a factor of $\Theta(\sqrt{d})$: let $\boldsymbol{K}$ be a hyperplane passing through $\bar{\boldsymbol{g}}$. For the proof, we divide space into thin shells of a hypersphere (like an onion) centered at $\bar{\boldsymbol{g}}$. We then argue that we are likely to land in a shell where we are about as likely to be in any one part of the shell as any other. Furthermore, in this shell, $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ can't be more than a small fraction of the overall volume of the shell.

Without loss of generality, let $\bar{\boldsymbol{g}}$ be the origin. Recall that the probability density function of $\boldsymbol{g}$ is given by

$$\mu(\boldsymbol{x}) = \left(1/\sqrt{2\pi}\right)^d e^{-\|\boldsymbol{x}\|^2/2}$$

Fix $\gamma > 0$. Let $S_R = \{\boldsymbol{x} : R \leq |\boldsymbol{x}| \leq (1 + \frac{\gamma}{d})R\}$ be the thin shell.

We would like to be able to argue that, if $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ is a small fraction of the volume of S_R, then if we condition on $\boldsymbol{g}$ landing within S_R, we are unlikely to land in $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$. The concept of *bias* allows us to make this argument. Define the *bias* of a region X by

$$bias(X) = \frac{\max_{\boldsymbol{x} \in X} \mu(\boldsymbol{x})}{\min_{\boldsymbol{x} \in X} \mu(\boldsymbol{x})}$$

Then we can say that, for any $Y \subset X$,

$$\Pr[\boldsymbol{g} \in Y | \boldsymbol{g} \in X] \leq \frac{vol(Y)}{vol(X)} \cdot bias(X)$$

For S_R, we calculate

$$bias(S_R) = \frac{e^{-R^2/\sigma^2}}{e^{-(1+\gamma/d)^2 R^2/\sigma^2}} = e^{(2\gamma/d + \gamma^2/d^2)R^2/\sigma^2}$$

We upper bound the probability of landing in $\mathbf{bdry}(\boldsymbol{K}, \epsilon)$ using

$$\Pr[\boldsymbol{g} \in \mathbf{bdry}(K, \epsilon) | \boldsymbol{g} \in S_R] \leq \frac{vol(\mathbf{bdry}(\boldsymbol{K}, \epsilon) \cap S_R)}{vol(S_R)} \cdot bias(S_R)$$

Let $\boldsymbol{B}$ be a ball of radius $(1 + \frac{\gamma}{d})R$. Let $\boldsymbol{K'}$ be the convex closure of $\mathbf{bdry}(\boldsymbol{K}, \epsilon) \cap S_R$. Clearly $\boldsymbol{K'} \subset \boldsymbol{B}$. We can upper bound $vol(\mathbf{bdry}(\boldsymbol{K}, \epsilon) \cap S_R)$ by $\epsilon \cdot vol_{d-1}(\mathbf{bdry}(\boldsymbol{K'}))$, and by fact 6, this is at most $\epsilon \cdot vol_{d-1}(\boldsymbol{B})$. The exact formulas for the volume and surface area of a sphere are

$$vol(S_R) = \frac{2((1 + \frac{\gamma}{d})R)^d \pi^{d/2}}{d\Gamma(d/2)} - \frac{2R^d \pi^{d/2}}{d\Gamma(d/2)}$$

$$vol_{d-1}(\boldsymbol{B}) = \frac{2((1 + \frac{\gamma}{d})R)^{d-1} \pi^{d/2}}{\Gamma(d/2)}$$

which yields

$$\frac{vol(\mathbf{bdry}(\boldsymbol{K}, \epsilon) \cap S_R)}{vol(S_R)} bias(S_R) \leq$$

$$\frac{d\epsilon}{R}\cdot\frac{(1+\frac{\gamma}{d})^{d-1}}{(1+\frac{\gamma}{d})^{d}-1}\cdot e^{\frac{\gamma}{d}(1+\gamma/d)^2(2+\gamma/d)R^2/\sigma^2}$$

To complete the proof, we sum over all the possible shells S_R that $\boldsymbol{g}$ might land in. This is done in the following formula.

$$\begin{aligned}
\Pr[\boldsymbol{g}\in\mathbf{bdry}(\boldsymbol{K},\epsilon)] &\le \sum_{k,R=(1+\frac{\gamma}{d})^k}\Pr[\boldsymbol{g}\in\mathbf{bdry}(\boldsymbol{K},\epsilon)\mid g\in S_R]\Pr[g\in S_R]\\
&\le \sum_{k,R=(1+\frac{\gamma}{d})^k}\Pr[\boldsymbol{g}\in S_R]\cdot\frac{d\epsilon}{R}\cdot\frac{(1+\frac{\gamma}{d})^{d-1}}{(1+\frac{\gamma}{d})^{d}-1}\cdot e^{\frac{\gamma}{d}(1+\gamma/d)^2(2+\gamma/d)R^2/\sigma^2}\\
&\le \mathbf{E}_{\{\boldsymbol{g},|\boldsymbol{g}|=\sigma\sqrt{cd}\}}\left[\frac{\sqrt{d}\epsilon}{\sqrt{c}\sigma}\cdot\frac{(1+\frac{\gamma}{d})^{d}}{(1+\frac{\gamma}{d})^{d}-1}\cdot e^{\gamma(1+\gamma/d)^4(2+\gamma/d)c}\right]
\end{aligned}$$

We use the identity
$\mathbf{E}\boldsymbol{g}[f(\boldsymbol{g})]=\int_{x=0}^{\infty}\Pr\boldsymbol{g}[f(\boldsymbol{g})>x]dx$ to upper bound that last expectation. Also, let $1/\gamma_1=\frac{(1+\frac{\gamma}{d})^d}{(1+\frac{\gamma}{d})^d-1}$ and let $\gamma_2=\gamma(1+\gamma/d)^4(1+\gamma/(2d))$. Then that last expectation is just $\frac{\sqrt{d}\epsilon}{\sigma\gamma_1}\mathbf{E}[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}]$. We compute the upper bound as follows:

$$\begin{aligned}
\mathbf{E}[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}] &= \int_{x=0}^{\infty}\Pr_{\{g,|g|=\sigma\sqrt{cd}\}}[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}>x]dx\\
&= \int_{x=0}^{\infty}\Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}>x,\ c\ge 1]+\Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2 c}>x,\ c<1]dx\\
&\le \int_{x}\Pr[e^{2\gamma_2 c}>x\text{ and }c\ge 1]+\Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2}>x\text{ and }c<1]dx\\
&= \int_{x=e^{2\gamma_2}}^{\infty}\Pr[e^{2\gamma_2 c}>x]dx+\int_{x=e^{2\gamma_2}}^{\infty}\Pr[\frac{1}{\sqrt{c}}e^{2\gamma_2}>x]dx\\
&= \int_{x=e^{2\gamma_2}}^{\infty}\Pr[c>\frac{\ln x}{2\gamma_2}]dx+\int_{x=e^{2\gamma_2}}^{\infty}\Pr[c<\frac{e^{4\gamma_2}}{x^2}]dx\\
&\le \int_{x=e^{2\gamma_2}}^{\infty}e^{\frac{d}{2}(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}}dx+\int_{x=e^{2\gamma_2}}^{\infty}e^{\frac{d}{2}(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx\\
&\le \int_{x=e^{2\gamma_2}}^{\infty}e^{(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}}dx+\int_{x=e^{2\gamma_2}}^{\infty}e^{(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx
\end{aligned}$$

Where on the last step we observe that $1-c'+\ln c'\le 0$ and we assume that $d\ge 2$. We now proceed to analyze the right-hand term.

$$\begin{aligned}
\int_{x=e^{2\gamma_2}}^{\infty}e^{(1-c'+\ln c')}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx &\le \int_{x=e^{2\gamma_2}}^{\infty}e^{1+\ln c'}|_{c'=\frac{e^{4\gamma_2}}{x^2}}dx\\
&= e\int_{x=e^{2\gamma_2}}^{\infty}\frac{e^{4\gamma_2}}{x^2}dx\\
&= e^{2\gamma_2+1}
\end{aligned}$$

For the lefthand term we make the change of variables $x = e^{2\gamma_2 \alpha}$. Continuing:

$$\begin{aligned}
\int_{x=e^{2\gamma_2}}^{\infty} e^{(1-c'+\ln c')}|_{c'=\frac{\ln x}{2\gamma_2}} dx &= \int_{\alpha=1}^{\infty} e^{1-\alpha+\ln\alpha} 2\gamma_2 e^{2\gamma_2\alpha} d\alpha \\
&= 2\gamma_2 e \int_{\alpha=1}^{\infty} \alpha e^{(2\gamma_2-1)\alpha} d\alpha \\
&= 2\gamma_2 e \left[\frac{\alpha}{2\gamma_2 - 1} e^{(2\gamma_2-1)\alpha} - \frac{1}{(2\gamma_2-1)^2} e^{(2\gamma_2-1)\alpha} \right]_{\alpha=1}^{\infty} \\
\text{(since } \gamma_2 < 1/2\text{)} \quad &= 2\gamma_2 e^{2\gamma_2} \left[\frac{1}{(2\gamma_2-1)^2} - \frac{1}{2\gamma_2 - 1} \right]
\end{aligned}$$

Our final bound on $\Pr[\boldsymbol{g} \in \mathbf{bdry}(K, \epsilon)]$ is thus

$$\frac{\epsilon\sqrt{d}}{\sigma} \left(\frac{e^{2\gamma_2}}{\gamma_1} \right) \left(e + \frac{4(\gamma_2 - \gamma_2^2)}{(2\gamma_2 - 1)^2} \right)$$

Letting $\gamma = .1$, we derive that this is at most $\frac{45\epsilon\sqrt{d}}{\sigma}$. This concludes the lemma proof. □

Bibliography

[Adl 83] I. Adler, “The expected number of pivots needed to solve parametric linear programs and the efficiency of the self-dual simplex method,” Technical Report, University of California at Berkeley, May 1983.

[AKS 87] I. Adler, R. M. Karp, and R. Shamir, “A simplex variant solving an m x d linear program in $O(min(m^2, d^2))$ expected number of pivot steps,” in *Journal of Complexity*, 3, 1987, pp372-387.

[AM 85] I. Adler and N. Megiddo, “A simplex algorithm whose average number of steps is bounded between two quadratic functions of the smaller dimension,” in *Journal of the ACM*, 32(4), October 1985, pp871-895.

[Agm 54] S. Agmon, “The relaxation method for linear inequalities,” in *Canadian Journal of Mathematics*, 6(3), 1954, pp382-392.

[Bal 93] K. Ball, “The reverse isoperimetric problem for gaussian measure,” in *Discrete and Computational Geometry*, 10(4), 1993, pp411-420.

[Bie 01] D. Bienstock, “Potential function methods for approximately solving linear programming problems: Theory and Practice,” Computational Optimization Research Center (CORC) at Columbia University TR-2001-06, and Center for Operations Research and Econometric (CORE at U. Catholique de Louvain, Belgium) Lecture Series, ISSN-0771 3894 (2001).

[BR 76] R. Bhattacharya and R. Rao, *Normal approximation and asymptotic expansion*, 1976, pp23-38.

[BD 02] A. Blum and J. Dunagan, “Smoothed Analysis of the Perceptron Algorithm for Linear Programming,” In *Proceedings of the 13th ACM-SIAM Symposium on Discrete Algorithms (SODA 2002)*, San Francisco, 2002, pp905-914. Invited to appear in a special issue of the *Journal of Algorithms*.

[BFKV 99] A. Blum, A. Frieze, R. Kannan and S. Vempala, “A Polynomial-Time Algorithm for Learning Noisy Linear Threshold Functions,” In *Algorithmica*, 22(1), 1999, pp35-52.

[Blu 90] L. Blum, “Lectures on a theory of computation and complexity over the reals (or an arbitrary ring),” in *Lectures in the Sciences of Complexity II*, edited by E. Jen, published by Addison-Wesley, 1990, pp1-47.

[Bor 77] K. H. Borgwardt, “Untersuchungen zur Azymptotik der mittleren Schrittzahl von Simplexverfahren in der linearen Optimierung,” PhD Thesis, Universitat Kaiserslautern, 1977.

[Bor 80] K. H. Borgwardt, “The simplex method: a probabilistic analysis,” in *Algorithms and Combinatorics*, 1, published by Springer-Verlag, 1980.

[Chv 83] V. Chvătal, *Linear programming*, W.H. Freeman, New York, 1983.

[CEMST 93] K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S. Teng, “Approximating center points with iterated radon points,” In *Proceedings of the 9th ACM Symposium on Computational Geometry (SOCG ’93)*, San Diego, CA, 1993, pp91-98. To appear in *International Journal of Computational Geometry & Applications*.

[CP 01] F. Cucker and J. Peña, “A primal-dual algorithm for solving polyhedral conic systems with a finite-precision machine,” Submitted to *SIAM Journal on Optimization*, 2001.

[Dan 51] G. B. Dantzig, “Maximization of a linear function of variables subject to linear inequalities,” in *Activity Analysis of Production and Allocation*, edited by T. C. Koopmans, 1951, pp339-347.

[DG 99] S. Dasgupta, A. Gupta, “An elementary proof of the Johnson-Lindenstrauss Lemma,” International Computer Science Institute, Technical Report 99-006.

[DG 92] D. L. Donoho and M. Gasko, “Breakdown properties of location estimates based on halfspace depth and projected outlyingness,” In *The Annals of Statistics*, 20(4), 1992, pp1803-1827.

[DST 02] J. Dunagan, D. Spielman and S. Teng, “Smoothed Analysis of Renegar’s Condition Number for Linear Programming,” To appear in *Proceedings of the 7th SIAM Conference on Optimization (SIOPT 2002)*, Toronto, 2002.

[DV 01] J. Dunagan and S. Vempala, “Optimal Outlier Removal in High-Dimensional Spaces,” In *Proceedings of the 33rd ACM Symposium on the Theory of Computing (STOC ’01)*, Crete, 2001, pp627-636. Invited to appear in a special issue of *The Journal of Computer and System Sciences (JCSS).*

[FE 00a] R. Freund and M. Epelman, “Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system,” in *Mathematical Programming*, 88(3), 2000, pp451-485.

[FE 00b] R. Freund and M. Epelman, “Condition number complexity of an elementary algorithm for resolving a conic linear system,” Technical Report O.R. Working Paper 319-97, MIT, 2000.

[FE 01] R. Freund and M. Epelman, “A new condition measure, pre-conditioners, and relations between different measures of conditioning for conic linear systems,” Submitted to *SIAM Journal on Optimization*, 2001.

[FM 00] R. Freund and S. Mizuno, "Interior point methods: Current status and future directions," in *High Performance Optimization*, edited by H. Frenk et al., published by Kluwer Academic Publishers, 2000, pp441-466.

[FN 01] R. Freund and M. Nuñez, "Condition-measure bounds on the behavior of the central trajectory of a semi-definite program," in *SIAM Journal on Optimization*, 11(3):818–836, 2001.

[FO 02] R. Freund and F. Ordoñez, "IPM Practical Performance on LPs and the Explanatory Value of Complexity Measures," To appear in *Proceedings of the 7th SIAM Conference on Optimization (SIOPT 2002)*, Toronto, 2002.

[FV 99] R. Freund and J. Vera, "On the complexity of computing estimates of condition measures of a conic linear system," Operations Research Center Working Paper, MIT, 1999, submitted to *Mathematics of Operations Research*, 1999.

[FV 00] R. Freund and J. Vera, "Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm," in *SIAM Journal on Optimization*, 10(1), 2000, pp155-176.

[Gar 02] R. J. Gardner, "The Brunn-Minkowski Inequality," submitted for publication. Available at `http://www.ac.wwu.edu/~gardner/`.

[GW 93] *Handbook of convex geometry*, chapter 1.2, edited by P. M. Gruber, J. M. Wills, published by Elsevier Science Publishers, 1993.

[GDK 63] B. Grunbaum L. Danzer and V. Klee, "Helly's theorem and its relatives," in *Convexity (Proceedings of the Symposia on Pure Mathematics 7)*, American Mathematical Society, 1963, pp101-180.

[Hai 83] M. Haimovich, "The simplex algorithm is very good! : On the expected number of pivot steps and related properties of random linear programs," Technical Report, Columbia University, April 1983.

[KM 72] V. Klee and G. J. Minty, "How good is the simplex algorithm?", in *Inequalities - III*, edited by O. Shisha, published by Academic Press, 1972, pp159-175.

[Kha 79] L. G. Khachiyan, "A Polynomial Algorithm in Linear Programming," in *Doklady Akademia Nauk SSSR*, 1979, pp1093-1096.

[Kar 84] N. Karmarkar, "A New Polynomial Time Algorithm for Linear Programming," in *Combinatorica*, 4, 1984, pp373-395.

[LKS 95] L. Lovász, R. Kannan and M. Simonovits, "Isoperimetric problems for convex bodies and a localization lemma," In *Discrete Computational Geometry* 13, 1995, pp541-559.

[LKS 97] L. Lovász, R. Kannan and M. Simonovits, "Random walks and an $O^*(n^5)$ volume algorithm for convex bodies," In *Random Structures and Algorithms* 11(1), 1997, pp1-50.

[MY 95] R. Maronna and V. Yohai, "The behaviour of the Stahel-Donoho robust multivariate estimator," In *Journal of the American Statistical Association* 90(429), pp330-341, 1995.

[Meg 86] N. Megiddo, "Improved asymptotic analysis of the average number os steps performed by the self-dual simplex algorithm," in *Mathematical Programming*, 35, 1986, pp140-172.

[MP 69] M. Minsky and S. Papert, *Perceptrons: An Introduction to Computational Geometry*, published by The MIT Press, 1969.

[Mur 80] K. G. Murty, "Computational complexity of parametric linear pgramming," in *Mathematical Programming*, 19, 1980, pp213-219.

[Ren 94] J. Renegar, "Some perturbation theory for linear programming," in *Math. Programming*, 65(1, Ser. A), 1994, pp73-91.

[Ren 95a] J. Renegar, "Incorporating condition measures into the complexity theory of linear programming," in *SIAM J. Optim.*, 5(3), 1995, pp506-524.

[Ren 95b] J. Renegar, "Linear programming, complexity theory and elementary functional analysis," in *Math. Programming*, 70(3, Ser. A), 1995, pp279-351.

[Sma 82] S. Smale, "The problem of the average speed of the simplex method," in *Proceedings of the 11th International Symposium on Mathematical programming*, August 1982, pp530-539.

[Sma 83] S. Smale, "On the average number of steps in the simplex method of linear programming," in *Mathematical Programming*, 27, 1983, pp241-262.

[ST 01] D. Spielman and S. Teng, "Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time," In *Proceedings of the 33rd Annual ACM Symposium on the Theory of Computing (STOC '01)*, 2001, pp296-305. Available at `http://math.mit.edu/∼spielman/SmoothedAnalysis/`

[ST 02] D. Spielman and S. Teng, "Models and applications of smoothed analysis," in submission.

[Tod 86] M. J. Todd, "Polynomial expected behavior of a pivoting algorithm for linear complementarity and linear programming problems," in *Mathematical Programming*, 35, 1986, pp173-192.

[Tod 91] M. J. Todd, "Probabilistic models for linear programming," in *Mathematics of Operations Research*, 16(4), 1991, pp671-693.

[Ver 96] J. Vera, "Ill-posedness and the complexity of deciding existence of solutions to linear programs," in *SIAM Journal on Optimization*, 6(3), 1996.